Mappeinnlevering 2

Fakultet for biovitenskap, fiskeri og økonomi.

Kandidatnummer 5, SOK-2009 Høst 2023

09-11-2023

Innholdsfortegnelse.

Oppgave 1:

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

Jeg vil se om det er en sammenheng mellom par sine utdanninger, og ser da på "Husband" sin utdanning mot "Wife" sin utdanning. Jeg kjører så en lineær regresjon mellom og vi ser at det er en trend som kan tyde på at de med høy utdanning også har en partner med høy utdanning.

Antagelser ved lineær regresjon.

I et dataset med et n verdier av x og y så regner vi regresjonslinjen ved $y_i = \alpha + \beta x + \epsilon_i$ der α blir skjeringspunktet i y aksen når $x = 0$, β er helningen på regresjonslinjen der en hver enhetsøkning i x, så øker y med β. ϵ_i er residualen for hver observasjon som da er forskjellen mellom en hver gitt $\alpha + \beta x$ og den observerte verdien i datasettet. Enhver gitt $\hat{y}$ viser oss da ethvert estimert punkt der summen av de kvadrerte residualene er minimert.

Antagelsene om ϵ_i er linearitet, normalitet, homoskedastisitet og uavhengighet.

Jeg forkarer de andre begrepene senere men jeg ikke helt hvordan jeg skal teste for uavhengighet men denne antagelsen er at feilene vi får ikke er korrelert med hverandre

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

Tabell 1: Lineært regresjonsresultat for ektefellers år med utdanning

	Avhengig variabel
	Mannens utdanning
Kvinnens utdanning	0.811^{***}
	(0.038)
Constant	2.530^{***}
	(0.478)
Observations	753
R^2	0.374
Adjusted R^2	0.374
Residual Std. Error	2.391 (df = 751)
F Statistic	449.615^{***} (df = 1; 751)
Note:	*p<0.1; **p<0.05; ***p<0.01

Koeffisienten til kvinners utdanning forteller oss at at for hvert ekstra år med utdanning som kvinnen har så er det en økning på 0.811 år i mannens utdanning.

"Constant" viser at dersom konen har 0 år med utdanning så er det forventet at mannen har 2.530år med utdanning.

Verdiene i parentes er standardfeilen til estimatene.

Det er 753 observasjoner i datasettet.

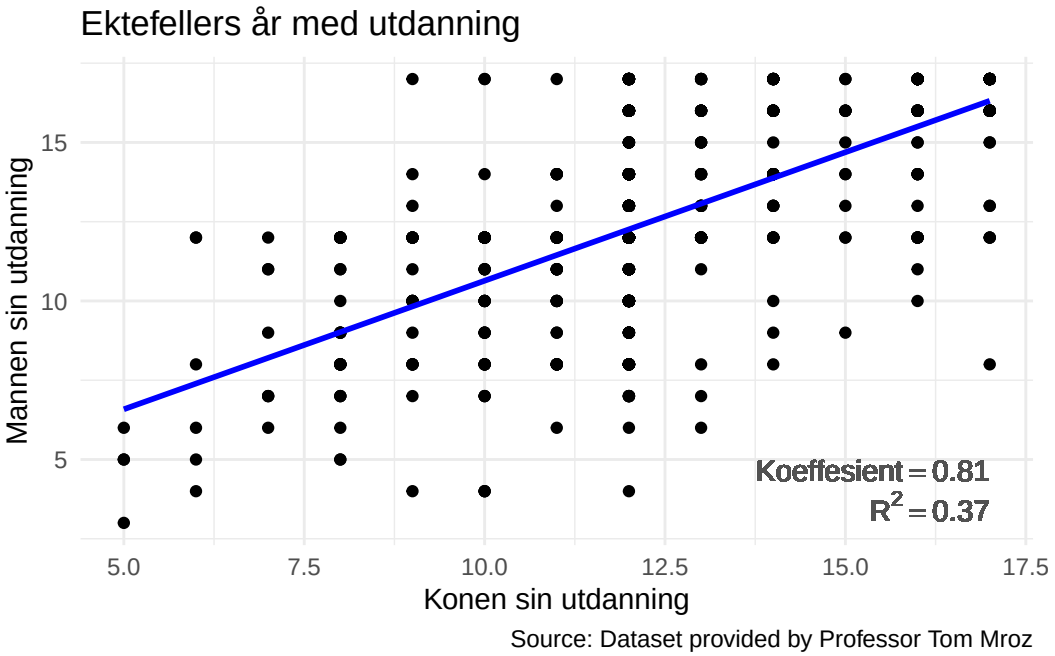
R^2 verdien på 0.374 forteller oss at 37.4% av variasjonen i mannen sin utdanning kan forklares av kvinnens utdanning i modellen. Vi kan også se AdjustedR^2 som er mer interessant i multippel regresjon da verdien tar hensyn til antallet variabler da denne "straffer" deg når det legges til flere variabler. I dette tilfellet er det bare de 2 variablene så denne blir lik R^2.

Residual STD. Error er standardavviket av feilene i modellen. 2.391 viser at det gjennomsnittlig er 2.391år forskjell mellom den preikerte verdien som modellen viser, og de faktiske datapunktene. df=751 angir at det er 751 "Degrees of freedom" som er antallet uavhengige variabler som er fri til å variere ved tilfeldig trekning.

F Statistic tester hvor godt kvinnens utdanning forklarer mannens utdanning. F verdien er den forklarte variansen per degrees of freedom delt på den uforklarte variansen per degrees of freedom. Det at denne er høy forteller oss at den forklarte variansen er høyere enn den uforklarte variansen og denne har 3 stjerner som betyr at den er signifikant med p verdi på under 0.01.

