BOKMÅL

NHH

SKOLEEKSAMEN MET4/INT010EKS

Vår, 2017

Dato: 16. mai 2017

Start: 09:00 – 12:00

Antall timer: 3

Foreleser/emneansvarlig kan kontaktes av eksamensvakt på telefon: 97599593

TILLATTE HJELPEMIDLER:

Alle trykte/egenskrevne, kalkulator

Ordbok: èn tospråklig ordbok tillatt

Antall sider, inkludert forside: 11

Alle delspørsmål teller likt ved sensur. **Husk at dere må skrive slik at sensor kan lese det som er skrevet.**

Oppgave 1

En nytilsatt postdoktor ved NHH har lang reisetid til og fra jobb. Han kjører bil hver dag, men lar seg særlig irritere over ettermiddagsrushet, som ofte fører til at postdoktoren må stå mye i kø. Det er to naturlige kjøreruter mellom hjem og arbeid. Den ene (*Osveien*) er kortere, men er ofte mer trafikkert enn den andre (*Hamre*).

Postdoktoren bestemmer seg for å undersøke om det er systematiske forskjeller i reisetid for de to kjørerutene. Han kjører hver rute i tre uker (15 arbeidsdager), og noterer hver dag reisetid i minutter mellom arbeid og hjem.

Han leser datasettet inn i en statistisk programpakke, og får ut følgende deskriptive statistikk (*1st. Qu* og *3rd. Qu* er første og tredje kvantil henholdsvis):

	Mean	St.dev	Min	1st. Qu.	Median	3rd. Qu.	Max	n
osveien	56.47	9.97	45.00	49.00	53.00	61.00	76.00	15
hamre	54.60	3.72	49.00	52.00	54.00	57.00	62.00	15

(a) Test om variansen i reisetid er den samme for de to kjørerutene.

(b) Test om forventet reisetid er den samme for de to kjørerutene. Begrunn eventuelle valg du gjør underveis.

Det finnes en tredje reiserute (*Fanafjellet*), som består av mye smal og dårlig vei, men som i stedet har veldig lite trafikk. Postdoktoren kjører også denne veien 15 ganger, og noterer på samme måte ned reisetid i minutter. Han gjennomfører så en variansanalyse for å sammenligne kjøretiden for alle tre rutene, og får ut følgende utskrift:

```
Analysis of Variance, response = reisetid, treatment = reiserute:

                    Sum of squares       df      Mean square

  Treatment                563.911        2          281.956
  Residual                 1773.07       42          42.2159
  Total                    2336.98       44          53.1131

  F(2, 42) = 281.956 / 42.2159 = 6.6789 [p-value 0.0030]

  Level         n       mean     std. dev

  1            15    56.4667       9.9704
  2            15       54.6       3.7187
  3            15    62.8667       3.6619

  Grand mean = 57.9778
```

(c) Hva er det som testes her, og hva blir konklusjonen? Gjør rede for forutsetningene, og vurder om de er oppfylt i denne situasjonen.

Et alternativ for postdoktoren er å ta buss til og fra jobb. På *bt.no* kunne man 24. januar 2017 lese følgende overskrift:

Andelen snikere gikk opp: *Skyss kontrollerte færre, men knep likevel flere snikere i 2016 sammenlignet med året før.*

Tallene som ligger bak denne påstanden er oppgitt i tabellen under, og gjelder alle kollektivreiser i Hordaland. *Antall avvik* er antall gebyr som billettkontrollørene har skrevet ut:

	2015	2016
Antall kollektivreiser	54.4 mill.	56.5 mill.
Antall kontrollerte	426 382	342 085
Antall avvik	13 064	12 904

(d) Test om andelen som sniker på kollektivtransporten i Hordaland virkelig har økt fra 2015 til 2016.

Postdoktoren er ikke alltid like ærlig, og vurderer muligheten for å snike på bussen når han skal til og fra jobb. Gebyret dersom han blir tatt er kr. 1150.

(e) Bruk tallene for 2016, og estimer postdoktorens forventede månedlige utgift til snikegebyr. Sammenhold svaret med prisen for månedskort på den aktuelle strekningen, som for tiden er kr. 960. Antall arbeidsdager i en måned kan du anta er 20.

