Oppgave 1

Vi har den lineære regressjonen $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

y_i er den avhengige variabelen.
x_i er den uavhengige variabelen.
β_0 er konstant leddet.
β_1 er koeffisienten.
ϵ_u er feilleddet/støyleddet, vi antar for en tilfeldig i så er$E(\epsilon_i) = 0$

Vi ønsker å finne y_i kurven som minimerer summen av den kvadrerte residualene (Ordinary Least squares). Vi kaller den faktisk sanne verdien av den avhengige vairabelen for y og resultat fra våres etsimator for $\hat{y}$

Vi ønsker å minimere $\sum_{i=1}^{n}(y_i - \hat{y_i})^2 = OLS$

Vi har to paramtere som vi må bestemme i $\hat{y}$, nemlig β_0 og β_1.

For å minimere disse må vi ta den partiellderiverte av både β_0 og β_1 og sette dem lik 0.

$\frac{\partial \text{OLS}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$ $\frac{\partial \text{OLS}}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$

For å vite at parameteren vi får, $\hat{\beta_0}$ og $\hat{\beta_1}$, faktisk minimerer de kvadrerte residualene sjekker vi andreordensbetingelsen.

$$Hessian(OLS) = \begin{bmatrix} \frac{\partial^2 \text{OLS}}{\partial \beta_0^2} & \frac{\partial^2 \text{OLS}}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \text{OLS}}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \text{OLS}}{\partial \beta_1^2} \end{bmatrix} \geq 0$$

Dette forsikrer oss om at Hessematrisen er positive semidefinite hvilket betyr at OLS funksjonen er konveks og dermed at vi minimerer de kvadrerte residualene.

Vi er interessert i å finne $\hat{\beta_0}$ og $\hat{\beta_1}$. Vi har fra før:

$\frac{\partial \text{OLS}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$
$\frac{\partial \text{OLS}}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$

hvilket vi kan skrive som:

$(1)\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X}$
$(2)\hat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$

Her er (1) den første momentbetingelsen og (2) den andre momentbetingelsen bare løst for β_0 og β_1 respektivt.

Vi har tre antakelser:

$1)E(u|x) = 0$
2) i.i.d, altså at utvalget er uavhengig og identiske distrbuerte til den faktiske populasjonen.
$3)0 < E(x^4) < \infty$ og $0 < E(y^4) < \infty$, X og Y har begrenset kurtosis. Med andre ord kan vi si at store ekstremverdier er usannsynlig

Hvis disse holder så finner vi:

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$
$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Når vi har $\bar{Y}$ og $\bar{X}$ kan vi enkelt finne $\hat{\beta_1}$ fra andre momentbetingelse (2) og deretter $\hat{\beta_0}$ fra første momentbetingelse (1).

Slik finner vi $\hat{Y} = \hat{B_0} + \hat{B_1}X$

Oppgave 2

a)

En dummyvariabel, også kalt indikatorvariabel, er en binær variabel som tar verdien 0 eller 1 for å indikere tilstedeværelsen av en effekt/variabel eller ikke tilstedeværelse. En dummyvariabel kan være student med 0 for nei og 1 for ja når vi undersøker inntektselastisitet i en befolkningen.

Når den uavhengige variabelen er kontinuerlig representerer koeffisienten endringen i den uavhengige variabelen på den avhengige variabelen. Det forteller oss noe om sammenhengen mellom den avhengige og uavhengige variabelen. Et eksempel på dette kan sammenhengen mellom utdanningsnivå og inntekt. Inntekt er en avhengige variabel som avhenger av den uavhengige variabelen utdanningsnivå og koeffisenten representerer endringen av et ekstra år med utdanning på inntekt.

Når den uavhengige variabelen er binær representerer koeffisenten effekten av tilstedeværelsen av den uavhengige variabel på den avhengige variabelen. Det sier noe om effekten og forksjellen mellom tilstedeværelsen av den uavhengige variabelen på den avhengige variabelen. Et eksempel på dette kan være sammenhengen mellom inntekt og hvorvidt et individ har høyere utdanning eller ikke. Inntekt er en avhengige variabel som avhenger av den uavhengige og binære variabelen for høyere utdanning. Hvis ingen høyere utdanning så er den uavhengige variabelen 0 og vice versa. Koeffisienten representerer effekten av høyere utdanning på inntekt.