På bunnen kan vi se "Note" som viser hva de forskjellige stjernene betyr. Dette forteller oss P-verdien for forskjellige konfidensnivåer. P verdi forteller hva sannsynligheten er for at nullhypotesen er sann. Altså sjangsen for at resultatet vi har kan forekomme tilfeldig. Så for en gitt $\widehat{\beta}$(estimert β) så er da sjangsen for at den "faktiske" β verdien er 0.

Da vi ser at det er 3 stjerner så betyr dette at det er over en 99% sannsynlighet at det observerte resultatet skyldes en effekt og ikke har kommet fra tilfeldig trekning. Altså at nullhypotesen kan forkastes.



Jeg ser da at jeg har en høy koffesient så helningen er høy, og jeg har en R^2 som også er veldig høy. Dette forteller meg at modellen tyder på at det er en sammenheng mellom utdannelsen til parterene i ekteskap.

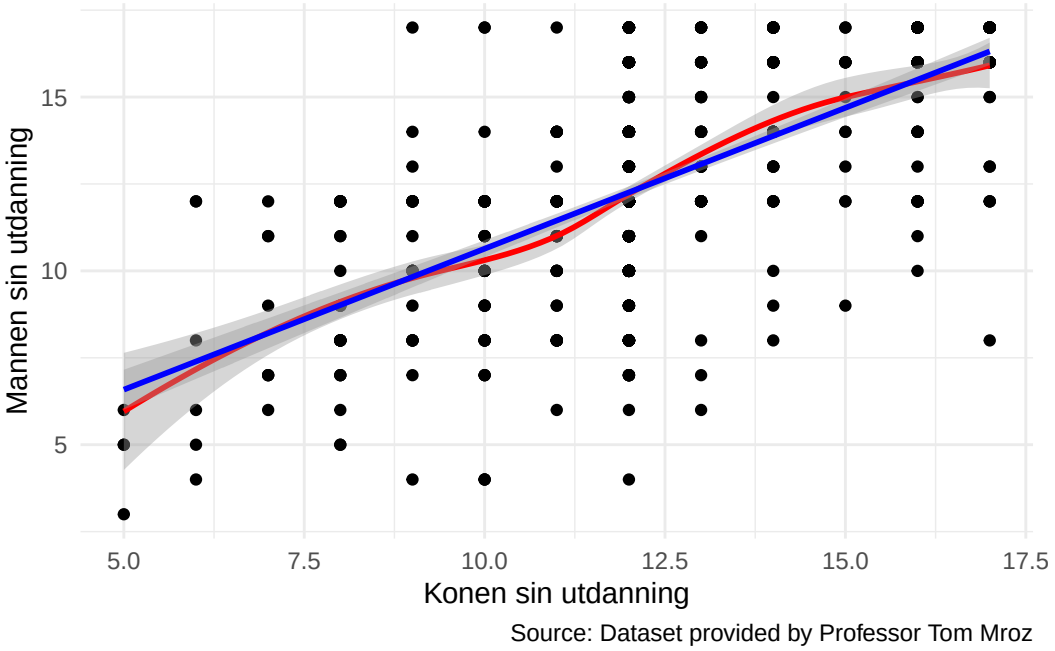
Dette viser da hvor godt utdanningen til kvinner påvirker mannens utdanning. Så om kvinnen tar et ekstra år med utdanning så får mannen 0.8år utdanning.

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.

Linearitet

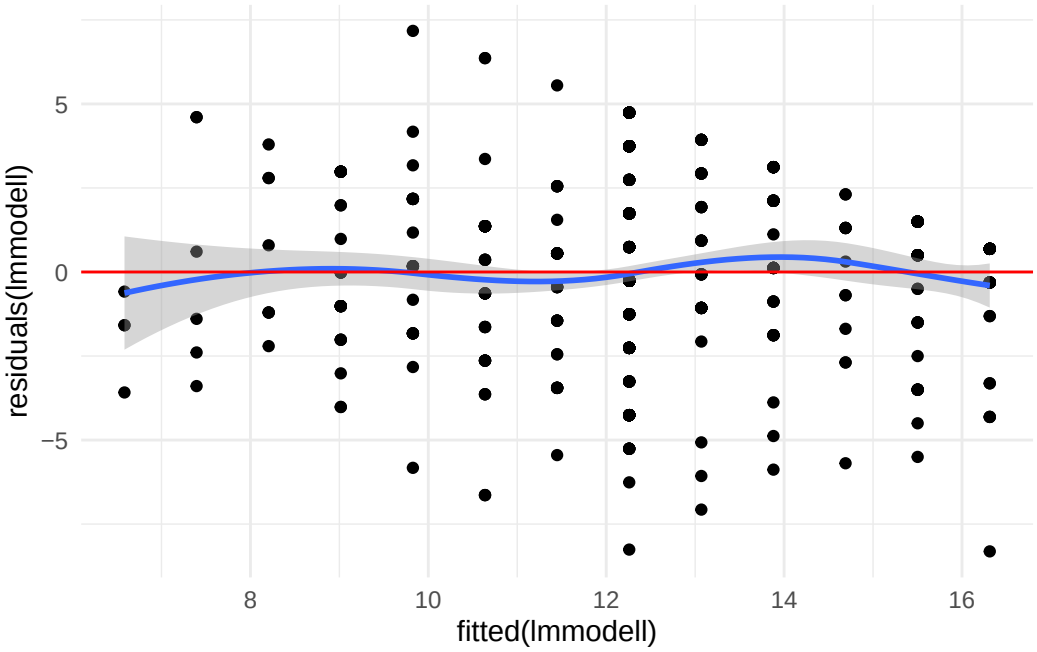
Linearitets antagelsen tilsir at de forventede utfallene er lineært relatert til de uavhengige variablene.

For å teste linearitet så tar jeg først en visuell undersøkelse av dataen og tegner opp en loess linje. Jeg tegner dette i tillegg til lineær regresjonslinjen da loess linjen er en lokalvektet glattet linje som lager en jevn linje igjennom de lokale datapunktene.



