BOKMÅL

NHH

HJEMMEEKSAMEN MET4

Høst 2020

Dato: 19. november 2020

Tidsrom: 09:00 - 13:00

Antall timer: 4

BESVARELSEN SKAL LEVERES I WISEFLOW

På våre nettsider finner du informasjon om hvordan du leverer din besvarelse: `https://www.nhh.no/for-studenter/eksamen/innlevering-individuelt-og-i-gruppe/`

Kandidatnummer blir oppgitt på StudentWeb i god tid før innlevering. Kandidatnummer skal være påført på alle sider øverst i høyre hjørne (ikke navn eller studentnummer). Ved gruppeinnlevering skal alle gruppemedlemmers kandidatnummer påføres.

Samarbeid mellom individer eller grupper om utarbeidelse er ikke tillatt, og utveksling av egenprodusert materiale til andre individer eller grupper skal ikke forekomme. En besvarelse skal bestå av individets, eller gruppens egne vurderinger og analyse. All kommunikasjon under hjemmeeksamen er å anse som fusk. Alle innleverte oppgaver blir behandlet i Urkund, NHHs datasystem for tekst- og plagiatkontroll

UTFYLLENDE BESTEMMELSER OM EKSAMEN

`https://www.nhh.no/globalassets/for-studenter/forskrifter/utfyllende-bestemmelser-til-forskrift-om-fulltidsstudiene-ved-nhh.pdf`

Antall sider, inkludert forside og vedlegg: 8

Antall vedlegg: 3 (Alle vedlegg følger etter oppgavene)

Oppgave 1

Mythbusters er et klassisk program på Discovery Channel, som går ut på å undersøke om ulike myter er sanne eller ikke. I en episode (sesong 3, episode 4) ønsket programlederne å teste om gjesping er smittsomt ved å sette opp følgende eksperiment: 50 personer ble intervjuet under påskudd av at de skulle rekruttere nye medhjelpere til programmet. Intervjuobjektene ble fordelt tilfeldig i to grupper, der personen som gjennomførte intervjuet gjespet tydelig under intervjuene i den første gruppen, men gjespet ikke i det hele tatt under intervjuene i den andre gruppen. Det ble så notert ned om intervjuobjektet gjespet i løpet av intervjuet. Resultatet ble som følger:

- Gruppe 1 (Intervjuer gjespet): **10 av 34 (29.4%) deltakere gjespet**.
- Gruppe 2 (Intervjuer gjespet ikke): **4 av 16 (25.0%) deltakere gjespet**.

Programleder Jamie Hyneman slo fast at forskjellen på 4.4 prosentpoeng mellom de to gruppene var statistisk signifikant, og at vi dermed kan slå fast at gjesping er smittsomt.

(a) Gjennomfør en hypotesetest for å kontrollere påstanden. Har programlederen rett?

Da denne oppgaven ble gitt på eksamen ved en annen institusjon, svarte en student på følgende måte:

Besvarelse 1

Dette er en test for sammenligning av to andeler. La p_1 og p_2 være de sanne andelene som vil gjespe i løpet av intervjuet i de to gruppene, og la $\widehat{p}_1 = 0.25$ og $\widehat{p}_2 = 0.294$ være de to observerte andelene. Den estimerte andelen som gjesper under nullhypotesen om ingen forskjell mellom gruppene er $\widehat{p} = (4 + 10)/(16 + 34) = 0.28$. Vi ønsker å teste

$$\mathrm{H}_0 : p_1 = p_2 \qquad \text{mot} \qquad \mathrm{H}_1 : p_1 \neq p_2$$

Testobservatoren er gitt ved

$$Z = \frac{\widehat{p}_2 - \widehat{p}_1}{\sqrt{\widehat{p}(1-\widehat{p})}} = \frac{0.294 - 0.25}{\sqrt{0.28(1-0.28)}} = 0.09.$$

Testobservatoren er tilnærmet normalfordelt under nullhypotesen, så kritisk verdi for tosidig test ved 5% signifikansnivå er 1.96. Observert verdi av testobservatoren er 0.09, så vi forkaster ikke nullhypotesen.

(b) Hvordan vil du vurdert dette svaret dersom du var sensor? Begrunn svaret ditt.

En annen student svarte dette:

Besvarelse 2

Vi skal teste for likhet mellom to andeler.

$$\widehat{p} = \frac{4+10}{16+34} = 0.28.$$

Testobservatoren er

$$Z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\widehat{p}(1-\widehat{p})}} = \frac{0.25 - 0.294}{\sqrt{\left(\frac{1}{34} + \frac{1}{16}\right)0.28(1-0.28)}} = -0.32.$$

Siden $-0.32 < 1.96$ er nullhypotesen om ingen forskjell bevist.