Salgs- og markedssjef Hanne Alver Krum uttalte til den samme nyhetssaken at hun ikke tror at det er flere reisende som unnlater å løse billett:

– *Vi er ikke sikker på at tallene viser en reell økning, for vi er blitt flinkere til å treffe med kontrollene. Nå sikter vi oss bevisst inn på tidspunkt og avganger der vi vet at det er flere reisende uten periodebillett, sier Krum.*

(f) Drøft kort om denne ekstra informasjonen får innvirkning på konklusjonene dine i spørsmål (d) og (e)?

Du får opplyst at det i 2014 ble kontrollert 574 005 passasjerer, og at det ble registrert 11 709 avvik.

(g) Bruk data fra 2014, 2015 og 2016 til å predikere hvor stor andel av de kontollerte passasjerene som blir tatt for sniking i 2017 ved hjelp av eksponentiell glatting, med en glattefaktor på 0.5.

Oppgave 2

Nasjonale prøver i grunnskolen ble innført i 2004, og gjennomføres hver høst på 5., 8., og 9. trinn. På Utdanningsdirektoratet sine nettsider kan man lese at *formålet med nasjonale prøver er å gi skolene kunnskap om elevene sine grunnleggende ferdigheter i lesing, regning og engelsk. Informasjonen fra prøvene skal danne grunnlag for underveisvurdering og kvalitetsutvikling på alle nivå i skolesystemet.*

Vi skal i denne oppgaven se nærmere på data fra en slik prøve: nasjonale prøven i lesing som ble gjennomført for 5. trinn i skoleåret 2015-2016. Hver elev har oppnådd en poengsum i intervallet 0-100, og vi har observert gjennomsnittsresultatet for hver kommune, i tillegg til at vi har hentet inn diverse andre karakteristika ved kommunene. Data fra nasjonale prøver er tilgjengelig gjennom *skoleporten.udir.no*.

Vi har følgende variable i datasettet:

Navn	Forklaring
`lesing`	Gjennomsnittlig skår på leseprøven for elevene på 5. trinn
`log(antall)`	Logaritmen til antall elever på 5. trinn i kommunen
`fritak`	Andel elever som av ulike grunner ble fritatt fra å delta på prøven (i prosent)
`driftsutgifter`	Brutto driftsutgifter for grunnskoleutdanning i kommunen (i 1000 kroner per elev)
`mobbing`	Andel *7.-klassinger* i kommunen som i skoletrivselundersøkelsen samme år rapporterte at de var blitt mobbet siste tre måneder (i prosent)
`bokmal`	Dummy som er 1 dersom kommunen har bokmål som administrasjonsspråk
`nynorsk`	Dummy som er 1 dersom kommunen har nynorsk som administrasjonsspråk
`nord`	Dummy som er 1 dersom kommunen ligger enten i Troms eller Finnmark
`log(folketall)`	Logaritmen til antall inbyggere i kommunen

Merk at målformen i en kommune kan være enten nynorsk, bokmål, eller nøytral.

I tabellen under finner du deskriptiv statistikk for disse variablene. Vi har data fra 425 kommuner, men legg merke til at vi for flere av variablene mangler data. Dette skyldes at det i mange kommuner er så få elever at det ville vært mulig å identifisere resultater for veldig små grupper, og til og med for enkeltindivider, dersom gjennomsnittet hadde blitt offentliggjort. Antall observasjoner for hver variabel er gitt i kolonnen helt til høyre.

	Gjennomsnitt	Standardavvik	Min	Median	Max	Antall obs.
lesing	48.44	2.45	38.00	49.00	58.00	385
log(antall)	4.05	1.26	1.10	3.97	8,78	425
fritak	3.71	4.47	0.00	3.10	22.90	229
driftsutgifter	104.00	22.63	67.81	98.78	250.40	425
mobbing	4.67	4.03	0.00	4.40	25.00	302
bokmal	0.37	0.48	0	0	1	425
nynorsk	0.27	0.44	0	0	1	425
nord	0.10	0.30	0	0	1	425
log(folketall)	8.51	1.18	5.30	8.44	13.40	425

Vi benytter lineær regresjon til å undersøke i hvor stor grad responsvariabelen `lesing` blir forklart av de andre variablene. Vi gjennomfører 6 forskjellige analyser, som er sammenfattet i tabellen under. Standardavviket til koeffisientene er gitt i parantes, og legg spesielt merke til raden "N", som angir antall observasjoner som inngår i hver analyse. Dersom minst en av variablene mangler verdi for en gitt kommune, blir denne kommunen tatt ut av analysen.