Loess linjen merket i rødt ser ut til å matche den lineære regresjonslinjen. Vi kan da se at det er ikke perfekt men den lineære modellen passer godt og loess linjen er ikke konsistent over eller under den linære linjen.

Jeg setter nå opp punktene til regresjonslinjen på y aksen og residualene på x aksen for å se etter brudd på lineæritet. Jeg legger igjen med en loess linje, og en rød horisontal linje.



Vi kan se at den lokalvektede linjen passer ganske godt med den horisontale linjen.

Normalitet

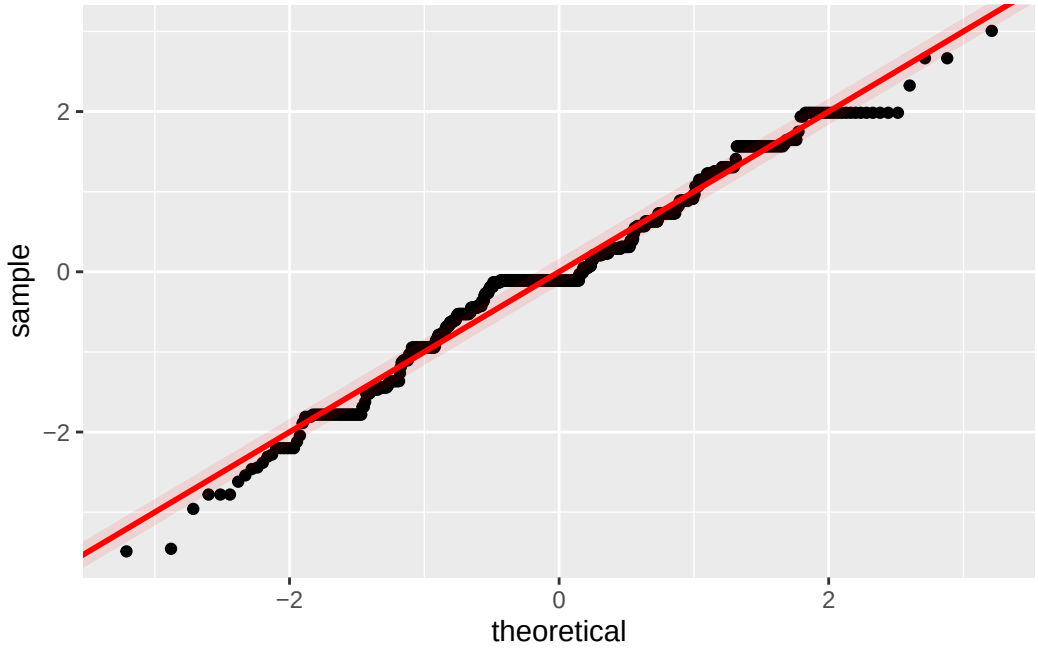
Normalitets antagelsen betyr at residualene er normalfordelt. Dette er nødvendig for å kunne estimere nøtaktige standardfeil. Vist i Tabell 1 som "Residual STD. Error"

Vi gjør da en Shapiro Wilk test

```
    Shapiro-Wilk normality test

data:  residuals(lmmodell)
W = 0.97587, p-value = 8.134e-10
```

Testen gir oss en svært lav P verdi som tyder på at residualene våre ikke er normalfordelt. P verdien her er sannynligheten for å observere en tilfeldig verdi som er 0.0000000008134 dersom nullhypotesen er sann som i testen er at residualene er normalfordelt. I økonomi kan det være normalt og ha en konfidensintervall på 0.95 så da testen gir en p verdi på under 0.05 så må jeg gjøre en visuall undersøkelse der jeg lager et kvantil-kvantil plot.



Dette var pussig. Vi fikk en figur der residualene markert med de svarte punktene følger den røde linjen som de ville gjort dersom residualene var normalfordelt.

Homoskedastisitet. Problem oppstår her.

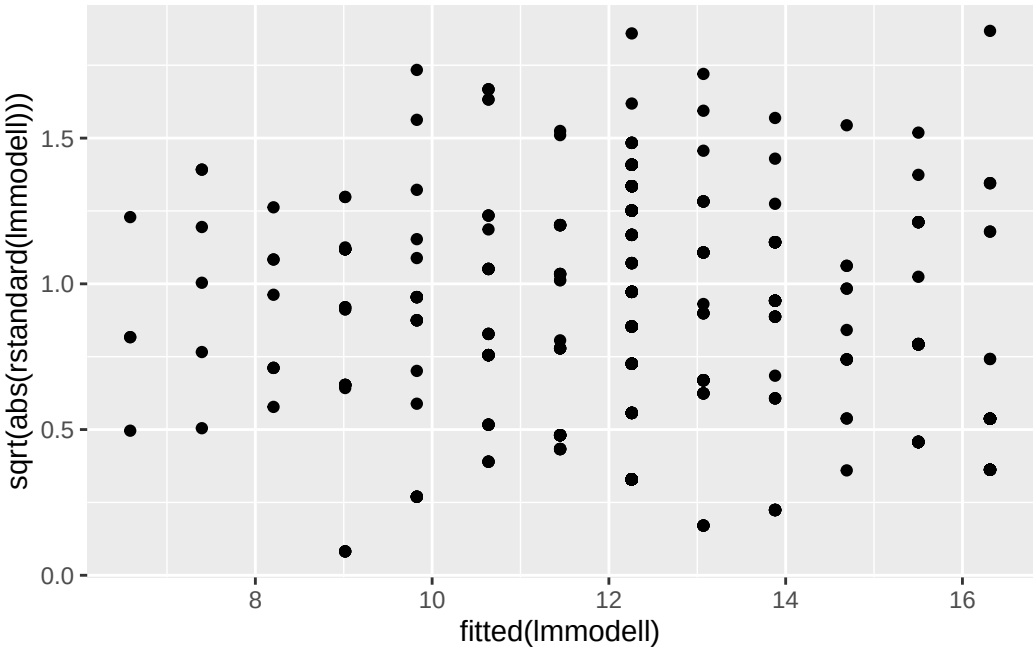
Homoskedastisitet betyr at residualene har lik varians for alle våre predikerte verdier i den lineære regresjonslinjen.

