OBLIG2Sok2009_36

36

a) Kjør en enkel lineær regresjonsanalyse. Velg avhengig og uavhengig variabel selv, og forklar hva du ønsker/har mulighet å finne ut av ved bruk av disse variablene.

Jeg ønsker å undersøke hvordan utdanningsnivå til foreldre påvirker den neste generasjonen. Jeg kan undersøke dette ved å først se på sammenhengen mellom mødres utdanning og døtrenes utdanningslengde. Jeg velger datterens utdanningslengde som den avhengige variabelen, og morens utdanningsnivå som den uavhengige variabelen. Jeg ønsker å forstå døtrenes utdanningslengde, og tror at mødres utdanning har en effekt på den avhengige variabelen.

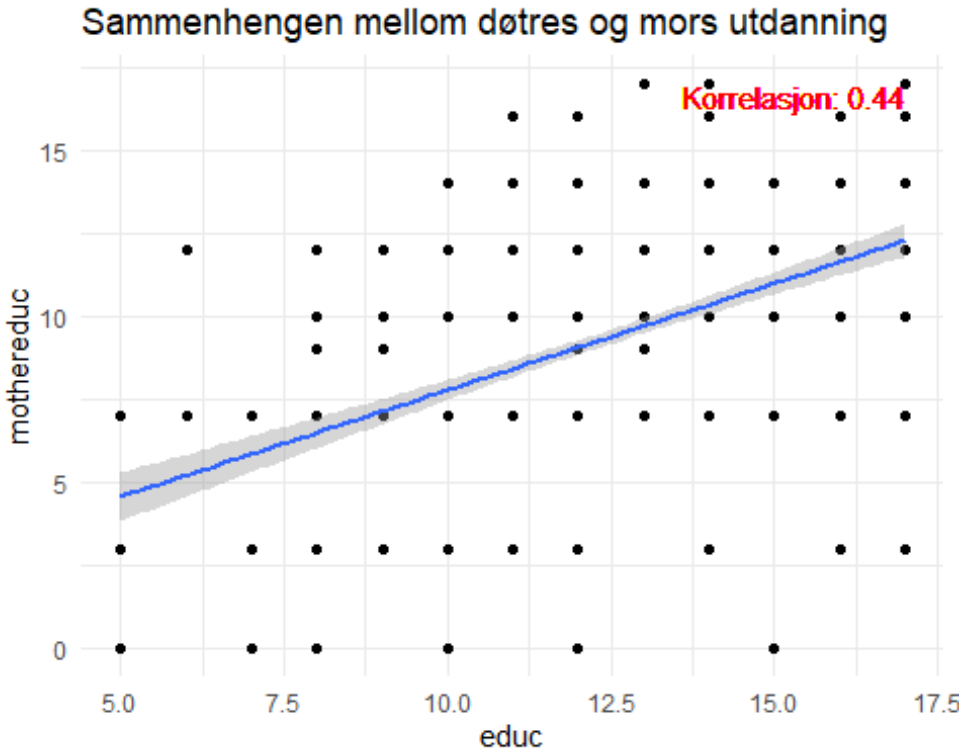
b) Vis og forklar resultatene dine. Bruk grafer, tabeller, og output til å forklare din modell, og ta med alle detaljer og statistikker som er relevante for din analyse.

```
Call:
lm(formula = mroz$educ ~ mroz$mothereduc)

Residuals:
    Min      1Q  Median      3Q     Max
-7.0972 -1.0972  0.3767  0.9028  6.5558

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.55981    0.21898   43.66   <2e-16 ***
mroz$mothereduc  0.29478    0.02224   13.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.054 on 751 degrees of freedom
Multiple R-squared:  0.1895,    Adjusted R-squared:  0.1884
F-statistic: 175.6 on 1 and 751 DF,  p-value: < 2.2e-16
```