	(1)	(2)	(3)	(4)	(5)	(6)
log(antall)			0.42**		-0.33	
			(0.14)		(0.76)	
fritak		0.10*				
		(0.04)				
driftsutgifter	-0.02*		0.00	0.00	-0.00	
	(0.01)		(0.01)	(0.01)	(0.01)	
mobbing	0.04		0.03	0.03	0.03	
	(0.03)		(0.03)	(0.03)	(0.03)	
bokmal	0.21		0.18	0.17	0.16	
	(0.27)		(0.27)	(0.27)	(0.27)	
nynorsk	-0.78*		-0.72*	-0.68*	-0.65*	
	(0.31)		(0.30)	(0.30)	(0.31)	
nord	-0.60		-0.69	-0.68	-0.66	
	(0.49)		(0.49)	(0.48)	(0.49)	
log(folketall)				0.44**	0.78	0.82***
				(0.14)	(0.78)	(0.10)
Konstantledd	50.81***	48.32***	46.96***	44.85***	43.47***	41.36***
	(0.80)	(0.24)	(1.53)	(2.09)	(3.82)	(0.93)
N	288	208	288	288	288	385
adj. R^2	0.07	0.027	0.10	0.10	0.10	0.13
S	1.97	2.56	1.94	1.94	1.95	2.28

Signifikanskoder: 0.1%: '***', 1%: '**', 5%: '*'

(a) Gi en kortfattet fortolkning av Analyse (1).

(b) Ta utgangspunkt i Analyse (1), og prediker testresultatet for en kommune i Finnmark med bokmål som administrasjonsspråk, der 5% av elevene på 7. trinn oppgir at de har blitt mobbet siste tre måneder, og som har brutto årlige driftsutgifter til grunnskoleutdanning per elev på kr 110 000.

(c) Vi ser at regresjonskoeffisienten til `driftsutgifter` i Analyse 1 er signifikant på 5% nivå. Sett opp nullhypotesen, og gjennomfør denne testen ved å bruke informasjon fra utskriften.

(d) Drøft kort diagnoseplottene til Analyse 1, som er gitt i Figur 1.

(e) Kommenter kort hva vi lærer av analysene 3, 4, 5 og 6 (du trenger ikke gi detaljerte fortolkninger av de enkelte koeffisientene, bare gi en kort overordnet vurdering).

(f) La Z_{ij} være testresultatet til elev nr. i i kommune nr. j, og la n_j være antall elever på 5. trinn i kommunen. La videre X_j være en egenskap ved kommune j, som vi bruker i følgende lineære regresjonsmodell:

$$Z_{ij} = \alpha + \beta X_j + \epsilon_{ij}, \qquad (*)$$

der residualene ϵ_{ij} er uavhengige og normalfordelte med forventning 0 og varians σ^2. La Y_j være gjennomsnittsresultatet til kommune j.

