12_sok_2009_h23_arbeidskrav2

12

```
# Clering the environment.
rm(list=ls())

# Loading packages.
suppressPackageStartupMessages({
  library(tidyverse)
  library(rjstat)
  library(janitor)
  library(readr)
  library(broom)
  library(flextable)
  library(stargazer)
  })
```

```
# Define custom theme function.
theme_mod_v1 <- function() {
  theme_minimal() +
    theme(
      plot.background = element_rect(fill = "white", color = NA),
      plot.margin = unit(c(5, 10, 5, 10), units = "mm"),
      plot.title = element_text(face = "bold.italic", color = "gray10"),
      axis.title = element_text(face = "bold", color = "gray10"),
      axis.text = element_text(color = "gray10"),
      legend.text = element_text(color = "gray10"),
      legend.title = element_text(face = "bold", color = "gray10"),
      panel.grid = element_line(color = "gray80")
    ) }
```

```
# Loading data.
load(url("http://www.principlesofeconometrics.com/poe5/data/rdata/mroz.rdata"))
```

Oppgavesett 1

Oppgave A)

Dersom en kjører en enkel lineær regresjonsanalyse av den avhengige variabelen «wage» (konas gjennomsnittlige timelønn, i 1975 dollar) og den uavhengige variabelen «educ» (konas utdannelse, i år) kan en undersøke sammenhengen mellom hvor lang utdanning kvinner (kona) har og deres gjennomsnittlige timelønn i 1975, altså hvordan utdanningsnivået påvirker timelønnen til kvinner i 1975.

```
# Selecting the data i need and filtering for women who worked in 1975.
df_01 <- mroz %>%
  select(educ, wage, lfp) %>%
  filter(lfp > 0)
```

```
# Running a linear regression model.
lm_model_01 <- lm(wage ~ educ, data = df_01)

# Shwoing the model.
stargazer(lm_model_01, type = "text", title = "LM modell", align = TRUE,
          dep.var.labels = c("Wage"), covariate.labels = c("Education"),
          no.space = TRUE, single.row = TRUE, keep.stat = c("n"))
```

```
LM modell
=========================================
                Dependent variable:
             ---------------------------
                        Wage
-----------------------------------------
Education          0.495*** (0.066)
Constant           -2.092** (0.848)
-----------------------------------------
Observations              428
=========================================
Note:         *p<0.1; **p<0.05; ***p<0.01
```

Fra tabellen ovenfor kan en se det estimerte konstantleddet (constant), som representerer lønnen når utdanningen er null enheter. Modellen estimerer at dersom kvinner i 1975 aldri går på skole, vil dem være nødt til å betale arbeidsgiver 2.092 dollar for å få jobb, altså vil ingen være villig å tilby dem jobb, noe som kan indikere at modellen fungerer dårligere med nullverdier.

En kan også se koeffisienten (education), som representerer hvor mye gjennomsnittlig lønn endrer seg for hver enhet økning i utdannelse. Altså vil lønnen til kvinner øke med ca. 0.49 dollar for hvert år de utdanner seg, ifølge denne modellen. Avslutningsvis ser en at modellen inneholder 428 observasjoner.

Oppgave B)

En kan nå se nærmere på modellen. Nedenfor kan en se de samme verdiene som i oppgaven ovenfor, men også standardfeilen (Std. Error) for estimatet av koeffisienten i en egen kolonne. Standardfeilen viser hvor stor usikkerhet det er rundt estimatet, mindre standardfeil, mer nøyaktig estimat.

```
# Extracting "Estimate" and "Std. error" from the lm.
model_01_coef <- coef(summary(lm_model_01))[ , c("Estimate", "Std. Error")]

# Showing values.
stargazer(model_01_coef, type = "text", title = "Std. Error", align = TRUE,
          no.space = TRUE, single.row = TRUE)
```

```
Std. Error
================================
            Estimate Std. Error
--------------------------------
(Intercept)  -2.092    0.848
   educ       0.495     0.066
--------------------------------
```

Som en kan se er standardfeilen til konstantleddet på ca. 0.84 dollar, så en kan anta at i denne modellen vil en lønn med null enheter utdannelse, ligge mellom ca. -1.25 til -2.93 dollar (differansen mellom konstantleddet og standardfeilen, begge veier). Standardfeilen til koeffisienten er på ca. 0.06 dollar, som vil si at i denne modellen vil gjennomsnittlig lønn endre seg med ca. 0,43 til 0,55 for hvert ekstra år med utdanning. Begge disse estimatene har relativt små standardfeil, noe som indikerer et nøyaktig estimat.Videre kan en se på p-verdien som er et mål på om kvinner utdannelse og timelønn er statistisk signifikant.

```
# Extracting p values from the lm.
model_01_p <- coef(summary(lm_model_01))[ , "Pr(>|t|)"]

# Showing values.
stargazer(model_01_p, type = "text", title = "P-value", align = TRUE,
```

```
           no.space = TRUE, single.row = TRUE)
```

```
P-value
================
(Intercept) educ
----------------
0.014        0
----------------
```