(c) Hvordan vil du vurdert dette svaret dersom du var sensor? Vurder de to besvarelsene mot hverandre.

Oppgave 2

Asylpolitikk er alltid et aktuelt tema i den offentlige debatten. En diskusjon i den sammenhengen er hvorvidt et land kan regulere antallet søknader om asyl ved å justere landets asylpolitikk. For eksempel kan man tenke seg at en strengere innvandringspolitikk vil føre til at potensielle asylsøkere heller velger å søke seg til andre land, og motsatt; at mindre streng politikk vil ha en tiltrekkende kraft på potensielle asylsøkere.

En artikkel fra 2016 i det prestisjetunge tidsskriftet *American Economic Review*[1] undersøker denne problemstillingen ved hjelp av regresjonsanalyse. Vi har data fra 2012 på antall asyslsøknader fra 49 avreiseland til 19 ankomstland, der vi har en observasjon av antall asylsøkere for hvert *par* av avreise- og ankomstland.

For å gjøre disse størrelsene sammenlignbare er antallet søknader normalisert med befolkningsstørrelsen i avreiselandet. Vi bruker logaritmen av denne andelen som responsvariabel. Vi nummererer hver land-par med indeksen i, så responsvariabelen i analysene vi skal se på i denne oppgaven er dermed $\log(Y_i)$, der Y_i er andelen av befolkningen i avreiselandet som søker asyl i ankomstlandet for land-par i dette året. Det vil si at vi kan skrive $Y_i = U_i/A_i$, der U_i er antall asylsøkere og A_i er antall innbyggere i avreiselandet.

Vi har flere kontrollvariabler i datasettet vårt. Alle variablene er presentert i Tabell 1 under.

(a) Bruk regresjonsutskriften i Vedlegg 1 til å vurdere påstanden om at en strengere asylpolitikk fører til færre asylsøkere. Hvor stor er den eventuelle effekten?

[1] Timothy J. Hatton: *Refugees, Asylum Seekers, and policy in OECD Countries.* American Economic Review (2016) Vol. 105, No. 5, pp. 441–45.

Variabel	Beskrivelse
`lnapps`	$\log(Y_i)$ der Y_i er andelen av befolkningen i avreiselandet som søker asyl i ankomstlandet.
`bdbest`	Uppsala-indeksen for antall drepte i krig i avreiselandet.
`fhcl`	Freedom House-indeksen for borgerrettigheter i avreiselandet.
`fhpr`	Freedom House-indeksen for politiske rettigheter i avreiselandet.
`lndist`	Logaritmen til avstanden fra avreise- til ankomstlandet.
`lngdpdest`	Logaritmen til bruttonasjonalprodukt per capita i ankomstlandet.
`lngdpsource`	Logaritmen til bruttonasjonalprodukt per capita i avreiselandet.
`lnsttot`	Logaritmen til antall innvandrere fra avreiselandet som allerede er bosatt i ankomstlandet.
`pt`	*Political terror scale*, et mål på nivået av menneskerettighetsbrudd i avreiselandet.
`unp`	Arbeidsledigheten i ankomstlandet.
`poltot`	Indeks som øker med hvor *streng* asylpolitikk ankomstlandet fører.

Tabell 1: Beskrivelse av variabler i asyldatasettet. Alle tall er fra 2012 bortsett fra `lnsttot` som er fra 2000/2001.

(b) Bruk residualplottene i Vedlegg 2 til å vurdere om forutsetningene for vanlig lineær regresjon (OLS) er oppfylt.

Per 2020 er verdien av forklaringsvariabelen `poltot` for Norge lik 3.5. Du får videre oppgitt at verdiene av resten av kontollvariablene (bortsett fra `poltot`) for land-paret (Norge, Syria) per 2020 er gitt slik at

$$\widehat{\beta}_0 + \widehat{\beta}_1\texttt{bdbest} + \widehat{\beta}_2\texttt{fhcl} + \widehat{\beta}_3\texttt{fhpr} + \widehat{\beta}_4\texttt{lndist} + \widehat{\beta}_5\texttt{lngdpdest} + \widehat{\beta}_6\texttt{lngdpsource} + \widehat{\beta}_7\texttt{lnsttot} + \widehat{\beta}_8\texttt{pt} + \widehat{\beta}_9\texttt{unp} = -10.3,$$

der koeffisientestimatene $\widehat{\beta}_0, \ldots, \widehat{\beta}_9$ er tilhørende koeffisientestimater hentet fra regresjonsutskriften i Vedlegg 1.