Jeg tester dette siden nøyaktige standardfeil er viktig for å kunne stole på resultatene som p verdiene vi får.

Jeg starter med en Breush-Pagan test. Denne testen gjør en egen regresjon med de predikerte verdiene som uavhengig variabel og tester om disse kan forklare variansen til residualene.

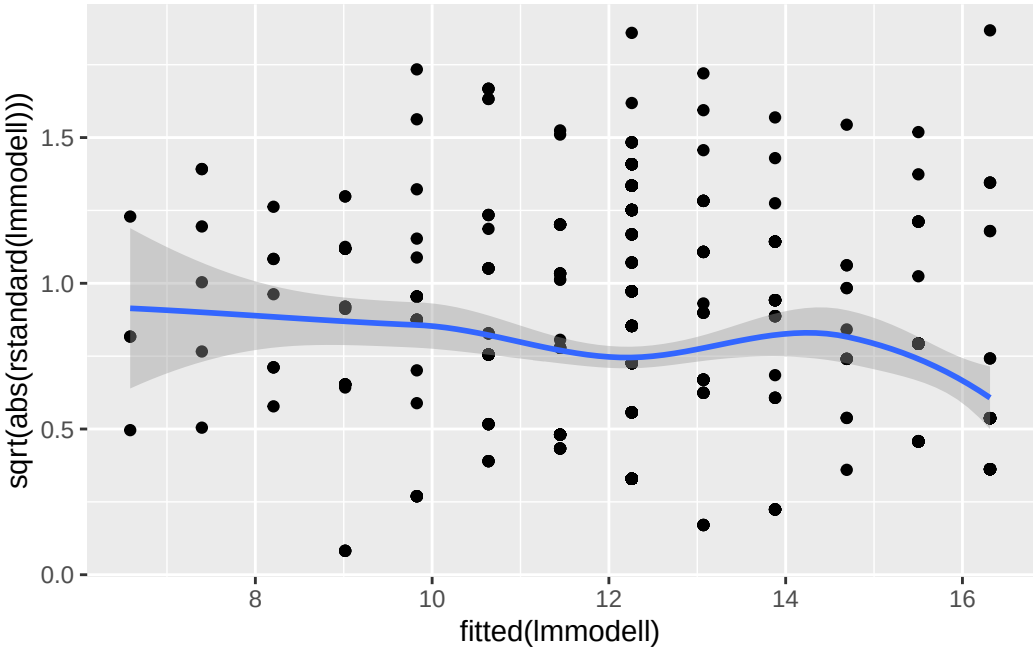
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 5.931168, Df = 1, p = 0.014875
```

Her fikk vi en P verdi som er på under 0.05 som ikke er bra. Dette forteller oss at sjangsen for at variansen på residualene ikke er konstant er veldig sannsynlig. Som bryter med homoskedastisitets antagelsen og at det da er heteroskedasisitet(?). Dette betyr at variansen på residualene ikke er lik for alle de predikerte verdiene.



Det vi her kan se er at det ikke ser ut til å være noe mønster. Så visuelt så kan det virke som at antagelsen om homoskedasisitet kan beholdes.

Jeg gjør da en siste visuell undersøkelse der jeg tar kvadratroten av de standariserte residualene mot de predikerte verdiene. Jeg håper da på en rett linje.



Det er en synkende varians som kan tyde på at antagelsen om homoskedasisitet er brutt. Den loess linjen vi får er fremdeles ganske flat men vi kan se at den synker, øker og synker igjen der vi ender på et punkt en del lavere enn startspunktet. Jeg vet ikke om dette tilsier at den er alvorlig brutt men dette kan bety at signifikansnivåene vi får ikke er nøyaktig og at det da er problemer som forskyver resultatet vi har.

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.

Da det så ut som det kunne være noe sammenheng mellom kvinnens og mannens utdanningsnivå så tar jeg nå med mannens foreldres utdanningsnivå for å se om dette kan forklare utdanningen til mannen.

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

Tabell 2: Multippel lineær regresjons menn

	Avhengig variabel
	Mannens utdanning
Kvinnens utdanning	0.811^{***}
	(0.038)
Mannens fars utdanning	0.002
	(0.030)
Mannens mors utdanning	-0.009
	(0.029)
Constant	2.590^{***}
	(0.559)
Observations	753
R^2	0.375
Adjusted R^2	0.372
Residual Std. Error	2.394 (df = 749)
F Statistic	149.522^{***} (df = 3; 749)
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabell 2 viste at det ikke ser ut til å være noe signifikant resultat på mannen sine foreldres utdanning. Så jeg tester mot kvinnes foreldres utdanning. Jeg gjøre det motsatte for kvinner for å se om jeg ser en klar effekt. Tabell 3 på neste side viser dette.

Jeg går ut ifra at jeg vil få problemer med uavhengighet da foreldres utdanning kanskje også har samme sammenheng.

Tabell 3: Lineære regresjonsresultater for ektefellers utdanningsvalg

	Avhengig variabel	
	Mannens utdanning	Kvinnens utdanning