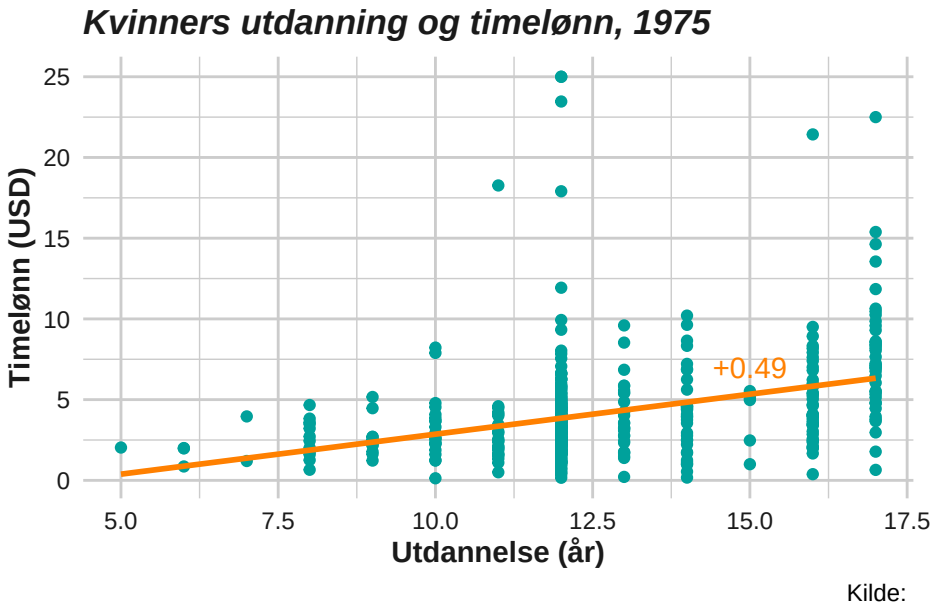
Som en kan se fra modellen ovenfor er p-verdien for kvinner utdannelse (educ) lik 0 (0.0000003485984) noe som antyder at sammenhengen mellom variablene er statistisk signifikant, altså at det er en sammenheng. En kan også se på korrelasjonen for å avgjøre styrken av forholdet mellom variablene. Korrelasjonens vil være mellom -1 og 1, hvor 1 representerer en sterk positiv sammenheng og -1 representerer en sterk negativ sammenheng. En verdi mellom -0.2 og 0.2 vil indikere svak eller ingen sammenheng.

```
# Calculating the correlation between the variables.
model_01_cor <- cor(df_01$wage, df_01$educ)

# Showing the correlation value.
stargazer(model_01_cor, type = "text", title = "Correlation", align = TRUE,
          no.space = TRUE, single.row = TRUE)
```

```
Correlation
=====
0.342
-----
```

Som en kan se fra verdien ovenfor er korrelasjonen mellom variablene ca. 0.342, noe som indikere en svak positiv korrelasjon, altså når den ene variabelen øker vil den andre også øke. Avslutningsvis kan en se på dataen grafisk. I figuren nedenfor kan en se timelønnen (1975 dollar) på y-aksen (avhengig variabel) og utdannelse (antall år) på x-aksen (uavhengig variabel) for kvinner i 1975. Det er også lagt inn en lineær regresjonsmodell (+0.49) som viser hvor mye timelønnen øker med per enhet utdannelse.

```r
# Creating figure.
ggplot(df_01, aes(educ, wage)) +
  geom_point(color = "#00A19B") +
  geom_smooth(method = lm, se = FALSE, color = "#FF8000", linewidth = 1) +
  labs(title = "Kvinners utdanning og timelønn, 1975",
       x = "Utdannelse (år)", y = "Timelønn (USD)", caption = "Kilde: ") +
  annotate("text", x = 15, y = 7, label = "+0.49", color = "#FF8000") +
  theme_mod_v1()
```



Som en kan se fra modellen ovenfor antyder den lineære regresjonsmodellen at høyere utdanning fører til høyere lønn, men en kan også observere at det er relativt stor spredning i dataen. Vi observerer for eksempel at noen av de med høyest utdanning har lik timelønn som de med lavest utdanning. Dette kan antyde at det er en svak sammenheng mellom utdannelse og timelønn for kvinner. Det vil kreve mer data og flere analyser for å kunne bekrefte årsakssammenheng.