Observerer at det er en positiv sammenheng mellom morens utdanning og døtres utdanningslengde ved stigningstallet. Vi kan også bruke stigningstallet til å predikere utdanningslengde fra den uavhengige variabelen. Stigningstallet (koeffisienten) i denne modellen er 0.29478, som betyr at for hver ekstra enhet, eller i dette tilfellet år av morens utdanning, øker datterens forventede utdanning med ca. 0.3 år - alt annet likt. Med andre ord, jo høyere utdanningsnivå moren har, desto høyere er forventet utdanningsnivå for datteren. Standardfeilen er 0.02224, dette er relativt lavt som betyr at man kan forvente et koeffisienten vil variere lite fra et utvalg til et annet . Standardfeilen er også lav i forhold til den forventede effekten, dette styrker hypotesen min.

Det er en veldig lav p - verdi, langt lavere enn det vanlige terskelnivået på 0.05. En så lav p - verdi indikerer at det er ekstremt usannsynlig at man vil observerer det sterke forholdet som kommer av modellen dersom det faktisk ikke er noen sammenheng mellom variablene. Dette betyr ikke at effekten er stor eller viktig men at den er statistisk signifikant. Vi har også en høy t - verdi, som er 13.25, dette betyr at estimatet av koeffisienten ligger 13.25 standardfeil unna null. En lav p - verdi med en høy t - verdi gir gode bevis på at mødres utdanning har effekt på døtres utdanning.

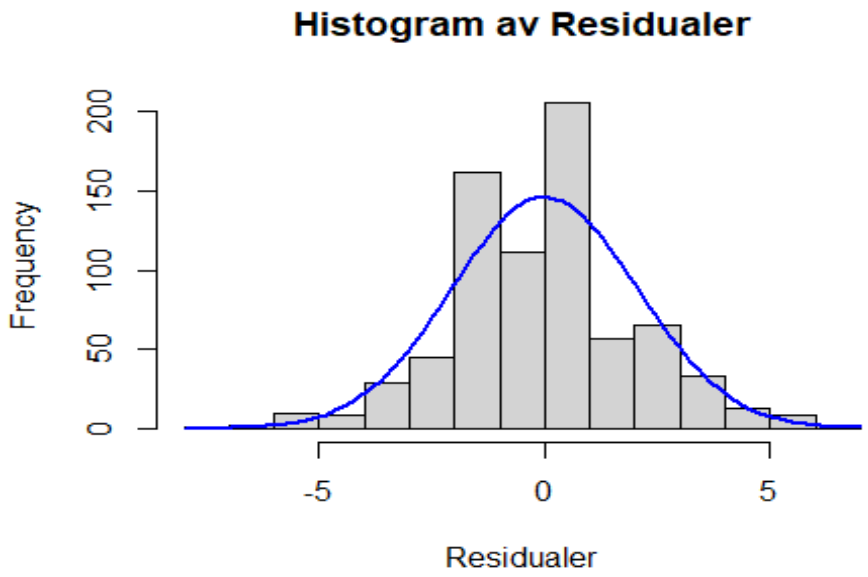
Standardfeilen til den predikerte verdien er 2.054, som betyr at på gjennomsnitttet er avviket mellom observerte verdier og predikerte verdier 2 år. Variansen er 0.1895, som betyr at omtrent 18.95% av variansen i døtres utdanningslengde kan forklares ved mødres utdanningsnivå alene. Samlet sett viser modellen at den uavhengige variabelen mødres utdanning har en signifikant positiv sammenheng med døtres utdanningslengde. Når mødres utdanning øker, forventes det at døtrenes utdanningslengde også øker.



Figuren over viser sammenhengen mellom døtrenes utdanningslengde og mødres utdanning. Observerer en korrelasjon på 0.44 (Pearsons r), som er en moderat signifikant korrelasjon.