	(1)	(2)
Kvinnens utdanning	0.734***	
	(0.044)	
Mannens utdanning		0.373***
		(0.022)
Mannens fars utdanning	−0.006	−0.022
	(0.030)	(0.022)
Mannens mors utdanning	−0.002	0.015
	(0.029)	(0.021)
Kvinnens fars utdanning	0.093***	0.101***
	(0.031)	(0.022)
Kvinnens mors utdanning	0.018	0.125***
	(0.032)	(0.023)
Constant	2.550***	5.642***
	(0.557)	(0.345)
Observations	753	753
R^2	0.386	0.453
Adjusted R^2	0.382	0.449
Residual Std. Error (df = 747)	2.374	1.692
F Statistic (df = 5; 747)	94.018***	123.767***
Note:		*p<0.1; **p<0.05; ***p<0.01

Vi kan se at for menn så får vi ingen signifikant resultat på hannes foreldes utdanning men den potentielle effekten ser uansett liten ut. Pussig at kvinnens fars utdanning haret signifikant resultat på hannes utdanning. Vi endrer ikke R^2 spesielt ved å legge til flere variabler og *Adjusted* R^2 viser fremdeles en marginal økning i forklaringskraft fra 0.374.

Men vi ser at for kvinner så ser det ut til å ha effekt hva hennes foreldres utdanning er og hva mannens utdanning er, så jeg fokuserer på denne effekten videre da vi fremdeles har en høy forklaringsgrad med signifikante resultater. Jeg sjekker hvordan effekt jeg får på R^2 og *Adjusted* R^2 ved å nå ta bort hennes manns foreldres utdanning da disse ikke ga et signifikant resultat og isolerer da hennes utdanning mot mannens utdanning og hennes foreldres utdanning i tabell 4.

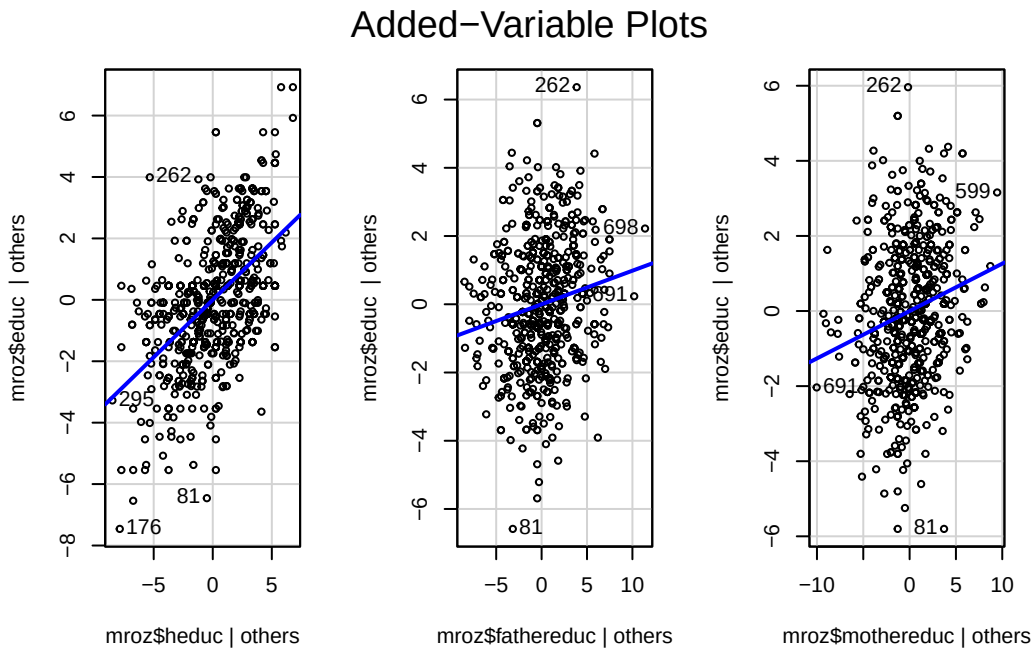
Tabell 4: Multippel lineær regresjons kvinner

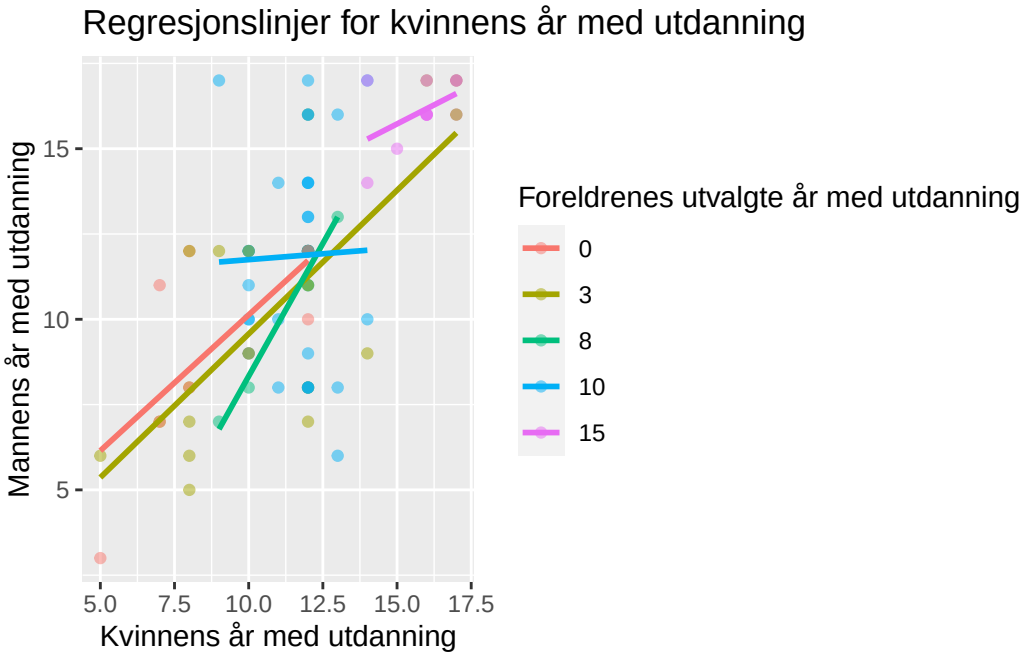
	Avhengig variabel
	Kvinnens utdanning
Mannens utdanning	0.374***
	(0.022)
Kvinnens fars utdanning	0.098***
	(0.022)
Kvinnens mors utdanning	0.126***
	(0.023)
Constant	5.586***
	(0.278)
Observations	753
R^2	0.452
Adjusted R^2	0.450
Residual Std. Error	1.691 (df = 749)
F Statistic	206.124*** (df = 3; 749)
Note:	*p<0.1; **p<0.05; ***p<0.01

R^2 og *Adjusted* R^2 har ikke hatt stor endring så denne modellen bruker jeg fremover.

Det vi her kan se er det er et signifikant resultat på kvinnens foreldres utdanning og mannens utdanning. Det vi kan se er at får hvert ekstra år med utdanning kvinnens far har så har hun 0.098 år ekstra utdanning, og hennes mors effekt er 0.126 år. Og får hvert ekstra år mannen har med utdanning så øker hennes med 0.374år.

De fine figurene vi får er så dette.





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

Jeg tester nå modellen i Tabell 4.

Kollinearitet

Jeg tester kollinearitet for å se om det er korrelasjon mellom de uavhengige variablene siden dette kan kunstig øke svingingene i modellen og svekker nøyaktigheten til modellen slik at vi ikke kan stole på p verdiene.

Jeg tester kollinearitet først ved å lage en korrelasjonsmatrise.