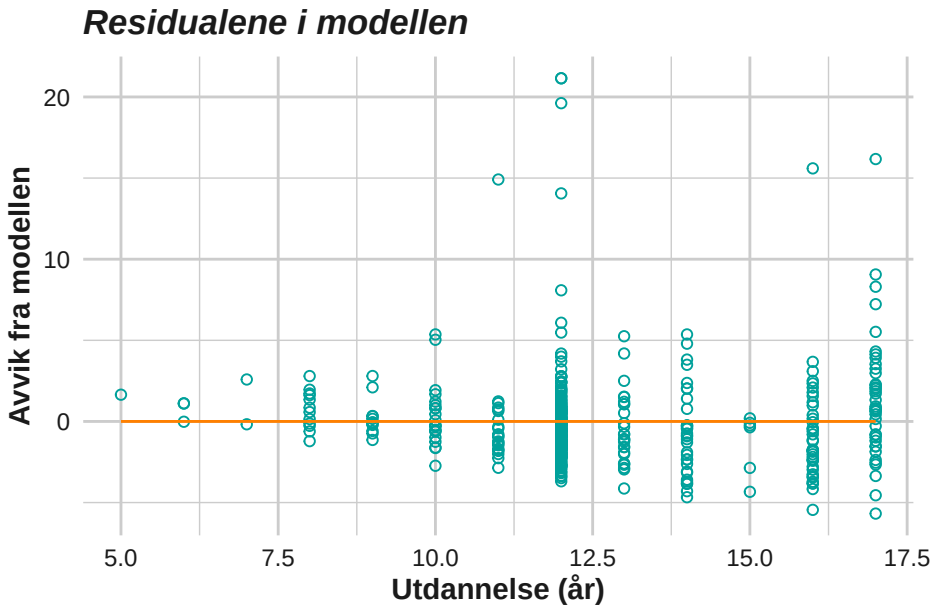
Oppgave C)

Når en skal lage modeller ved bruke av lineær regresjon er det basert på noen antakelser og dersom modellen bryter med disse antakelsene kan det påvirke gyldigheten til regresjonsanalysen. En av disse antakelsene er at det skal foreligge et lineært forhold mellom variablene. Dersom

forholdet mellom variablene ikke er lineært, vil modellen kunne overvurdere eller undervurdere kvinners utdannelses virkning på timelønnen.

En annen antakelse er at residualene skal være uavhengige. Residual i en regresjonsanalyse er all variasjon i en effektvariabel som en modell ikke klarer å fange opp (Wikipedia, 2021), altså forskjellen mellom den faktiske observerte verdien og verdien beregnet av modellen. Dersom feilen ikke er uavhengig, kan det føre til en automatisk korrelasjon i dataen. Nedenfor kan en se residualene visualisert:

```
# Extracting residuals from linear model.
res_01 <- resid(lm_model_01)

# Creating figure.
ggplot(data = NULL, aes(x = df_01$educ, y = res_01)) +
  geom_point(shape = 1, color = "#00A19B") +
  geom_line(aes(x = seq(5,17), y = 0), color = "#FF8000") +
  labs(title = "Residualene i modellen",
       x = "Utdannelse (år)", y = "Avvik fra modellen") +
  theme_mod_v1()
```

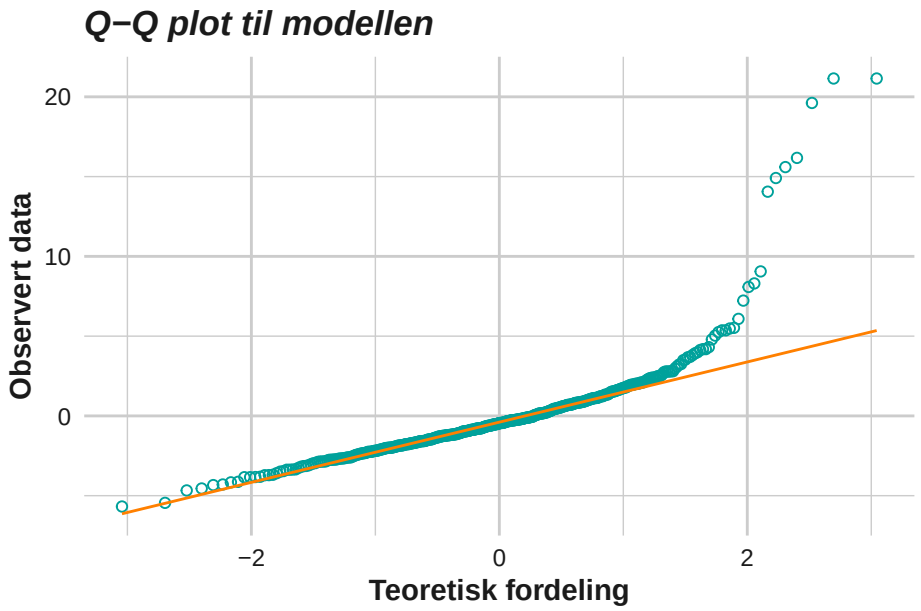


På figuren ovenfor kan en se avviket fra modellen på y-aksen og antall år utdannelse på x-aksen. Ved bruk av denne figuren kan en avgjøre om modellen har et lineært forhold og om residualene er uavhengige. Dersom en observerer et mønster som f.eks. først positive verdier,