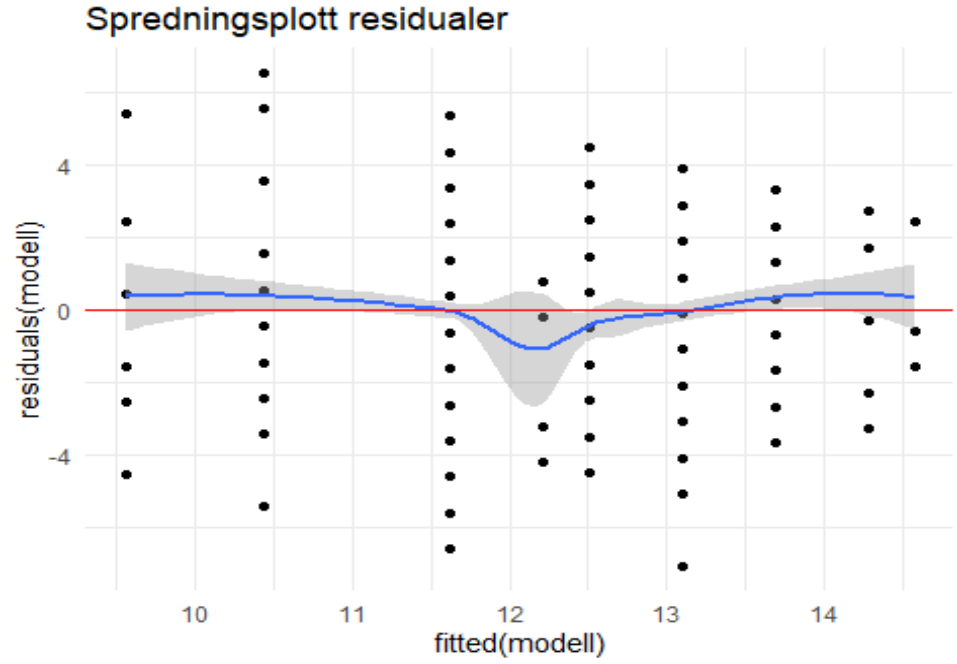
c) Undersøk hvorvidt modellen din bryter med antakelsene til lineær regresjon. Hvis ja, hva er konsekvensen av de eventuelle bruddene? Vis og forklar hvordan du testet/undersøkte.

Det er flere metoder som kan anvendes for å undersøke hvorvidt modellen bryter med antakelsene til lineær regresjon. Jeg vil undersøke residual plott. Residualer er forskjellene mellom de observerte verdiene av den avhengige variabelen (de faktiske dataene) og de av modellen - forventede verdier (de estimerte dataene).