```
                educ     heduc fathereduc mothereduc
educ       1.0000000 0.6119538  0.4424582  0.4353365
heduc      0.6119538 1.0000000  0.3666996  0.3244747
fathereduc 0.4424582 0.3666996  1.0000000  0.5730717
mothereduc 0.4353365 0.3244747  0.5730717  1.0000000
```

Vi ser at korrelasjonene er under 0.7 men de er fortsatt noe høye.

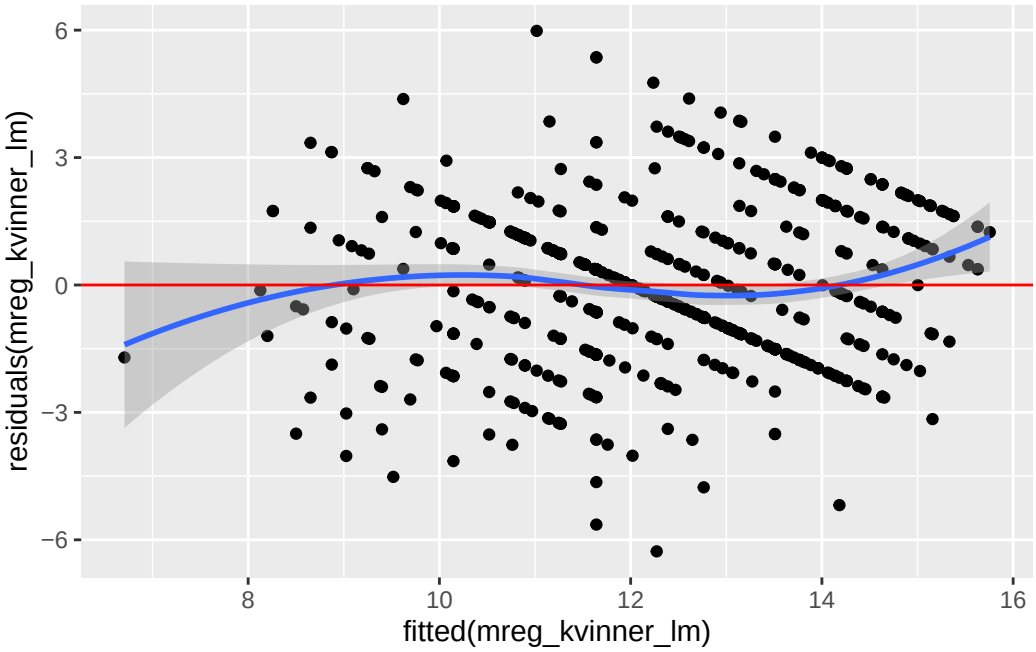
Jeg gjør nå en Variance Inflation Factor test for å få en målestokk for graden av kollinearitet.