så negative, så positive, osv. vil dette antyde avvik fra linearitet. En observerer fra figuren at det ikke foreligger et slik mønster noe som kan tyde på at antakelsen for linearitet er oppfylt. Siden en ikke observerer et mønster vil dette også indikere at residualene er uavhengige, noe som oppfyller antakelsen.

Videre kan en undersøke om modellen oppfyller antakelsen om at residualene er tilnærmet normalfordelt. Dette kan en undersøke ved bruk av et Q-Q plot (quantile – quantile plot). Et Q-Q plot sammenligner kvantiler fra den observerte dataen med kvantiler fra den teoretiske fordelingen (Wikipedia, 2023). Dette er visualisert nedenfor:

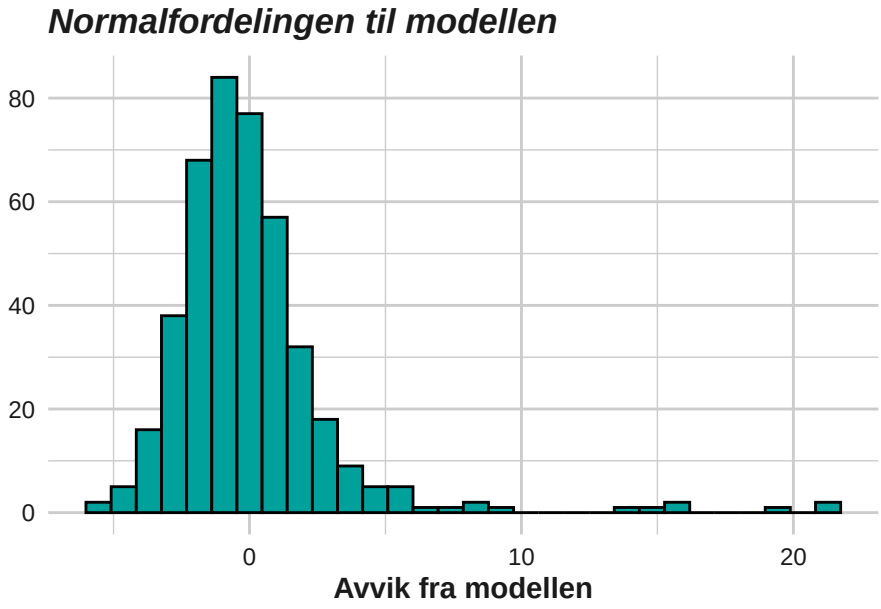
```
# Creating dataframe.
res_df_01 <- data.frame(res_01)

# Creating Q-Q plot.
ggplot(res_df_01, aes(sample = res_01)) +
  geom_qq(shape = 1, color = "#00A19B") +
  geom_qq_line(color = "#FF8000") +
  labs(title = "Q-Q plot til modellen", x = "Teoretisk fordeling",
       y = "Observert data") +
  theme_mod_v1()
```



Fra figuren ovenfor kan en se kvantiler fra den teoretiske fordelingen på x-aksen og kvantiler for den observerte dataen på y-aksen. En perfekt normalfordeling vil følge linjen i plottet, altså

at observasjonene følger den forventede fordelingen. Som en kan se fra figuren over avviker observasjonene fra linjen mot slutten av figuren. Dette antyder høyreskjev fordeling, når en ønsker en symmetrisk fordeling (bjelle form). Dette er visualisert nedenfor:

```
# Creating figure.
ggplot(res_df_01, aes(x = res_01)) +
  geom_histogram(fill = "#00A19B", color = "black") +
  labs(title = "Normalfordelingen til modellen",
       x = "Avvik fra modellen", y = " ") +
  theme_mod_v1()
```



En kan altså ikke si at residualene er tilnærmet symmetrisk normalfordelingen, noe som kan tyde på at modellen ikke er pålitelig.