Dersom den uavhengige variabelen er binær må vi lage konfidensintervall på en annen måte. For de kontinuerlige uavhengige variabelen er et konfidensintervall et mål på usikkerheten rundt verdien på koeffisenten gitt en p-verdi (Eks, p=0.05 for 95% konfidensintervall).For den binære uavhengige variabelen måler vi differeansen mellom gjennomsnittsverdien med tilstedeværelse av den uavhengige variabelen og gjennomsnittsverdien uten tilstedeværelse. Konfidensintervallet måler så usikkerheten rundt forskjellen mellom de to gjennomsnittsverdiene gitt en p-verdi.

Under hypotesetesting kan nullhypotesen H_0 og alternativhypotesen $H_a$se ganske lik ut for både en regressjon med binær - og kontinuerlig uavhengig variabel. I begge regresjonen bruker vi β som koeffisient. Vi sier at:
$H_0 : \beta = 0$
$H_a : \beta \neq 0$
Her blir tolkningen annerledes. Hypotesting for en regressjon med kontinuerlig uavhengig variabel tester om en liten endring i den uavhengige variabelen har en signifikant effekt på den avhengige variabelen. Hypotesting for en regressjon med binær uavhengig variabel tester om en det er en signifikant forskjell mellom tilstedeværelse eller ikke av den uavhenige variabel på den avhengige variabelen.

b)

Vi har fra oppgaven: $\ln(Y_i) = \beta_0 + \beta_1 D + u_i$

Vi skal tolke β_1 i denne modellen både tilnærmet og eksakt.

Når β_1 er "liten", for eksempel $\beta_1 \leq 0.2$ kan vi tolke β_1 som prosentvis endringen når vi har tatt logaritem av Y_i med tilstrekkelig nøyaktighet. Altså en $\beta_1 = 0.1$ vil da være en 10% økning i lønn dersom en er kvinne $(D = 1)$

Derimot for en eksakt tolkning av $\beta = 0.1$ må vi eksponesiere begge sider.

$Y_i = e^{0.1} \approx 1.1052$ For å finne faktisk prosentvis økning må vi trekke fra den gamlle verdien, dele på den gamle verdien, og gange med 100

$\%\Delta Y_1 = \frac{(e^{0.1}-1)}{1} * 100 = 0.1052$

Vi ser at den eksakte og tilnærmete tolkningen er veldig like for små nok verdier av β_1

c)

Vi har $Q = AL^{\alpha}K^{\beta}$

Vi transfomerer funksjonen ved å ta den naturlige logaritmen av den; $\ln(Q) = \ln(A) + \alpha\ln(L) + \beta\ln(K)$.

Vi har nå en lineær modell for parametrene α og β. Vi kan bruke lineær regresjon for å estimere α og β og konstanten $\ln(A)$ som vi kan kalle C.
$\ln(Q) = \alpha\ln(L) + \beta\ln(K) + C$

Vi kan med OLS metoden beskrevet i oppgave 1 finne de nøvendige estimatene:
$\hat{\alpha}, \hat{\beta}, \hat{C}$,slik at $ln(Q) = \hat{\alpha}\ln(L) + \hat{\beta}\ln(K) + \hat{C}$

I dette tilfellet blir $\hat{\alpha}$ den estimerte produksjonselastisiteten for arbeid L og $\hat{\beta}$ den estimerte produksjonselastisiteten for kapital K og $\hat{C}$ konstantleddet.

Anta at estimatene vi fikk ble $ln(Q) = 0.5 * ln(L) + 0.4 * ln(K)$ så vil en 1% økning i arbeid L gi 0.5% i produksjon Q og 1% økning i kapital K gi 0.4% økning i produksjon Q.