```
mroz$heduc mroz$fathereduc mroz$mothereduc
  1.181938        1.574622        1.523260
```

Jeg ser her at verdier jeg får er på rundt 1 til 1.6 og det forteller at det ikke er stor grad av kollinearitet.

Linearitet

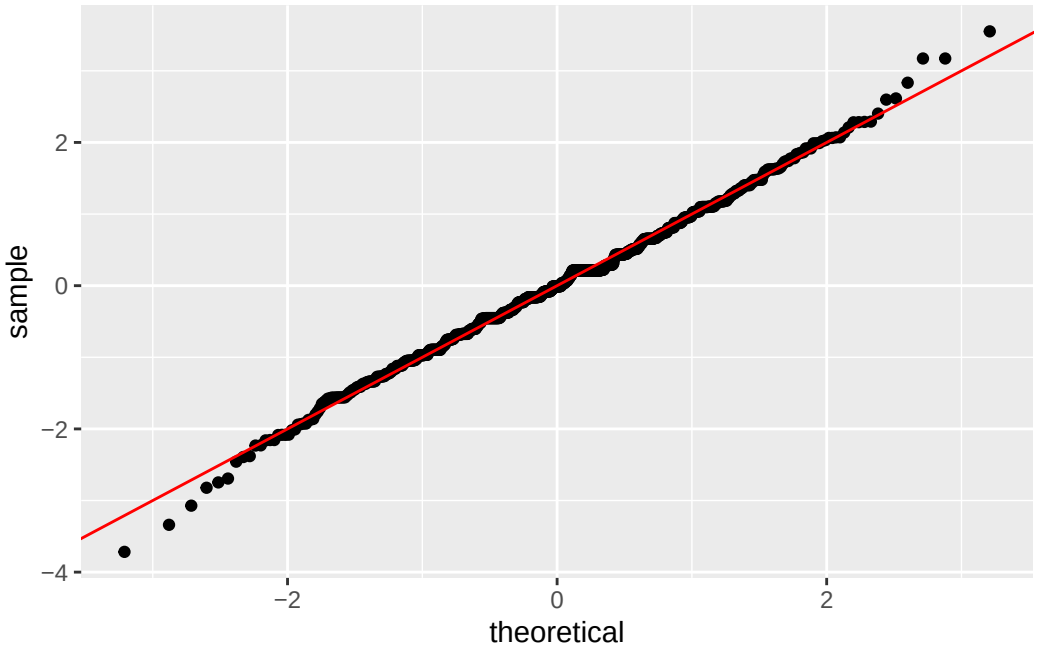
Jeg setter opp en figur med residualene på y aksen og de predikerte verdiene på x aksen.



Holder seg noe bra men avviker på start og slutt.

Normalitet

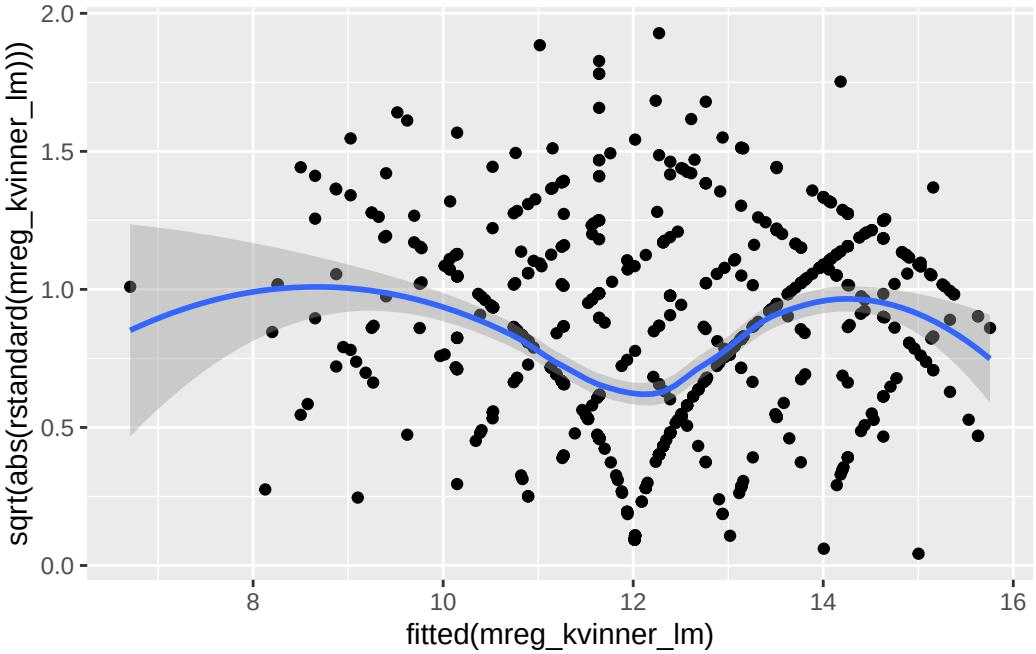
Jeg sjekker normaliteten i et kvantil-kvantil plot.



Igjen ser vi at den holder seg veldig bra men avviker på start og slutt.

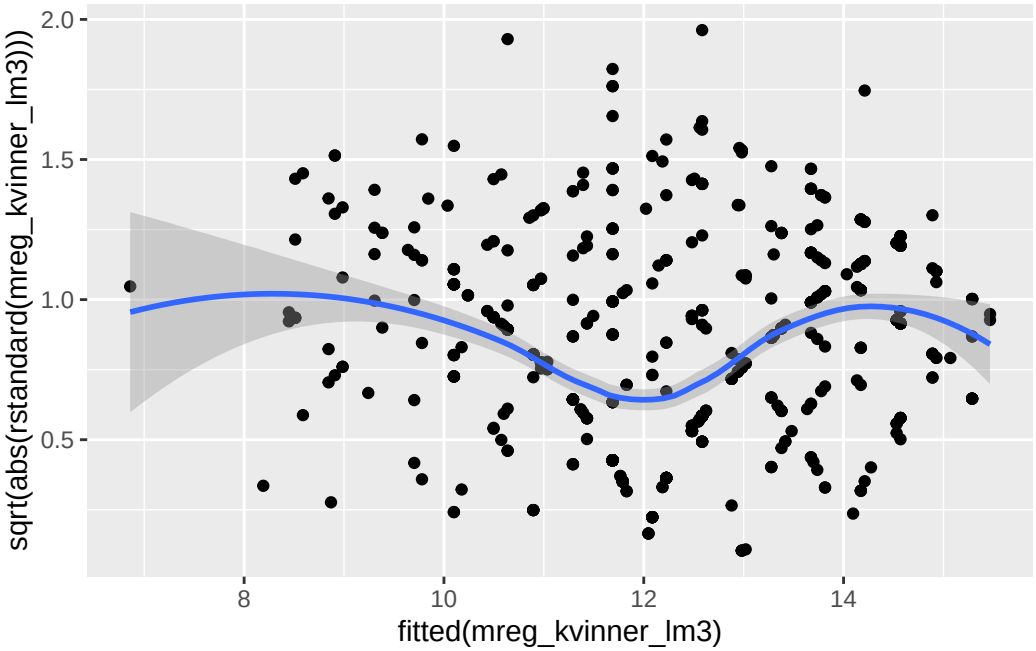
Homoskedasisitet

Jeg lager en figur der jeg tar kvadratroten av absoluttverdiene til de standardiserte residualene på y aksen og de predikerte verdiene på x aksen.



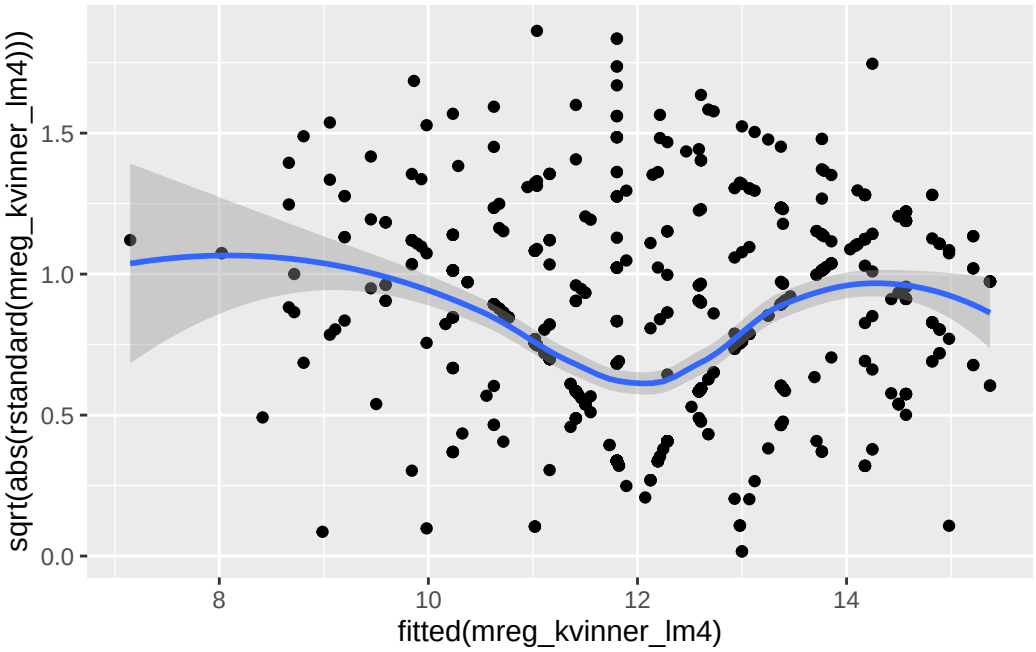
Her ser vi at det er et problem. Jeg vet ikke om dette betyr at det er et alvorlig brudd men om jeg hadde en linjal som så slik ut så ville jeg ikke prøvd å måle noe med den.

Jeg prøver å fjerne farens utdanning for å se effekten.