(c) Bruk informasjonen over til å predikere verdien av `lnapps` for land-paret (Norge, Syria) i 2020.

Befolkningen i Syria er 17.7 millioner per 2020.

(d) Prediker antall asylsøkere fra Syria til Norge i 2020 hvis vi legger til grunn regresjonsmodellen i Vedlegg 1.

Det originale datasettet er et paneldatasett med årlige observasjoner av alle variablene i perioden 1997 - 2012. En alternativ modellspesifikasjon er da

$$\log(Y_{it}) = \alpha_i + v_t + \beta_1\texttt{bdbest}_{it} + \beta_2\texttt{fhcl}_{it} + \beta_3\texttt{fhpr}_{it} + \beta_4\texttt{lndist}_{it} + \beta_5\texttt{lngdpdest}_{it} + \beta_6\texttt{lngdpsource}_{it} + \beta_7\texttt{lnsttot}_{it} + \beta_8\texttt{pt}_{it} + \beta_9\texttt{unp}_{it} + \beta_{10}\texttt{poltot}_{it} + \epsilon_{it},$$

hvor i og t nå angir henholdsvis land-par og årstall. Her betraktes α_i som faste effekter, mens v_t er dummyvariabler for hvert årstall. Den estimerte modellen er vist i Vedlegg 3.

(e) Hva representerer α_i og v_t i *denne spesifikke konteksten?*

Norge er ankomstland for alle landpar i hvor $i = 1, 2, ..., m$. For nettopp disse land-parene i 2012 er innbyggertallene i avreiselandene og kontrollvariablene (utenom `poltot`) slik at

$$\sum_{i=1}^{m} A_{i,2012} \exp(\widehat{\eta}_{i,2012}) = 13456.83,$$

hvor $A_{i,2012}$ var antall innbyggere i avreiselandet for land-par i i 2012, og hvor

$$\begin{aligned}\widehat{\eta}_{i,2012} =& \widehat{\alpha}_i + \widehat{v}_{2012} + \widehat{\beta}_1 \texttt{bdbest}_{i,2012} + \widehat{\beta}_2 \texttt{fhcl}_{i,2012} + \widehat{\beta}_3 \texttt{fhpr}_{i,2012} + \widehat{\beta}_4 \texttt{lndist}_{i,2012} \\ &+ \widehat{\beta}_5 \texttt{lngdpdest}_{i,2012} + \widehat{\beta}_6 \texttt{lngdpsource}_{i,2012} + \widehat{\beta}_7 \texttt{lnsttot}_{i,2012} + \widehat{\beta}_8 \texttt{pt}_{i,2012} + \widehat{\beta}_9 \texttt{unp}_{i,2012},\end{aligned}$$

der alle koeffisentestimatene kommer fra modellen i Vedlegg 3.

(f) Et politisk parti foreslo i 2012 å stramme inn asylpolitikken på en måte som tilsvarer å øke verdien av `poltot` fra 3.5 til 5. Bruk informasjonen over til å predikere hvor mange færre asylsøkere Norge ville fått i 2012 dersom forslaget ble vedtatt.

(g) Drøft kort eventuelle etiske problemstillinger knyttet til utregninger som den i oppgave (f) (Maks 100 ord).

Oppgave 3

Vi har følgende datasett med seks observasjoner bestående av en binær responsvariabel y og to forklaringsvariabler x_1 og x_2:

Tabell 2: Datasett

y	x1	x2
0	3	4
1	4	5
1	5	3
0	3	6
0	4	3
1	6	2

Du observerer så forklaringsvariablene $(x_1, x_2) = (3, 3)$ for et nytt individ.