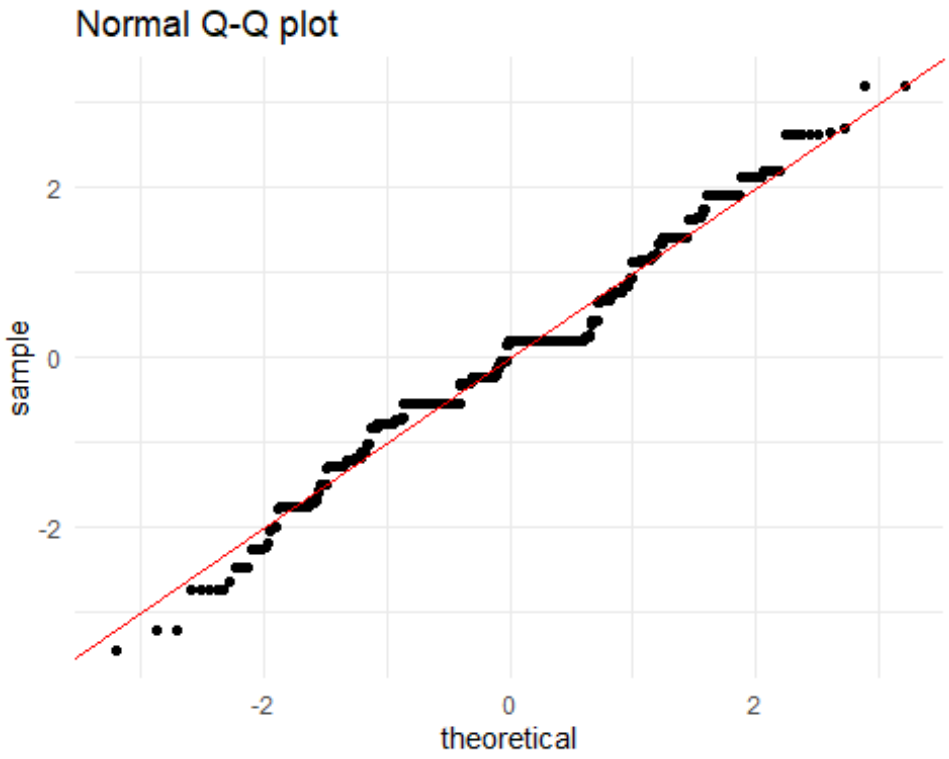


Over et det tegnet et histogram for residualene. Observerer at de fleste residualer er rundt null, men uten et symmetrisk mønster som antyder mulige avvik fra normalfordeling. Det er viktig at residualene er normalfordelte for å kunne estimere korrekte parametere.

Ved å konstruere et spredningsplott med de estimerte verdiene på x-aksen og residualene på y-aksen kan vi undersøke nærmere. Observerer at verdiene ligger ganske flatt, og ikke følger noen systematisk trend, dette gir visuelle indikasjoner på at residualene kan være normalfordelt. Observerer avvik mot midten av plottet, i likhet med histogrammet over. Basert på disse figurene kan vi ikke helt sikkert anslå om dataen er normalfordelt.



Til sist gir et kvantil-kvantil Q-Q plot en mer detaljert visualisering av om dataene følger en normalfordeling. X – aksen viser de teoretiske kvantilene, hvor residualene er i en standard normalfordeling om variansen er kjent. Y – aksen viser kvantilene til dataten. Dersom dataene følger en normalfordeling vil punktene følge den rette linjen. Observerer at punktene i Q-Q plottet ligger nær den rette linjen, uten store avvik, støtter antagelsen om normalfordeling, og indikerer at de observerte og de forventede fordelingene ligner nær hverandre



Samlet sett gir analysen blandete signaler om hvorvidt dataen er normalfordelt. Histogrammet viser at residualene er samlet rundt null, men mangler et symmetrisk mønster som er karakteristisk for normalformaldeling. Dette antyder mulige avvik. Spredningsplottet gir antydning til normalfordeling, da punktene konsentreres rundt null, men viser likevel avvik rundt midten. Dette er konsistent med histogrammet. Q-Q plottet derimot, tyder i større grad at dataene er normfordelt ettersom punktene ligger nær den rette linjen som presenterer den forventede fordelingen. Imidlertid, ettersom det er en viss inkonsistens (tilstede, men små) i resultatene burde man ikke se helt bort fra de avvikende observasjonene.

a) Kjør en multippel lineær regresjonsanalyse med minst to uavhengige variabler. Velg selv om du tilføyer en eller flere variabler til din tidligere analyse, eller om du lager en helt ny. Forklar hvorfor du har valgt denne kombinasjonen av variabler.

I denne oppgaven vil jeg å legge til fedres utdanning. Dette kan gi et mer helhetlig bilde av hvordan familien påvirker døtres utdanning. Ved å se på begge foreldre kan jeg undersøke den kombinerte effekten, noe som kan avdekke kjønnsrollemønster og kontrollere for

varians eller skjevheter som kan oppstå hvis man bare ser på en av foreldrene - det kan gi mer presise estimater.

b) Vis og forklar resultatene dine. Bruk grafer, tabeller, og output til å forklare din modell og hva modellen kan fortelle oss.

```
Call:
lm(formula = mroz$educ ~ mroz$mothereduc + mroz$fathereduc)

Residuals:
    Min      1Q  Median      3Q     Max
-7.4596 -1.3760 -0.0924  0.7241  6.9243

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      8.97566    0.22567  39.774  < 2e-16 ***
mroz$mothereduc  0.18328    0.02622   6.991 6.04e-12 ***
mroz$fathereduc  0.18342    0.02471   7.422 3.14e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.984 on 750 degrees of freedom
Multiple R-squared:  0.245, Adjusted R-squared:  0.243
F-statistic: 121.7 on 2 and 750 DF,  p-value: < 2.2e-16
```

