Mappeinnlevering 2

Fakultet for biovitenskap, fiskeri og økonomi.

Kandidatnummer 6, SOK-2009, Høst 2023

09-11-2023

Table of contents

Figurliste

Oppgave 1:

a) Kjøre en enkel lineær regresjonsanalyse

Datasettet vi har fått inneholder mye data på inntekt, arbeidserfaring og utdanningsnivå, spesielt for kvinner. Jeg har lyst å se om det kan finnes en korrelasjon mellom utdanning og slekt, og vil da se om det er en større korrelasjon om at barn som har utdannede foreldre også tar utdanning selv.

Derfor tar jeg tar den avhengige variabelen "educ" som direkte oversatt er "Kvinnens utdanningsnivå i år" for 1975 og den uavhengige variabelen "mothereduc" som også direkte oversatt er "Kvinnens mor sitt utdanningsnivå i år" og sjekker om det er en korrelasjon.

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

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 ***
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
```

b) Forklaring av resultater

Når jeg kjører en enkel lineær regresjonsanalyse så får jeg ut Table 1.

R^2 forklarer at 18.95% av kvinners utdanning kan forklares av den uavhengige variabelen som er kvinnens mor sin utdanning. Man kan også se at for hvert "x" ekstra år utdanning for mødre så forventer vi at kvinner tar 0.294 ekstra år med høyere utdanning. Intercepten er på 9.55 som sier hva verdien på kvinners utdanning er når mothereduc er 0.

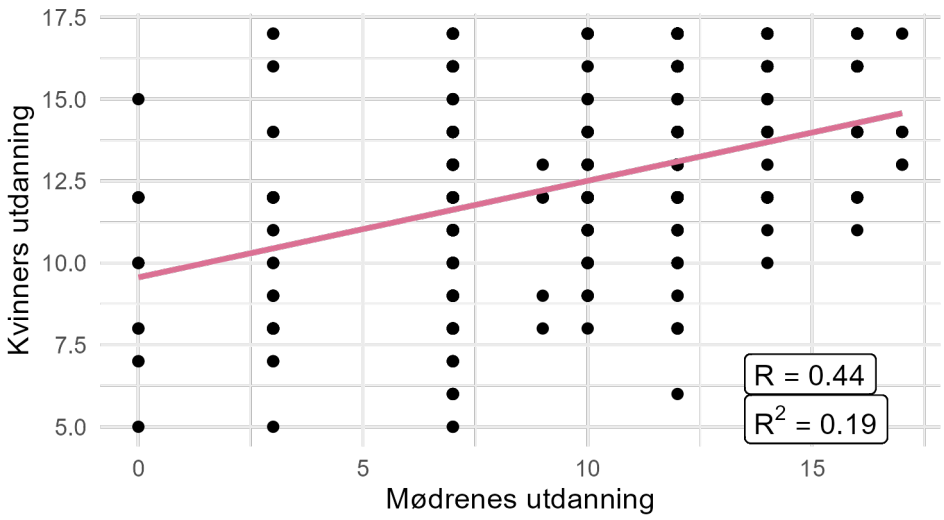
Std. Error eller standardfeilen forklarer hvor "sikker" man er på det man har estimert, om det er et høyt tall er man mer usikker på om estimatet er riktig utifra det utvalget man har tatt ifra en populasjon. En lavere standardfeil vil gi mindre usikkerhet.

Residual standard error er her 2.054 som betyr at på gjennomsnitt er avviket mellom observerte verdier og predikerte verdier på 2.054 år med utdanning. Degrees of freedom er antallet uavhengige variabler som er "fri" til å variere ved tilfeldig trekning.

	Avhengig variabel	
	Kvinners utdanning	
Intercept	9.55981***	(0.21898)
Mødres utdanning	0.29478***	(0.02224)
Observations	753	
R^2	0.1895	
Adjusted R^2	0.1884	
Residual Std. Error	2.054 (df = 751)	
F Statistic	175.6*** (df = 1; 751)	

Tabell 1: Lineær regresjonsanalyseresultat for kvinners utdanning og deres mødre.