Oppgavesett 2

Oppgave A)

Dersom en kjører en multippel lineær regresjonsanalyse av den avhengige variabelen «faminc» (familiens inntekt) og de uavhengige variablene «wage» (timelønn kvinner) og «hwage» (timelønn menn), kan en undersøke hvordan familiens inntekt påvirkes av lønnen til begge

kjønn. Forskjellen fra en enkel- og multippel lineær regresjonsanalyse er at en ser effekten av to uavhengige variabler på den avhengige variabelen, istedenfor en.

```r
# Selecting the data i need and filtering for men and women with wage > 0.
df_02 <- mroz %>%
  select(faminc, wage, hwage) %>%
  filter(wage > 0, hwage > 0)
```

```r
# Running a linear regression model.
lm_model_02 <- lm(faminc ~ wage + hwage, data = df_02)

# Showing model.
stargazer(lm_model_02, type = "text", title = "LM modell", align = TRUE,
          dep.var.labels = c("Family income"),
          covariate.labels = c("Wage (Women)", "Wage (Men)"),
          no.space = TRUE, single.row = TRUE, keep.stat = c("n"))
```

```
LM modell
=========================================
                Dependent variable:
             ---------------------------
                    Family income
-----------------------------------------
Wage (Women)    585.326*** (127.092)
Wage (Men)     2,068.462*** (117.805)
Constant       6,737.945*** (989.846)
-----------------------------------------
Observations             428
=========================================
Note:        *p<0.1; **p<0.05; ***p<0.01
```

Fra tabellen ovenfor kan en se det estimerte konstantleddet (constant), som representerer husholdningen inntekt dersom timelønnen til både kvinner og menn er lik null. Siden konstantleddet er lik ca. 6.738 (dollar, 1975) kan dette tyde på at husholdningene enten har tilgang på økonomiske goder som f.eks. trygd, eller at modellen ikke fungerer for husholdninger der mann og kvinner har timelønn lik null. Videre kan en se hvordan endringer i kvinners timelønn (wage (women)), når mannens timelønn holdes konstant, påvirker husholdningens inntekt. Altså vil husholdningens inntekt øke med ca. 585 (dollar, 1975) for hver enhet ekstra med kvinnelig timelønn, så fremst mannens timelønn holdes konstant. En kan også se det samme for hvordan mannens timelønn (wage (men)) påvirker, altså en økning på ca. 2.068 (dollar, 1975) per enhet timelønn. Avslutningsvis ser en at modellen inneholder 428 observasjoner.

Disse variablene ble valgt med bakgrunn i at det forventes en sterk korrelasjon mellom timelønnen til kjønnene og familiens inntekt. Men det kan også være andre faktorer som trygde eller andre goder som kan være med å påvirke resultatet. Resultatene vil også forventes å kunne vise forskjellene mellom kvinner og menn i 1975.

Oppgave B)

Nedenfor kan en se de samme verdiene som i oppgaven ovenfor, bare at standardfeilen (Std. Error) for estimatet av koeffisienten er lagt i en egen kolonne (forklaring på standardfeil i oppgavesett 1).

```
# Extracting "Estimate" and "Std. error" from the lm.
model_02_coef <- coef(summary(lm_model_02))[ , c("Estimate", "Std. Error")]

# Showing model.
stargazer(model_02_coef, type = "text", title = "Std. Error", align = TRUE,
          no.space = TRUE, single.row = TRUE)
```

```
Std. Error
================================
            Estimate  Std. Error
--------------------------------
(Intercept) 6,737.945  989.846
   wage      585.326   127.092
   hwage    2,068.462  117.805
--------------------------------
```

Standardfeilen rundt konstanten (intercept) er relativt høy (ca. 990) noe som kan indikere at modellen ikke fungerer godt for nullverdier, som diskutert i forrige oppgave. Standardfeilen rundt kvinners timelønn er også relativt høy (ca. 127), noe som kan indikere at det er store variasjoner i kvinners timelønn innvirkning på familiens inntekter. Timelønnen til mannen er innenfor et akseptabelt nivå (ca. 117), noe som kan indikere at modellen er relativt presis på disse verdiene.

Videre kan en se på p-verdien som er et mål på om timelønnen (begge kjønn) og husholdningens inntekt er statistisk signifikant.

```
# Extracting p-value from lm.
model_02_p <- coef(summary(lm_model_02))[ , "Pr(>|t|)"]
```

```r
# Shwoing figure.
stargazer(model_02_p, type = "text", title = "P-value", align = TRUE,
          no.space = TRUE, single.row = TRUE)
```

```
P-value
========================
(Intercept)  wage   hwage
------------------------
0           0.00001   0
------------------------
```

Som en kan se fra tabellen ovenfor er p-verdien for både kvinner og menn tilnærmet lik null. Dette betyr at det er statistisk sammenheng mellom variablene. En kan også se på korrelasjonen mellom variablene for å avgjøre styrken av forholdet mellom variablene.

```r
# Extracting correlation for both sex.
model_02_cor_1 <- cor(df_02$hwage, df_02$faminc)
model_02_cor_2 <- cor(df_02$wage, df_02$faminc)

# Creating a dataframe.
cor_02_df <- data.frame(model_02_cor_1, model_02_cor_2)

# Shwoing figure.
stargazer(cor_02_df ,type = "text", title = "Correlation", align = TRUE,
          no.space = TRUE, single.row = TRUE, summary = FALSE, rownames = FALSE,
          covariate.labels = c("Men", "Women"))
```