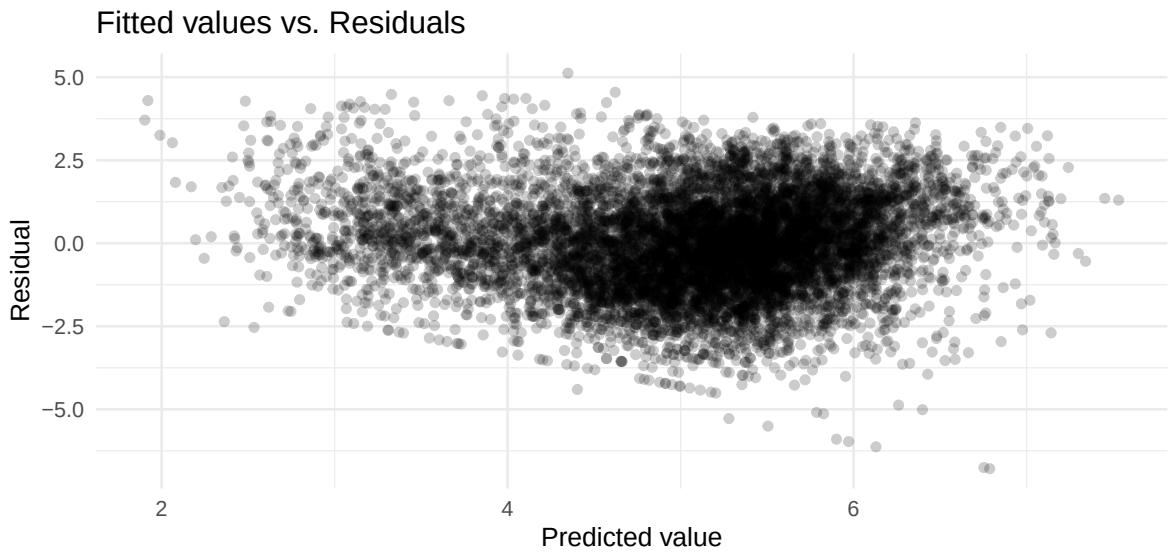
(a) Regn ut hva klassifiseringen av y blir for det nye individet ved å bruke k-nearest neighbor (KNN), med $k = 3$.

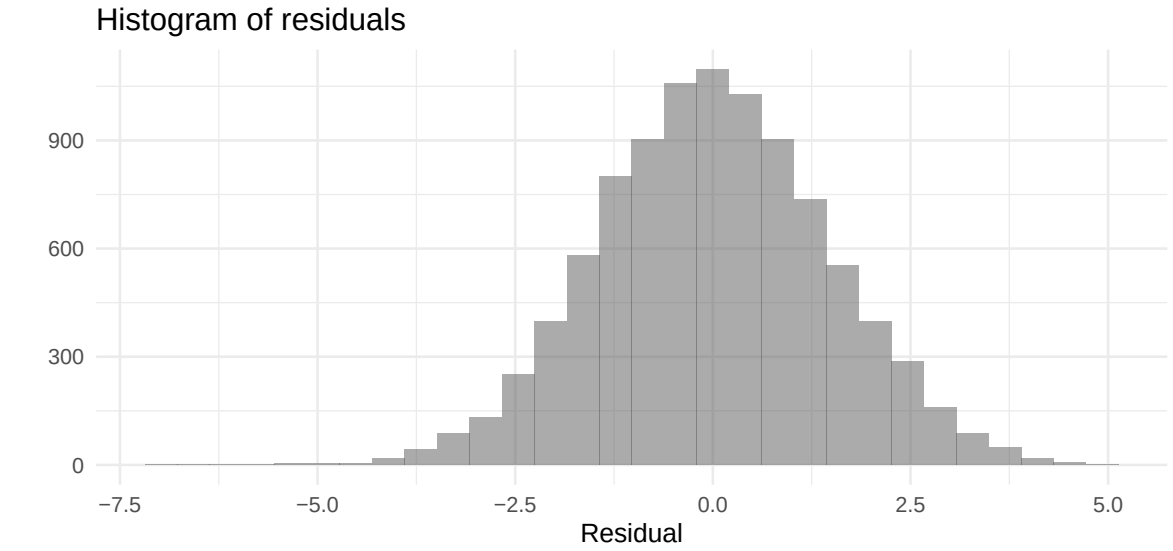
(b) Hvordan fungerer KNN når $k = n$, hvor n er antall observajoner i datasettet? Hva vil skje dersom $k = 6$ for dette datasettet?