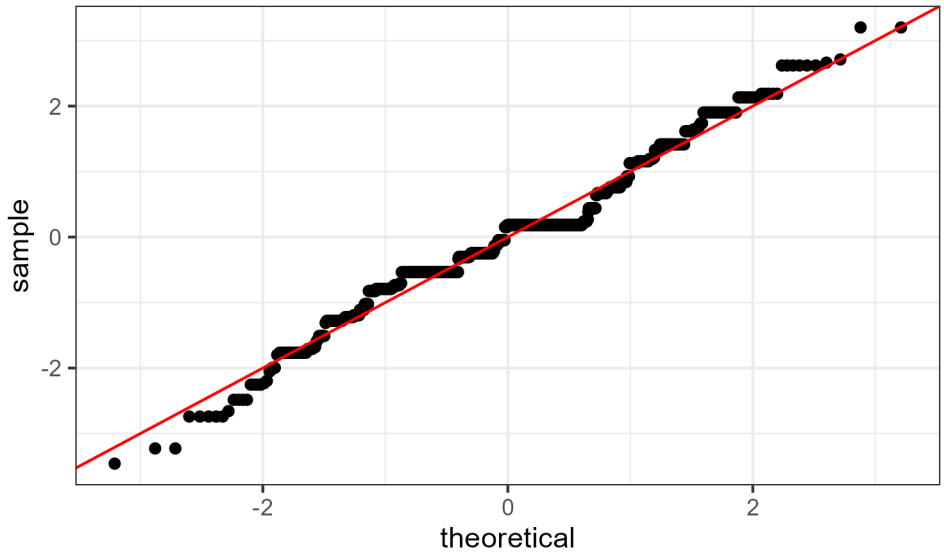
Tre stjerner eller "***" forteller meg tilslutt at resultatet er signifikant, som vil si at det er lite sannsynlighet for at resultatet har oppstått tilfeldig



Figur 1: Lineær regresjon mellom kvinner og kvinners mødre sin utdanning i år

I Figur 1 kan man se forholdet mellom den avhengige variabelen på y-aksen som er Kvinners utdanning mot den uavhengige variabelen mødrenes utdanning som er på x-aksen, og at korrelasjonen er positiv. Det vi tar ut av koeffisientene er at det er en "svak" mot "moderat" korrelasjon for verdien til korrelasjonskoeffisienten R som er på 0.44. Og som sagt tidligere sier R^2 at omtrent 19% av kvinners utdanning kan forklares av moren sin utdanning.

c) Bryter modellen med antakelsene



Figur 2: QQ - Quantile-Quantile plot

I (Quantile-Quantile) QQ-plottet sammenligner man fordelingen av residualene med en normal fordeling. Her ser vi fordelingen til "error termen" og ser at fordelingen ikke er normalfordelt, for da hadde den fulgt den røde linjen perfekt. Men det vi kan se er at resultatet er ganske

"nært" den røde linjen. For å se mer presist om modellen fungerer vil jeg heller gå videre å se på en multippel regresjon.

Oppgave 2:

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.

For å sjekke om forholdet mellom mine variabler er godt nok vil jeg legge til noen flere variabler for utdanning, og velger da å putte med to ekstra uavhengige variabler for å se om de kan gjøre regresjonen mer presis. Jeg velger da å plotte "educ" som den avhengige og "mothereduc" + "heduc" + "hfathereduc" som er mannens utdanning og faren til mannens utdanning i tillegg.

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

	Koeffisient	Standardfeil
Intercept	5.744757***	(0.331256)
Mødres utdanning (mothereduc)	0.178925***	(0.019631)
Partners utdanning (heduc)	0.397012***	(0.021875)
Fars utdanning (hfathereduc)	-0.008489	(0.019441)
Observations	753	
R^2	0.4373	
Adjusted R^2	0.435	
Residual Std. Error	1.714 (df = 749)	
F Statistic	194*** (df = 3; 749)	

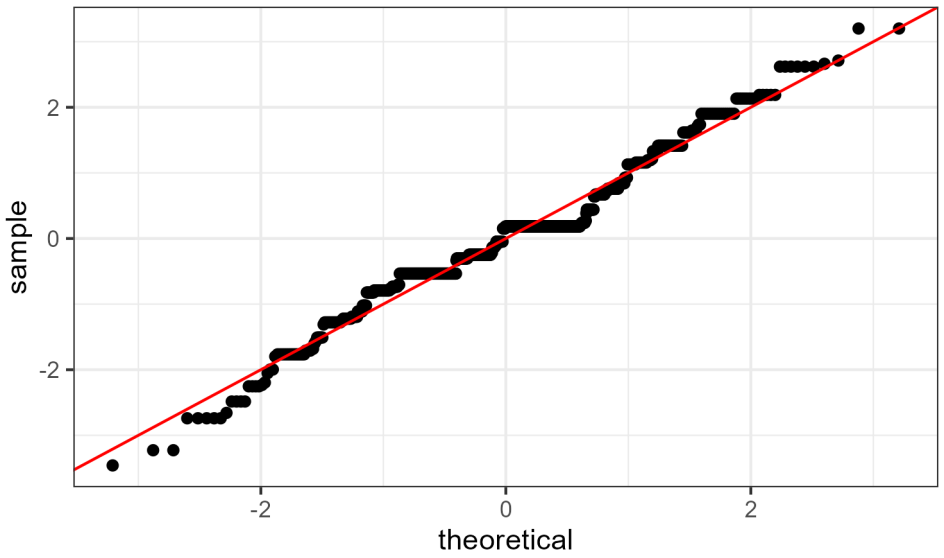
Tabell 2: Lineær multippel regresjonsanalyseresultat for kvinners utdanning.