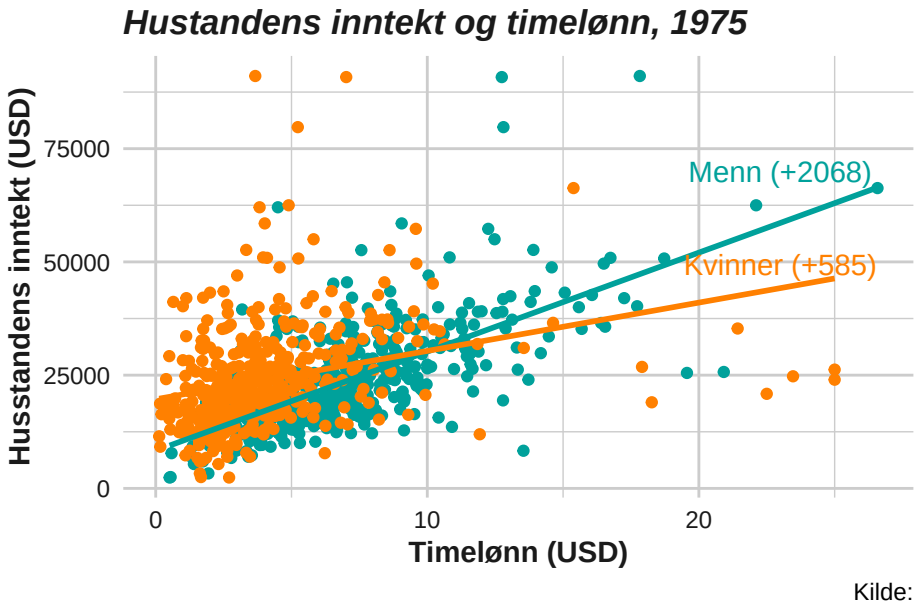
```
Correlation
===========
 Men  Women
-----------
0.669 0.303
-----------
```

Fra tabellen ser en at korrelasjonen mellom kvinner timelønn og husstandenes inntekt, dersom mannens inntekt holdes konstant er lik ca. 0.3. Dette antyder en svak positiv korrelasjon mellom variablene. Samtidig kan en se at mannens timelønn og husstandens inntekt, dersom kvinnens inntekt holdes konstant er lik ca. 0.67, noe som antyder en sterkere positiv korrelasjon.

Dette indikerer at det er større sammenheng mellom mannens inntekt og husstandens inntekt, enn kvinnens inntekt og husstandens inntekt.

Avslutningsvis kan en se på dataen grafisk. I figuren nedenfor kan en se timelønnen (dollar, 1975) på x-aksen (uavhengig variabel) og husstandens inntekt (dollar, 1975) på y-aksen (avhengig variabel) for begge kjønn i 1975. Det er også lagt inn to lineære regresjonsmodeller for kvinner (+585) og menn (+2068) som viser hvor mye husstandens inntekt øker per enhet timelønn.

```
# Creating figure.
fig_01 <- ggplot(df_02, aes(x = NULL, y = faminc)) +
  geom_point(data = df_02, aes(x = hwage), color = "#00A19B") +
  geom_point(data = df_02, aes(x = wage), color = "#FF8000") +
  geom_smooth(data = df_02, aes(x = hwage),
              method = lm, se = FALSE, color = "#00A19B") +
  geom_smooth(data = df_02, aes(x = wage),
              method = lm, se = FALSE, color = "#FF8000") +
  labs(x = "Timelønn (USD)", y = "Husstandens inntekt (USD)",
       title = "Hustandens inntekt og timelønn, 1975", caption = "Kilde:") +
  annotate("text", x = 23, y = 49500, label = "Kvinner (+585)",
           color = "#FF8000") +
  annotate("text", x = 23, y = 70000, label = "Menn (+2068)",
           color = "#00A19B") +
  theme_mod_v1()

# Shwoing figure.
fig_01
```