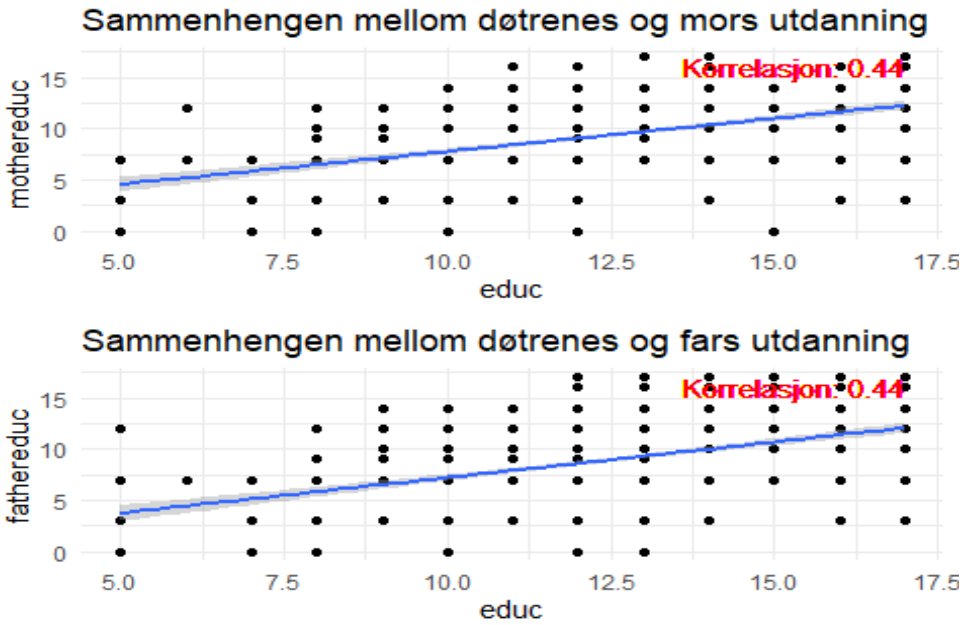
Figuren på neste side viser sammenhengen mellom den avhengige variabelen og de to uavhengige variablene. Legger merke til at det er en lik korrelasjon (Pearsons r).

Stigningstallet (koeffisienten) til den nye uavhengige variabelen er tilnærmet lik mødres utdanning, som kan indikere at kjønn ikke spiller en betydelig rolle når det gjelder utdanningslengde. Observerer at variansen i større grad kan forklares ved en multippel regresjon, den er 0.243, som betyr at 24.3% av variansen kan forklares av modellen.

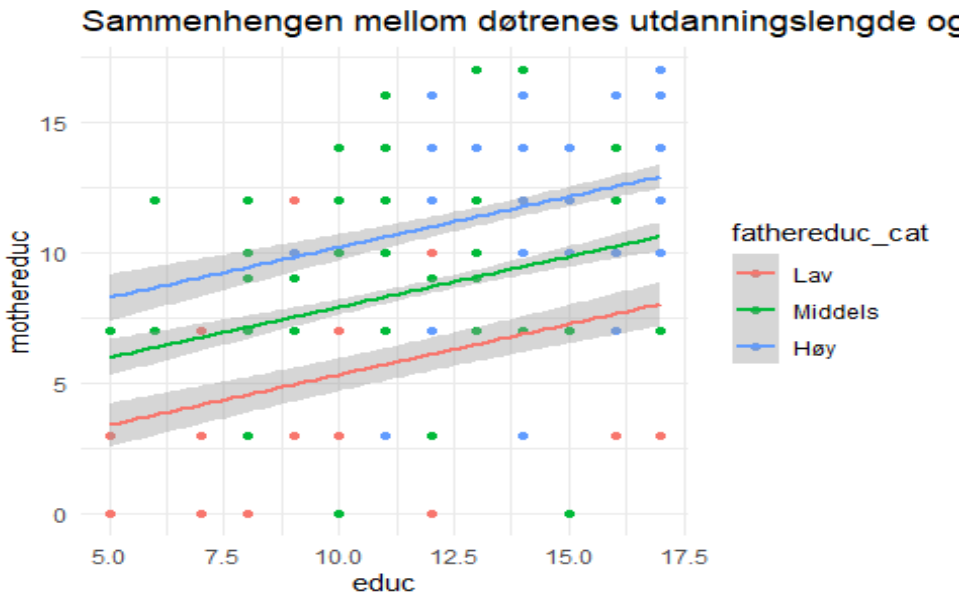
Intercept er på 8.9, som betyr at når foreldrene har null utdanning så har døtrene ca. 9 år utdanning. Dette tallet er litt lavere enn det so ble observert med bare en uavhengig variabel, som var 9.55.