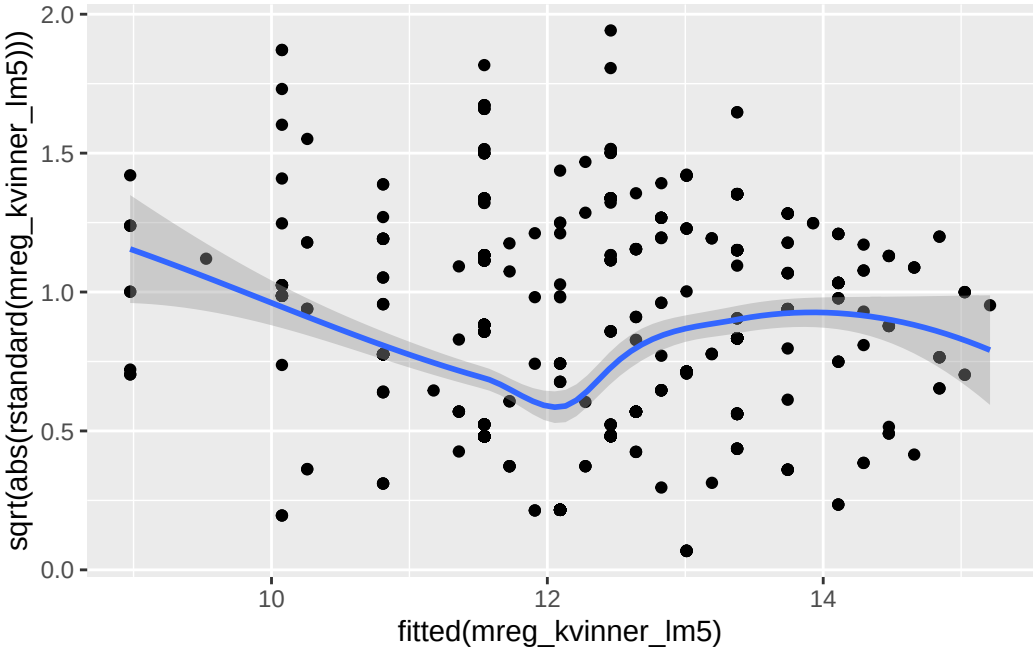


Den ble noe rettere men fremdeles tydelig buet.

Prøver modellen men denne gangen uten morens utdanning.



De virker til å ikke ha stor effekt så jeg prøver å fjerne mannens utdanning.



Det var pussig. Det ser ut som modellen ikke kan stoles på da det ikke kan påvises homoskedasisitet.

Gjør nå en NCV test igjen

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.3597919, Df = 1, p = 0.54862
```

Her fikk vi en P verdi som er på over 0.05 som er bra. Dette forteller oss at sjangsen for at variansen på residualene ikke er konstant er lite sannsynlig. Dette stemmer med homoskedastisitets antagelsen. Dette betyr at variansen på residualene kan være lik for alle de predikerte verdiene.

Konklusjon

Jeg kan da konkludere med at den lineære modellen mellom Menns utdanning og Kvinners utdanning virket som den kunne stoles på men i den multipple regresjonsmodellen så kan jeg ikke konkludere med mye da jeg ikke føler jeg kan stole på modellen da antagelsen om homoskedasisitet er alvorlig brutt.