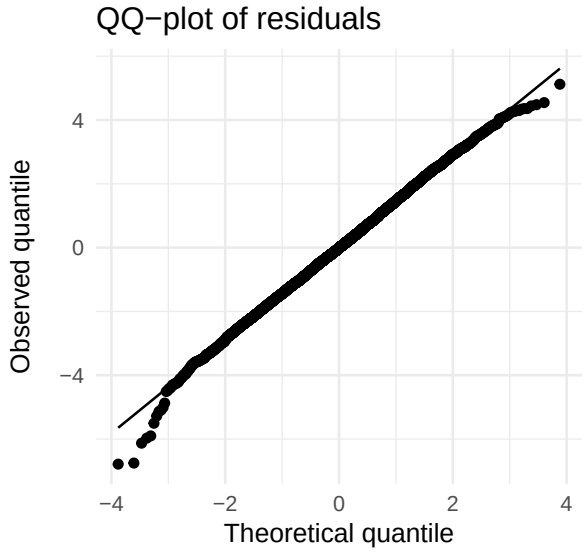
Vedlegg 1: Regresjonsutskrift

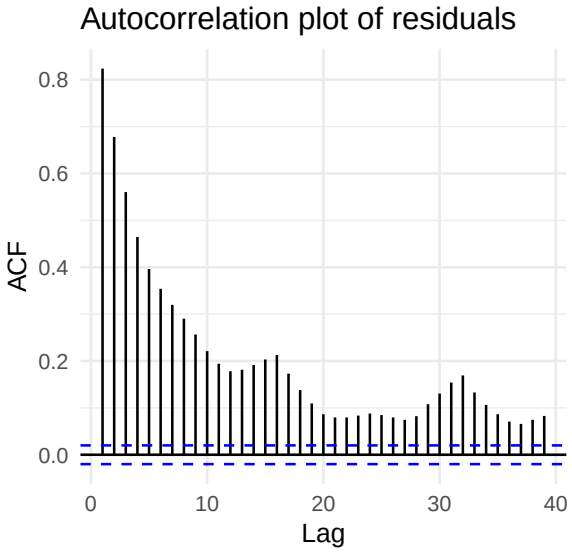
```
Call:
lm(formula = lnapps ~ pt + fhcl + fhpr + bdbest + lngdpsource +
    lnsttot + lndist + lngdpdest + unp + poltot, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2587 -0.9182 -0.0898  0.8323  4.2495

Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept) -2.75613064  3.50941361  -0.785   0.4326
pt           0.13786415  0.09198360   1.499   0.1345
fhcl         0.64786337  0.13559508   4.778 2.25e-06 ***
fhpr        -0.34547444  0.08776811  -3.936 9.28e-05 ***
bdbest       0.00005508  0.00002415   2.281   0.0229 *
lngdpsource -0.11462954  0.07363955  -1.557   0.1201
lnsttot      0.23142907  0.01959886  11.808  < 2e-16 ***
lndist      -0.62837657  0.09710837  -6.471 2.07e-10 ***
lngdpdest    0.64447185  0.32058186   2.010   0.0449 *
unp         -0.06167373  0.01522865  -4.050 5.82e-05 ***
poltot      -0.05396284  0.01649892  -1.998   0.0462 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.4 on 580 degrees of freedom
  (42 observations deleted due to missingness)
Multiple R-squared:  0.3488,    Adjusted R-squared:  0.3376
F-statistic: 31.07 on 10 and 580 DF,  p-value: < 2.2e-16
```

Vedlegg 2: Residualplott









Vedlegg 3: Regresjonsutskrift for paneldata

```
Oneway (individual) effect Within Model

Unbalanced Panel: n = 626, T = 6-16, N = 9610

Coefficients:
              Estimate Std. Error t-value Pr(>|t|)
year1998     0.2703738  0.0547938    4.93  8.2e-07 ***
year1999     0.5098071  0.0567478    8.98  < 2e-16 ***
year2000     0.7632323  0.0615734   12.40  < 2e-16 ***
year2001     0.9365277  0.0658655   14.22  < 2e-16 ***
year2002     1.0517368  0.0712065   14.77  < 2e-16 ***
year2003     1.0595820  0.0756011   14.02  < 2e-16 ***
year2004     0.8480908  0.0837786   10.12  < 2e-16 ***
year2005     0.7395502  0.0918983    8.05  9.5e-16 ***
year2006     0.7126442  0.1035050    6.89  6.2e-12 ***
year2007     0.6021288  0.1120705    5.37  7.9e-08 ***
year2008     0.6939668  0.1180102    5.88  4.2e-09 ***
year2009     0.8328582  0.1165391    7.15  9.6e-13 ***
year2010     0.8603227  0.1222780    7.04  2.1e-12 ***
year2011     0.9105812  0.1296684    7.02  2.3e-12 ***
year2012     1.0803078  0.1348168    8.01  1.3e-15 ***
pt           0.2210024  0.0196214   11.26  < 2e-16 ***
fhcl         0.2894999  0.0221929   13.04  < 2e-16 ***
fhpr        -0.0503066  0.0182382   -2.76   0.0058 **
bdbest       0.0000103  0.0000050    2.07   0.0387 *
lngdpsource -0.5334016  0.0676613   -7.88  3.6e-15 ***
lngdpdest    0.0660557  0.1989760    0.33   0.7399
unp         -0.0238331  0.0057291   -4.16  3.2e-05 ***
poltot      -0.0463667  0.0058515   -7.92  2.6e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    8140
Residual Sum of Squares: 7140
R-Squared:      0.123
Adj. R-Squared: 0.0599
F-statistic: 54.7712 on 23 and 8961 DF, p-value: <2e-16
```