Det er også verdt å nevne at de uavhengige variablene korrelerer. Ved bruk av en Pearsons R test er korrelasjonen estimert til 0.57, som indikerer en moderat til sterk positiv sammenheng. Dette betyr at det er en tendens til at jo høyere utdanning den ene forelderen, desto høyere er også utdanningen til den andre forelderen.

Standardfeilen til koeffisienten er lav i forhold til den forventede effekten, som styrker hypotesen min. Det er en høy t -verdi, og en veldig lav p - verdi. Dette indikerer statistisk signifikans. Standardfeilen til den predikerte verdien er 1.984, som betyr at på gjennomsnittet er avviket mellom observerte verdier og predikerte verdier er tilnærmet 2 år.



I figuren under har jeg kategorisert fars utdanning i tre, fedre med høy utdanning, middels og en med lav. Observerer at stigning er tilnærmet lik, slik at effekten er den samme for de tre kategoriene. Standardavviket er noe forskjellig, observerer at standardavviket er størst for fedre med lav utdanning, som indikerer at punktene avviker mer fra gjennomsnittet, dette kan være interessant å se nærmere på. Alle kategoriene viser en positiv sammenheng, som betyr at jo mer utdanning foreldrene har, desto høyere er forventet utdanning for døtrene.

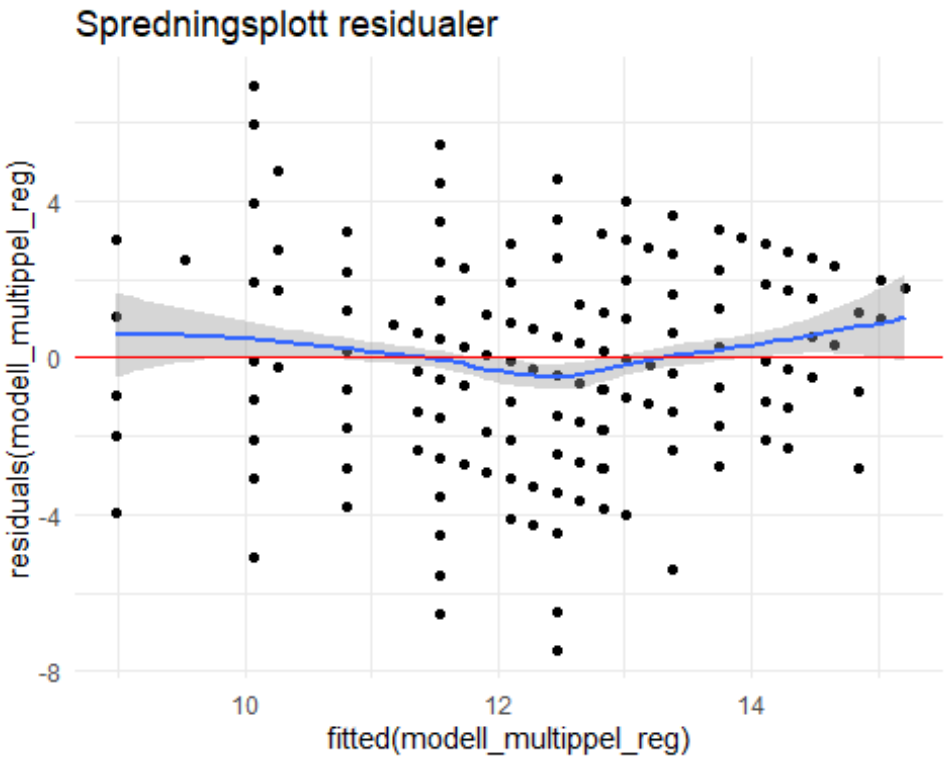


c) Test hvorvidt modellen din bryter med antakelsene til multippel lineær regresjon. Vis og forklar hvordan du testet/undersøkte

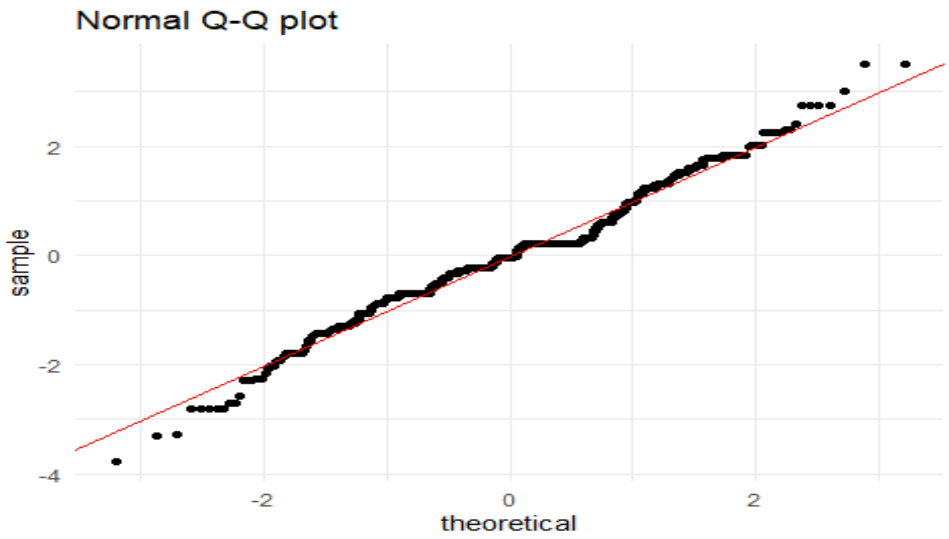
Bruker en Shapiro - Wilk test for normal fordeling for modellen med en uavhengig variabel og to uavhengige variabeler . Shapiro - Wilk test fungerer bra for relativt små samples, som er tilfellet i denne analysen. Observerer at begge p - verdiene er mye lavere enn 0.05, noe som tilsier at dataene ikke er normalfordelt

```
   Shapiro-Wilk normality test

data:  residualer
W = 0.97183, p-value = 7.264e-11


    Shapiro-Wilk normality test

data:  residualer_multippel
W = 0.97937, p-value = 8.123e-09
```