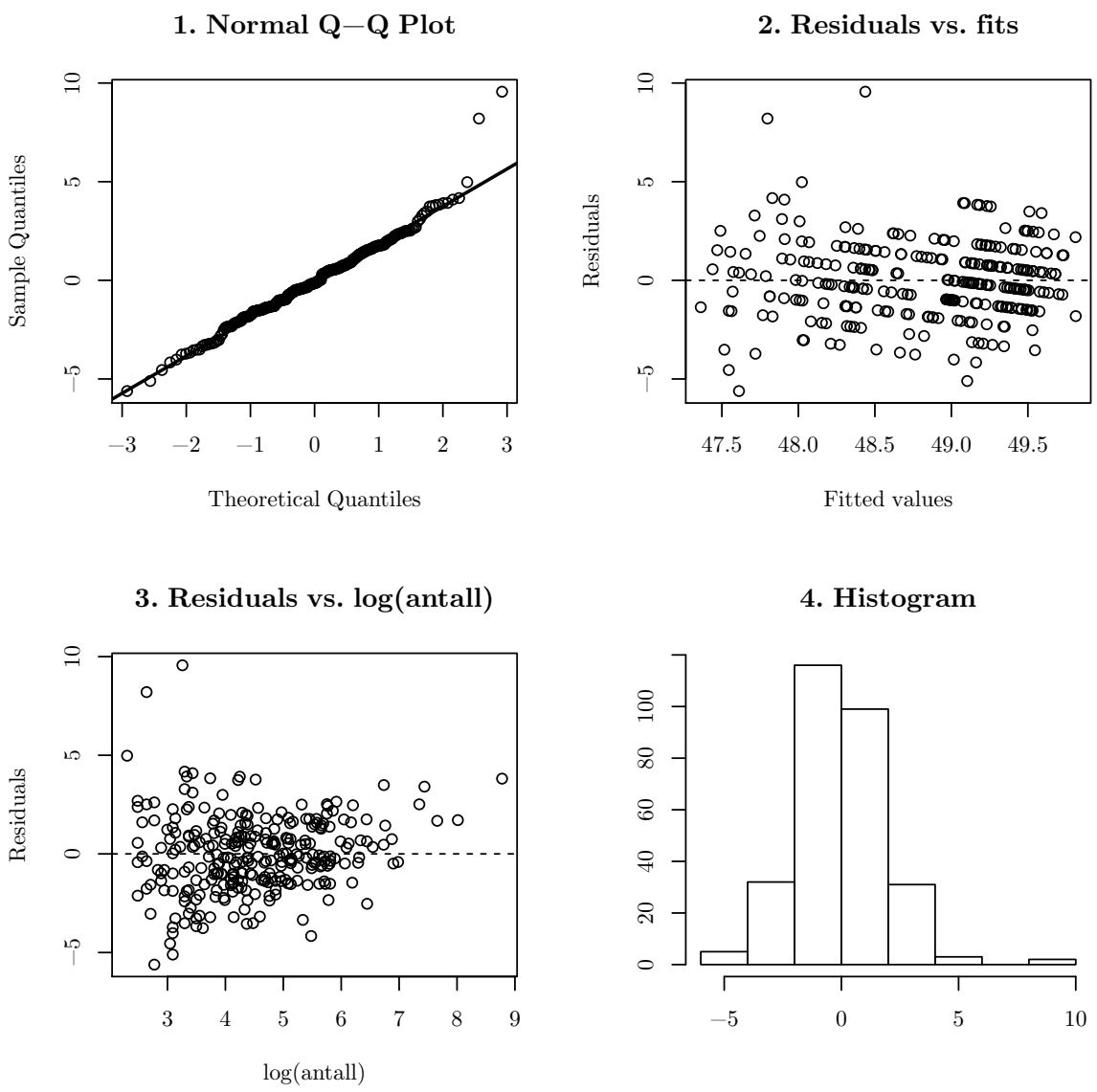
Lag en ny regresjonsmodell ved å ta gjennomsnittet over alle elevene i en kommune i modellen over $(*)$ slik at du får Y_j på venstre side, og vis formelt at feilleddet til denne regresjonsmodellen er heteroskedastisk. Hvordan kan dette resultatet hjelpe oss til å forstå diagnoseplott nr. 3 i Figur 1?

En journalist i en riksdekkende tabloidavis har fattet interesse for vår Analyse 2, og ønsker å lage en sak om at skoler gir fritak fra nasjonale prøver til svake elever i den hensikt å oppnå bedre resultater.