Ved flere typer forskjellige maskiner og typer kapital ville vi utvidet kapital $K = \sum_i^n k_i$ og arbeid $L = \sum_i^n l_i$ hvor det er n typer kapital og arbeid. På samme måte får vi også $\alpha = \sum_i^n \alpha_i$ og $\beta = \sum_i^n \beta_i$ slik beholder vi de samme egenskapene som originale modellen bare at nå kan vi isolere effekten av en spesifikt type kapital eller arbeid på produksjon Q.

Oppgave 3

a)

Vi har $\hat{\text{Lønn}} = 10.73 + 1.78 * \text{Mann}$ hvor

$$\text{Mann} = \begin{cases} 1, & \text{hvis mann} \\ 0, & \text{ellers} \end{cases}$$

$L\varnothing nn_{kvinner} = 10.73 + 1.78 * 0 = 10.73$, enheten er 30kr/time så vi har $10.73 * 30\text{kr/time} = 322\text{kr/time}$

$L\varnothing nn_{menn} = 10.73 + 1.78 * 1 = 12.51$, enheten er 30kr/time så vi har $12.51 * 30\text{kr/time} = 375\text{kr/time}$

Estimert lønnsforskjell er $L\varnothing nn_{menn} - L\varnothing nn_{kvinner} = 375 - 321 = 54$, det er estimert 54 kroner i timen lønnsforksjell.

Estimert prosentvis forskjell $\frac{54}{321} \approx 17\%$, det er estimert 17% forksjell i timelønn.

b)

```r
# Estimater fra regresjonen
beta_1 <- 1.78  # Koeffisienten for variabelen "Mann"
SE_beta_1 <- 0.29  # Standardfeilen for koeffisienten

# Beregn t-verdien
t_value <- beta_1 / SE_beta_1

# Finn antall observasjoner
n_menn <- 200  # Antall mannlige observasjoner
n_kvinner <- 240  # Antall kvinnelige observasjoner
df <- (n_menn + n_kvinner - 2)  # Frihetsgrader

# Beregn p-verdi
```

```r
p_value <- 2 * pt(-abs(t_value), df=df)

# Vis resultatene
cat("T-verdi:", t_value, "\n")
```

```
## T-verdi: 6.137931
```

```r
cat("Frihetsgrader:", df, "\n")
```

```
## Frihetsgrader: 438
```

```r
cat("P-verdi:", p_value, "\n")
```

```
## P-verdi: 1.874188e-09
```

```r
# Sjekk om lønnsforskjellen er signifikant på et 5% nivå
if(p_value < 0.05) {
  cat("Lønnsforskjellen er signifikant forskjellig fra 0.\n")
} else {
  cat("Lønnsforskjellen er ikke signifikant forskjellig fra 0.\n")
}
```

```
## Lønnsforskjellen er signifikant forskjellig fra 0.
```

c)

```r
beta_1 <- 1.78  # Koeffisienten for variabelen "Mann"
SE_beta_1 <- 0.29  # Standardfeilen for koeffisienten

# Beregn t-verdien
t_value <- beta_1 / SE_beta_1

# Finn antall observasjoner
n_menn <- 200  # Antall mannlige observasjoner
n_kvinner <- 240  # Antall kvinnelige observasjoner
df <- (n_menn + n_kvinner - 2)  # Frihetsgrader

# Beregn p-verdi
alpha <- 0.05
t_critical <- qt(1 - alpha/2, df)

# Konfidensintervall beregning
lower_bound <- beta_1 - t_critical * SE_beta_1
upper_bound <- beta_1 + t_critical * SE_beta_1

# Vis konfidensintervallet
cat("95% konfidensintervall for lønnsforskjellen: (", lower_bound, ",", upper_bound, ")\n")
```

```
## 95% konfidensintervall for lønnsforskjellen: ( 1.210035 , 2.349965 )
```

d)

Dersom tilstedeværelse av kvinne er 1 for den binære variabelen så må vi regne ut vår nye β_0 og β_1.