R^2 forklarer at 43.73% av kvinners utdanning kan forklares av den uavhengige variabelen som er kvinnens mor sin utdanning. Man kan også se at for hvert "x" ekstra år utdanning for mødre så forventer vi at kvinner tar 0.17 ekstra år med høyere utdanning og enda mer med 0.39 for mannens utdanning, og et negativt forhold til faren til mannen med -0.08 omtrent. Intercepten er på 5.74 som sier hva verdien på kvinners utdanning er når de andre variablene er 0.

Det som også er interresant å se på er Adjusted R^2 i multippel regresjon og her er koeffisienten ganske lik R^2 med få desimaler, som betyr at forklaringsgraden til alle variablene er ganske presis, selv om man har tatt med flere. Hadde denne Adjusted R^2 vært mye lavere enn R^2 hadde man ikke kunne sagt at variablene hadde "hjulpet" med forklaringsgraden.

Residual standard error er her 1.714 som betyr at på gjennomsnitt er avviket mellom observerte verdier og predikerte verdier på 1.714 år med utdanning. Degrees of freedom er antallet uavhengige variabler som er "fri" til å variere ved tilfeldig trekning.

Tre stjerner eller "***" forteller meg tilslutt at resultatet er signifikant, som vil si at det er lite sannsynlighet for at resultatet har oppstått tilfeldig. Dette gjelder for morens utdanning, mannens utdanning men ikke for faren til mannens utdanning, denne er ikke signifikant.



Figur 3: QQ - Quantile-Quantile plot

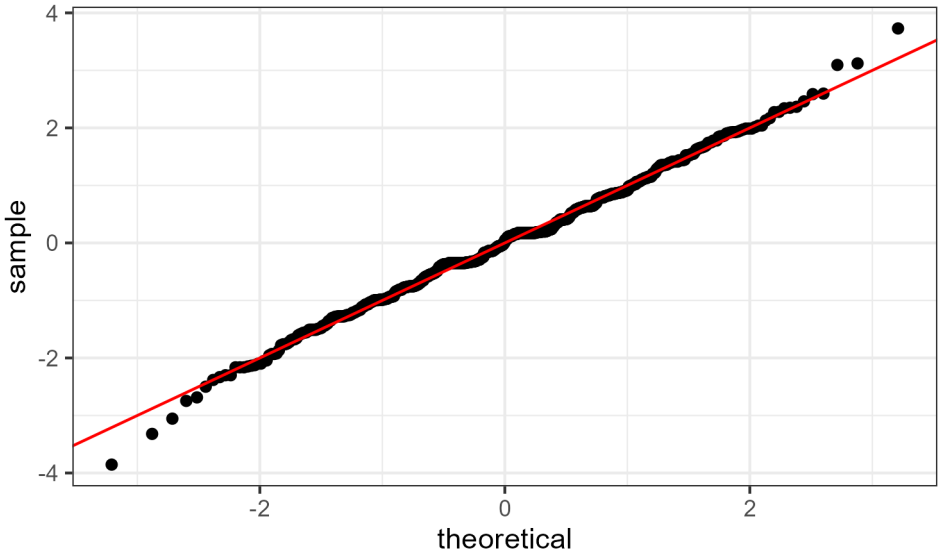
I Figur 3 kan man se forholdet mellom den avhengige variabelen på y-aksen som er Kvinners utdanning mot den uavhengige variabelen mødrenes utdanning som er på x-aksen, og at korrelasjonen er positiv. Og som sagt tidligere sier R^2 at omtrent 43% av kvinners utdanning kan forklares av moren sin utdanning.

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

```
    Shapiro-Wilk normality test

data:  residuals(model_2)
W = 0.99542, p-value = 0.02481
```

I (Quantile-Quantile) QQ-plottet sammenligner man fordelingen av residualene med en normal fordeling. Her ser vi fordelingen til "error termen" og ser at fordelingen ikke er normalfordelt, for da hadde den fulgt den røde linjen perfekt. Men det vi kan se er at resultatet er ganske "nært" den røde linjen. Som vil si at den ikke bryter med antakelsene mine.



Figur 4: QQ - Quantile-Quantile plot

Appendiks

Bruk av KI: ChatGPT 4 inkludert advanced data analysis.

Noe hjelp med hypotesen

Noe hjelp med c og tabeller