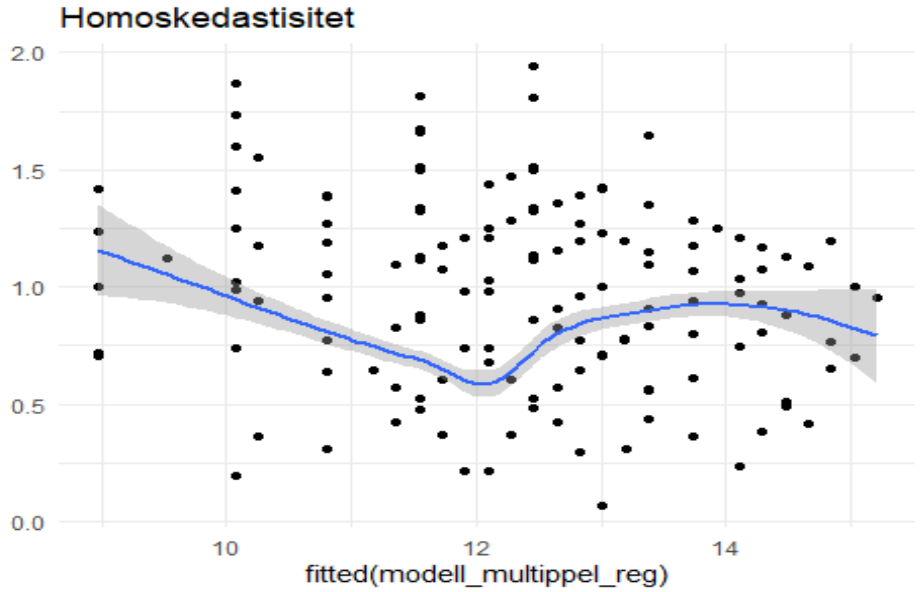
Ved å konstruere et spredningsplott med estimerte verdiene på x-aksen og residualene på y-aksen, observeres det at verdiene i stor grad konsentreres rundt null, som imotsetning til Shapiro – Wilk testen indikerer normalfordeling. Sammenlignet med regresjonen med en uavhengig variabel ser vi også at avvikene er mindre rundt midten av plottet. Dette tyder på at når vi inkluderer fars utdanning så ligger de predikerte verdiene nærmere de observerte.



I Q-Q plottet under ser vi at punktene ligger nær den rette linjen, indikerer også en normalfordeling. Observerer at punktene varierer mindre rundt linjen enn med bare en uavhengig variabel, som er konsistent med spredningsplottet fra forrige side. Ser også en liten antydning til for høye (høye) verdier, og for lave (lave) verdier.



Grafen som er presentert nedenfor er utarbeidet for å evaluere homoskedasitet, som vil si konstant i varians av feilleddene. Dette er utfordrende å avgjøre visuelt, men det kommer ikke frem noen klare mønstre i plottet, og linjen fremstår ikke som rett. I tillegg kjøres det en Breusch Pagan test. Test resultatet (finnes i vedlegg) viser en p – verdien litt over 0.05, noe som tyder på homoskedasitet. Dette indikere at vi kan ha tillit til standardfeilene, og som konsekvens av dette, de estimerte parameterne.



Samlet sett gir modellen blandende resultater når det gjelder normalfordelingen. Shapiro – Wilk testen indikerer med sine p – verdier at dataene ikke følger normalfordeling. Dette er i konflikt med resultatene fra den visuelle undersøkelsen av residualene i spredningsplottet og Q-Q plottet, som begge antyder at dataen er normalfordelt.

Dataen ser til å følge antakelsen om homoskedastisitet, som understøttes av Breusch – Pagan testen, og de visuelle undersøkelsene. Det impliserer at det ikke finne bevis for at variansen i feilleddene endrer seg med de uavhengige variablene, og at vi derfor kan stole på de estimerte parameterne.

Konklusjonen er at selv om det er noen avvik, så støtter modellen antakelsene nok, slik at vi kan anta at de estimerte parameterne er pålitelige. Dette antyder at modellen spesielt dersom man inkluderer fars utdanning, kan gi et presist estimat på døtrenes utdanning. Det er likevel lurt å være forsiktig i tolkningen av dataen, og det kan derfor være nytte med ytterligere analyser.

Kildeliste

Forelesning 8, Sok 2008 av Chris Rune Andersen.
Link : https://uit.instructure.com/courses/31419/files?preview=2649695

Forelesning 9, Sok 2008 av Chris Rune Andersen.
Link : https://uit.instructure.com/courses/31419/files?preview=2661339

Forelesning 10, Sok 2008 av Chris Rune Andersen.
Link: https://uit.instructure.com/courses/31419/files?preview=2672245

Forelesningsnotater gjort av student (2023).