Konstant leddet β_0 blir gjennomsnittslønn for menn:
$L\varnothing nn_{menn} = 10.73 + 1.78 * 1 = 12.51$

Koeffisienten β_1 blir effekten av tilstedværelse av kvinne, det blir den negative effekten av det vi så tidligere: $L\varnothing nn_{kvinner} = 12.51 - 1.78 * 1 = 10.73$

Vi har då: $\hat{\text{Lønn}} = 12.51 - 1.78 * \text{Kvinne}$, hvor

$$\text{kvinne} = \begin{cases} 1, & \text{hvis kvinne} \\ 0, & \text{ellers} \end{cases}$$

Oppgave 4

Vi har fra oppgave teksten, korrekte spesifikasjonen $(1) : Y = \beta_0 + \beta_1 X + \beta_2 A + u$, og den spesifikasjonen vi valgte $(2) : Y = \beta_0 + \beta_1 X + u$

Ettersom vi ikke har spesifiert leddet $\beta_2 A$ vil modellen vil dette være endel av feilleddet i våres modell (2) $Y = \beta_0 + \beta_1 X + (u + \beta_2 A)$

Fra (1) kan vi skrive $A = \frac{Y - \beta_0 - \beta_1 X - u}{\beta_2}$

fra gjennomgangen våres av OLS så blir $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}}$, der
$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ er utvalgskovarians
$S_{xx} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ er utvalgsvarians

Når utvalget n går mot uendelig vil vi ha de sanne verdiene for variansen og kovariansen for populasjonen slik at:
$\lim_p \hat{\beta}_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{\text{Cov}(X, \beta_0 + \beta_1 X + \beta_2 A + u)}{\text{Var}(X)} = \frac{\beta_1 \text{Var}(X) + \beta_2 \text{Cov}(X,A)}{\text{Var}(X)} = \beta_1 + \beta_2 \frac{\text{Cov}(X,A)}{\text{Var}(X)}$, hvor

$\frac{\text{Cov}(X,A)}{\text{Var}(X)} = \frac{\text{Corr}(X,A) \cdot \text{std}(X)\text{std}(A)}{\text{std}(X)\text{std}(X)}$, slik at

$\lim_p \hat{\beta}_1 = \beta_1 + \beta_2 \cdot \text{Corr}(X, A) \cdot \frac{\text{std}(A)}{\text{std}(X)}$

En utelatt forklaringsvariabel er en variabel som påvirker den avhengige variabel og/eller uavhengig variabel som ikke er med i regressjonsmodellen. Siden denne variabel ikke er med i modellen vil estimatene av koeffisenten blir skjev. I eksempel vi har brukt blir β_1 skjev fordi vi har en utelatt variabel A. Dette betyr at vi for at upresist mål av effekten utdanning $\beta_1 X$ har på lønn Y siden vi mangler $\beta_2 A$.

Vi ser at fra: $\hat{\beta}_1 = \beta_1 + \beta_2 \cdot \text{Corr}(X, A) \cdot \frac{\text{std}(A)}{\text{std}(X)}$ at hvorvidt vi får negativ skjevhet eller positiv skjevhet avhenger av korrelasjonen mellom utdanning og evner og evner sin påvirkning på lønn. Vi kan se dette fra ligningen ved å se at fortegnet for leddet $\beta_2 \cdot \text{Corr}(X, A) \cdot \frac{\text{std}(A)}{\text{std}(X)}$ avhenger produktet $\beta_2 \cdot \text{Corr}(X, A)$ hvor:

$$\text{Fortegn på skjevheten} = \begin{cases} \text{Positiv} & \text{hvis } \beta_2 > 0 \text{ og } \text{Corr}(X, A) > 0, \\ \text{Positiv} & \text{hvis } \beta_2 < 0 \text{ og } \text{Corr}(X, A) < 0, \\ \text{Negativ} & \text{hvis } \beta_2 > 0 \text{ og } \text{Corr}(X, A) < 0, \\ \text{Negativ} & \text{hvis } \beta_2 < 0 \text{ og } \text{Corr}(X, A) > 0. \end{cases}$$