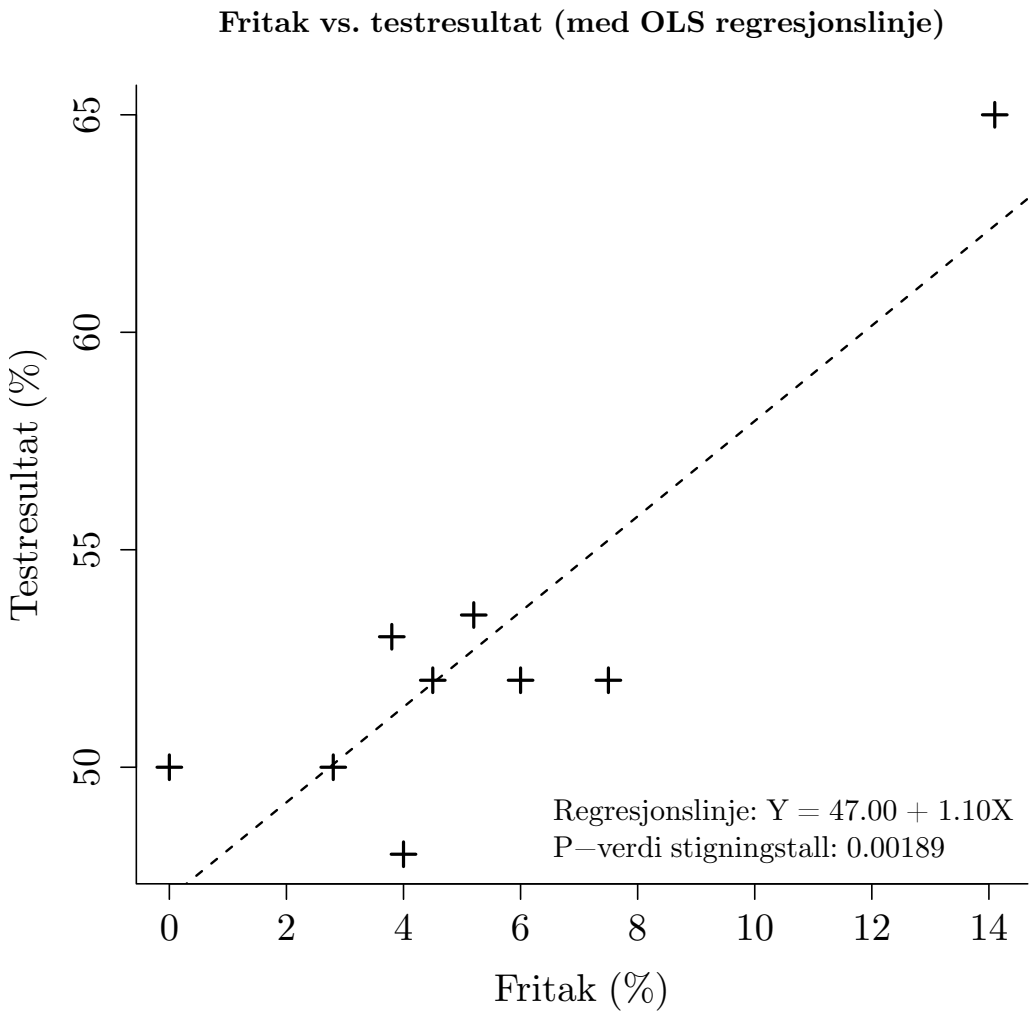
(g) Kommenter kort om vår Analyse 2 kan understøtte en slik påstand.

Journalisten har samlet inn egne data fra ni grunnskoler i Oslo, der fritaksgraden (i prosent) er registrert sammen med testresultatet i lesing på 5. trinn. Hun har fått gjennomført en regresjon med fritaksgrad som forklaringsvariabel og testresultat som responsvariabel i en statistisk programvarepakke. Figur 2 viser den estimerte regresjonslinjen, sammen med formel for linjen og p-verdi for det estimerte stigningstallet.

(h) Hva er fortolkningen av regresjonskoefisienten til `fritak`? Kan du ved hjelp av Figur 2 peke på et potensielt problem med analysen av denne regresjonen og hvordan journalisten bør gå frem for å løse det?



Figur 1: Diagnoseplott til Analyse (1)



Figur 2: Fritaksgrad vs. testresultat for ni grunnskoler i Oslo, sammen med regresjonslinjen. Formelen for regresjonslinjen og p-verdi for stigningstallet kan leses av i nederste høyre hjørne.