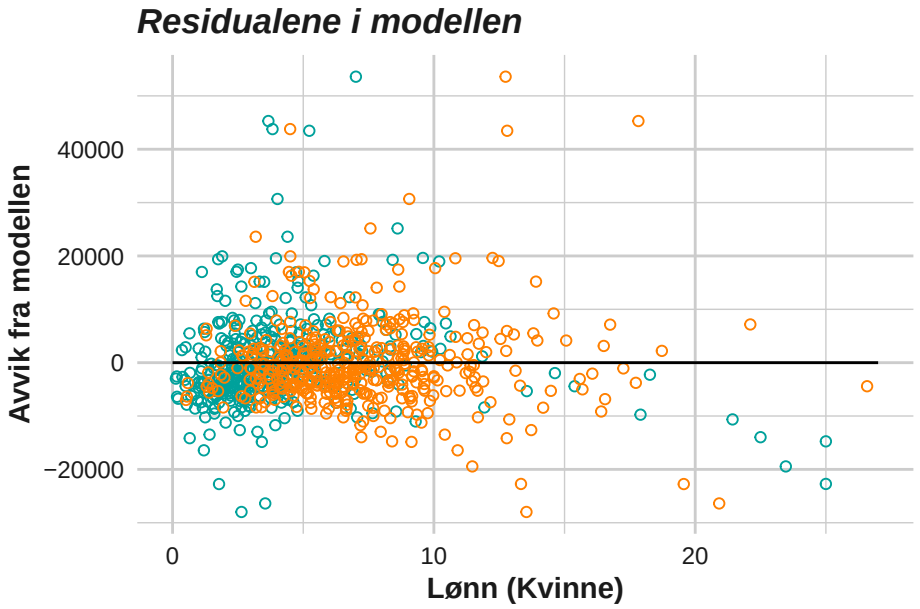
Som en kan se fra modellen ovenfor vil en enhet økning i timelønn ha større påvirkning på husstandens inntekt hos menn enn hos kvinner, etter at individet krysser em timelønn på ca. 10 USD. En kan også se at høyere timelønn fører til høyere inntekter for husstanden for begge kjønn i de fleste observasjonene, som antatt i oppgave A. Men også her kan en observerer en stor spredning i dataen. Dette kan antyde at modellen er upresis med mange ekstremverdier som vil påvirke modellen. Det vil kreve mer data og flere analyser for å kunne bekrefte årsakssammenheng.

Oppgave C)

Antakelsene for en enkelt lineær- og multippel regresjons modell er ganske lik. Det finnes flere måter og teste en multippel regresjons modell, men i denne oppgaven vil en fokusere på de samme som i forrige oppgavesett. Disse antakelsene er altså at det foreligger et lineært forhold mellom variablene, residualene skal være uavhengige og normalitet. Dersom noen av disse antakelsene ikke er oppfylt vil det kunne påvirke modellen (se oppgavesett 1), og gjøre den ugyldig.

Antakelsen om linearitet og uavhengighet kan en undersøke ved å plotte residualene og uavhengig variabel (timelønn) mot hverandre, og observere avviket fra null.

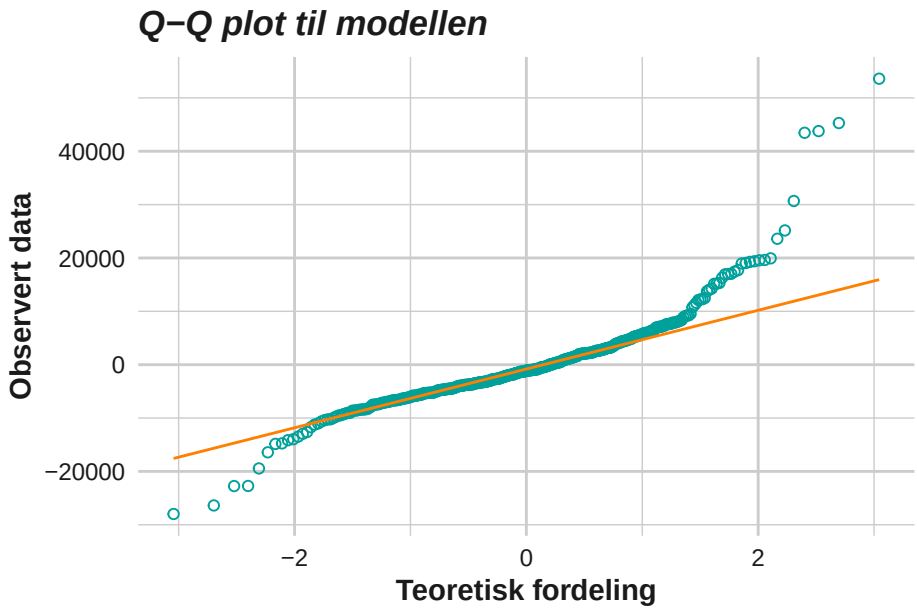
```r
# Extracting residuals from linear model.
res_02 <- resid(lm_model_02)

# Creating figure.
ggplot(data = NULL, aes(x = NULL, y = res_02)) +
  geom_point(aes(x = df_02$wage), shape = 1, color = "#00A19B") +
  geom_point(aes(x = df_02$hwage), shape = 1, color = "#FF8000") +
  geom_line(aes(x = seq(0,27), y = 0), color = "black") +
  labs(title = "Residualene i modellen",
       x = "Lønn (Kvinne)", y = "Avvik fra modellen") +
  theme_mod_v1()
```