Oppgave 5

```r
# Skalar
s <- 5

# Rekkevektor
r_vec <- c(1, 2, 3)

# Kolonnevektor
```

```r
c_vec <- matrix(c(1, 2, 3), nrow = 3, ncol = 1)

# Matrise
A <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 3, ncol = 2)

# Printing
print(s)
```

```
## [1] 5
```

```r
print(r_vec)
```

```
## [1] 1 2 3
```

```r
print(c_vec)
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
```

```r
print(A)
```

```
##      [,1] [,2]
## [1,]    1    4
## [2,]    2    5
## [3,]    3    6
```

```r
# Transpose
t_r_vec <- t(matrix(r_vec, nrow = 1))
t_A <- t(A)

# Printing
print(t_r_vec)
```

```
##      [,1]
## [1,]    1
## [2,]    2
## [3,]    3
```

```r
print(t_A)
```

```
##      [,1] [,2] [,3]
## [1,]    1    2    3
## [2,]    4    5    6
```

```r
# Dimensjon
dim_A <- dim(A)
print(dim_A)
```

```
## [1] 3 2
```

```r
# Matrise A
A <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 3, ncol = 2)

# Matrise B
B <- matrix(c(6, 5, 4, 3, 2, 1), nrow = 3, ncol = 2)

# Addering
```

```r
C <- A + B

# Printing
print(C)
```

```
##      [,1] [,2]
## [1,]    7    7
## [2,]    7    7
## [3,]    7    7
```

```r
# Skalar
s <- 3

# Multiplikasjon
D <- A * s

# Printing
print(D)
```

```
##      [,1] [,2]
## [1,]    3   12
## [2,]    6   15
## [3,]    9   18
```

```r
print("Multiplikasjon av en (5x2) matrise med en (2x4) matrise gir en matrise med dimensjonen 5×4.")
```

```
## [1] "Multiplikasjon av en (5x2) matrise med en (2x4) matrise gir en matrise med dimensjonen 5×4."
```

```r
# Matrise A (3x2)
A <- matrix(c(1, 2, 3, 4, 5, 6), nrow = 3, ncol = 2)

# Matrise B (2x2)
B <- matrix(c(7, 8, 9, 10), nrow = 2, ncol = 2)

# Multiplikasjon
E <- A %*% B

# Printing
print(E)
```

```
##      [,1] [,2]
## [1,]   39   49
## [2,]   54   68
## [3,]   69   87
```

```r
# Matrise A (2x2)
A <- matrix(c(4, 7, 2, 6), nrow = 2, ncol = 2)

# Invers
A_inv <- solve(A)

# Printing
print(A_inv)
```

```
##      [,1] [,2]
## [1,]  0.6 -0.2
## [2,] -0.7  0.4
```

```r
# Simuler data
n <- 100
k <- 3
X <- matrix(rnorm(n * k), n, k)
beta_true <- c(1, 0.5, -0.5)
epsilon <- rnorm(n)

Y <- X %*% beta_true + epsilon

# OLS-estimator
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% Y

# Printing result
print(beta_hat)
```

```
##            [,1]
## [1,]  0.9181043
## [2,]  0.4844285
## [3,] -0.5076022
```

```r
# Lag en 10x3 matrise
set.seed(123)
A <- matrix(rnorm(30), nrow = 10, ncol = 3)

# Print matrisen
print(A)
```

```
##              [,1]       [,2]       [,3]
##  [1,] -0.56047565  1.2240818 -1.0678237
##  [2,] -0.23017749  0.3598138 -0.2179749
##  [3,]  1.55870831  0.4007715 -1.0260044
##  [4,]  0.07050839  0.1106827 -0.7288912
##  [5,]  0.12928774 -0.5558411 -0.6250393
##  [6,]  1.71506499  1.7869131 -1.6866933
##  [7,]  0.46091621  0.4978505  0.8377870
##  [8,] -1.26506123 -1.9666172  0.1533731
##  [9,] -0.68685285  0.7013559 -1.1381369
## [10,] -0.44566197 -0.4727914  1.2538149
```

```r
# Beregn varians-kovarians-matrisen
cov_matrix <- cov(A)

# Print varians-kovarians-matrisen
print(cov_matrix)
```

```
##            [,1]       [,2]       [,3]
## [1,]  0.9097040  0.5718955 -0.3604070
## [2,]  0.5718955  1.0775964 -0.5481997
## [3,] -0.3604070 -0.5481997  0.8664058
```