På figuren ovenfor kan en se avviket fra modellen på y-aksen (residualene) og timelønnen på x-aksen. Ved bruk av denne figuren kan en avgjøre om modellen har et lineært forhold og om residualene er uavhengige. Fra modellen kan en ikke observere noen mønster, noe som kan indikere at antakelsen om linearitet og uavhengighet er oppfylt.

Videre kan en undersøke om modellen oppfyller antakelsen om at residualene er tilnærmet normalfordelt. Dette kan gjøres ved bruk av et Q-Q plot (forklaring oppgavesett 1):

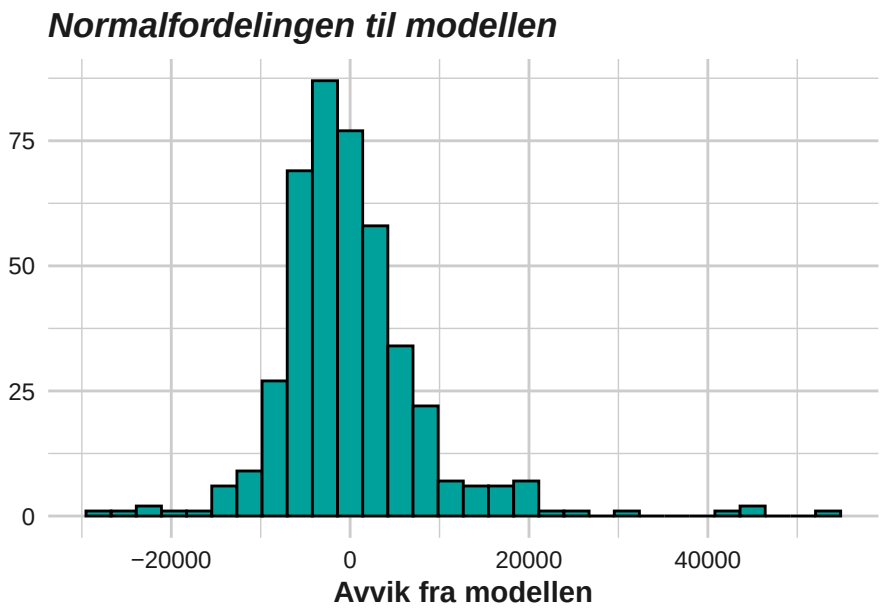
```r
# Creating dataframe.
res_df_02 <- data.frame(res_02)

# Creating Q-Q plot.
ggplot(res_df_02, aes(sample = res_02)) +
  geom_qq(shape = 1, color = "#00A19B") +
  geom_qq_line(color = "#FF8000") +
  labs(title = "Q-Q plot til modellen", x = "Teoretisk fordeling",
       y = "Observert data") +
  theme_mod_v1()
```



Fra figuren ovenfor kan en se kvantiler fra den teoretiske fordelingen på x-aksen og kvantiler for den observerte dataen på y-aksen. En observerer at både starten og slutten av figuren avviker fra den perfekte normalfordelingen (oransje linje). Fra figuren kan en anta at normalfordelingen vil være noe høyreskjev fordelt, med lange tynne «haler» på begge sider, noe som avviker fra symmetrisk normalfordeling. Dette er visualisert nedenfor:

```r
# Creating figure.
ggplot(res_df_02, aes(x = res_02)) +
  geom_histogram(fill = "#00A19B", color = "black") +
  labs(title = "Normalfordelingen til modellen",
```

```
      x = "Avvik fra modellen", y = " ") +
theme_mod_v1()
```



En kan altså ikke si at residualene er tilnærmet symmetrisk normalfordeling, noe som kan tyde på at modellen ikke er pålitelig.

Litteraturliste

Skovlund, E. (2020, 26. oktober). *Enkel lineær regresjon.* Hentet 02.11.2023 fra https://tidsskriftet.no/2020/10/medisin-og-tall/enkel-lineaer-regresjon

UiO. (I.D.). *Multippel regresjon.* Hentet 05.11.2023 fra https://www.uio.no/studier/emner/matnat/math/STK1000/h15/forelesninger/uke45.pdf

Wikipedia. (2021, 07. september). *Residual.* Hentet 02.11.2023 fra https://no.wikipedia.org/wiki/Residual

Wikipedia. (2023, 21. oktober). *Q-Q plot.* Hentet 02.11.2023 fra https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot