STK4900 V24 | Obligatorisk innlevering 1 | Adrian Duric

Oppgave 1

a)

For å finne det forventede antallet menn høyere enn 2 meter, finner vi først sannsynligheten for at en tilfeldig trukket mann er >200cm høy. Så ganger vi det med antallet samples (1 million) for å få forventet antall.

```
mu_w = 168
sigma_w = 6.5
mu_m = 181
sigma_m = 8

# Finner sannsynlighet for å trekke en mann over 200cm
prob = 1 - pnorm(200, mean=mu_m, sd=sigma_m)

# Finner forventet antall menn over 200cm per 1M menn
prob = prob * 1000000
cat("Forventet antall menn over 200cm per 1 million menn:", prob)
```

```
## Forventet antall menn over 200cm per 1 million menn: 8774.475
```

b)

Regner ut de kumulative sannsynlighetene $F(155.26)$ og $F(180.74)$ for kvinnenes normalfordeling:

```
cat("Sannsynlighet for å trekke 1 kvinne med høyde <= 155.26cm:", pnorm(155.26, mean=mu_w, sd=sigma_w),
```

```
## Sannsynlighet for å trekke 1 kvinne med høyde <= 155.26cm: 0.0249979
```

```
cat("Sannsynlighet for å trekke 1 kvinne med høyde <= 180.74cm:", pnorm(180.74, mean=mu_w, sd=sigma_w))
```

```
## Sannsynlighet for å trekke 1 kvinne med høyde <= 180.74cm: 0.9750021
```

Fra utskriften ser vi at de to verdiene gir oss nesten nøyaktig 2.5%- og 97.5%-persentilene for høyden til kvinner på øya. Det vil si at 95% av kvinner samplet fra fordelingen vil være mellom disse to verdiene.

Vi bruker at $P(a \leq X \leq b) = F(b) - F(a)$ for å finne sannsynligheten for at én tilfeldig trukket kvinnes høyde er innenfor intervallet $[155.26, 180.74]$. Det er rimelig å anta at de 10 tilfeldig trukkede kvinnenes høyder er uavhengige av hverandre, så vi opphøyer sannsynligheten med antall sampler (n=10) for å finne sannsynligheten for at alle de 10 er innenfor intervallet.

```
# Regner sannsynlighet for at en tilfeldig trukket kvinne har høyde innenfor intervallet
prob = pnorm(180.74, mean=mu_w, sd=sigma_w) - pnorm(155.26, mean=mu_w, sd=sigma_w)

# Regner sannsynlighet for å trekke 10 kvinner innenfor intervallet
prob = prob^10
cat("Sannsynlighet for å trekke 10 kvinner med høyde i [155.26, 180.74]:", prob)
```

```
## Sannsynlighet for å trekke 10 kvinner med høyde i [155.26, 180.74]: 0.5987635
```

c)

Vi bruker at $E(a + \sum_{i=1}^{n} b_i X_i) = a + \sum_{i=1}^{n} b_i E(X_i)$ og at $V(a + \sum_{i=1}^{n} b_i X_i) = \sum_{i=1}^{n} b_i^2 V(X_i)$ i det følgende:

Vi definerer kvinnens høyde som K og mannens som M, og antar at disse er uavhengige. Så regner vi ut $P(K > M)$ eller $P(Z > 0)$, hvor $Z = K - M$. Fordi vi har middelverdi og standardavvik for kvinner og menn, kan vi finne de samme verdiene for Z.

```
# Finner middelverdi og standardavvik for differanse mellom kvinner og menn
mu_z = mu_w - mu_m
sigma_z = sqrt(sigma_w^2 + sigma_m^2)

# Finner sannsynligheten for at kvinnens høyde er større enn mannens, dvs. P(Z > 0)
prob = 1 - pnorm(0, mean=mu_z, sd=sigma_z)
cat("Sannsynlighet for at kvinnen er høyere enn mannen:", prob)
```

```
## Sannsynlighet for at kvinnen er høyere enn mannen: 0.1036211
```

d)

Siden vi vet at alle samplene er enten bare menn eller bare kvinner, kan vi uttrykke null- og den alternative hypotesen som H_0 : alle de 25 menneskene er kvinner, og H_A : alle de 25 menneskene er menn. Å bestemme hvordan disse setningene skal uttrykkes matematisk, er derimot ikke helt rett frem.

Forsøker først å uttrykke nullhypotesen som $H_0 : \overline{x} = \mu_w$:

```
n = 25
x_avg = 172

# Beregner z-statistikk
z = (x_avg - mu_w) / (sigma_w / sqrt(n))

# Beregner den ensidede P-verdien P(Z > z) = 1 - P(Z <= z)
cat("P-verdi (H0: alle de 25 menneskene er kvinner):", 1 - pnorm(z))
```

```
## P-verdi (H0: alle de 25 menneskene er kvinner): 0.001045746
```

Dersom alle de 25 menneskene var bare kvinner, var altså $\overline{x} = 172cm$ en veldig usannsynlig måling (ca 0.1% sanns. for at den ville forekommet). Hadde man bare hatt informasjon om tetthetsfordelingen til kvinner, ville det vært naturlig å forkaste nullhypotesen og konkludere med den alternative hypotesen om at alle er menn. Men det virker ikke helt intuitivt, siden $\overline{x}$ jo ligger nærmere μ_w enn μ_m.

Vi ser dette tydeligere ved å gjøre ting andre veien, altså anta nullhypotesen $H_0 : \overline{x} = \mu_m$:

```
n = 25
x_avg = 172

# Beregner z-statistikk
z = (x_avg - mu_m) / (sigma_m / sqrt(n))

# Beregner den ensidede P-verdien P(Z <= z)
cat("P-verdi (H0: alle de 25 menneskene er menn):", pnorm(z))
```

```
## P-verdi (H0: alle de 25 menneskene er menn): 9.275399e-09
```

Her får vi en P-verdi som er flere størrelsesordener lavere enn tidligere, og som derfor også burde forkastes. Men om begge nullhypotesene forkastes, ville ikke oppgavens vilkår vært innfridd; vi vet jo at alle de 25 menneskene er enten bare menn eller bare kvinner. Altså må $\overline{x}$ "stamme fra" enten μ_w eller μ_m, gitt at menneskene er samplet fra en av de gitte normalfordelingene for enten menn eller kvinner alle sammen.

Det vi uansett kan konkludere med, er at den gitte målingen av $\overline{x}$ var veldig usannsynlig å forekomme uansett hvilket kjønn de 25 menneskene har. Likevel leser vi fra størrelsesordenen på til P-verdiene at det er betydelig mer sannsynlig at de er kvinner enn at de er menn. Eksempelet viser også at hvilken informasjon man har er relevant for hvordan man skal tolke en P-verdi (hadde vi ikke visst at alle er av samme kjønn, hadde det vært lett å si at gruppen på 25 personer mest sannsynlig består av både menn og kvinner).

Oppgave 2

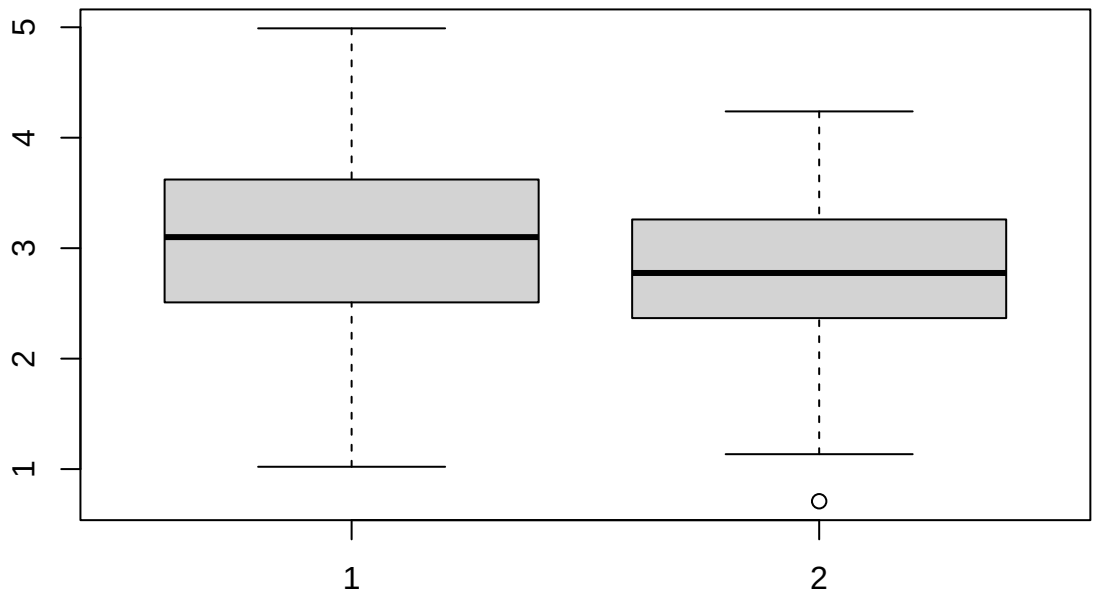
a)

Lager boxplots for røykende og ikke-røykende mødre:

```r
# Forbereder data
babies = matrix(scan("mothersbabies.txt",skip=7),byrow=T,ncol=7)
y = babies[ ,1] # birthweight, in kg
x1 = babies[ ,2] # mother's weight prior to pregnancy
x2 = babies[ ,3] # age
x3 = babies[ ,4] # indicator for smoking (1 for yes, 0 for no)
x4 = babies[ ,5] # indicator for ethnic 1, black
x5 = babies[ ,6] # indicator for ethnic 2, neither white nor black
x6 = babies[ ,7] # indicator for ethnic 0, white

ysmoke = y[x3 == 1]
ynosmoke = y[x3 == 0]

# Lager boxplots
boxplot(ynosmoke, ysmoke)
```



Fra boxplottene ser vi antydninger til at vekten til babyen generelt sett er høyere for ikkerøykende mødre enn for røykende; medianvekten er høyere, samt de resterende kvartilverdiene og spesielt maksimalverdien.

b)

For å kunne utføre t-testen antar vi at data fra de to gruppene er samplet fra normalfordelinger med potensielt forskjellige middelverdier og standardavvik. Vi antar at de sanne middelverdiene og standardavvikene ikke er kjente, men estimeres ut fra dataene.

```
# Gjennomfører t-test
t.test(ysmoke, ynosmoke)
```

```
##
##  Welch Two Sample t-test
##
## data:  ysmoke and ynosmoke
## t = -2.7095, df = 170, p-value = 0.00743
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.48695979 -0.07646677
## sample estimates:
## mean of x mean of y
##  2.773243  3.054957
```

Vi leser fra t-testen at P-verdien blir 0.00743, som er lavt (under 95% konfidensnivå). Vi kan derfor forkaste

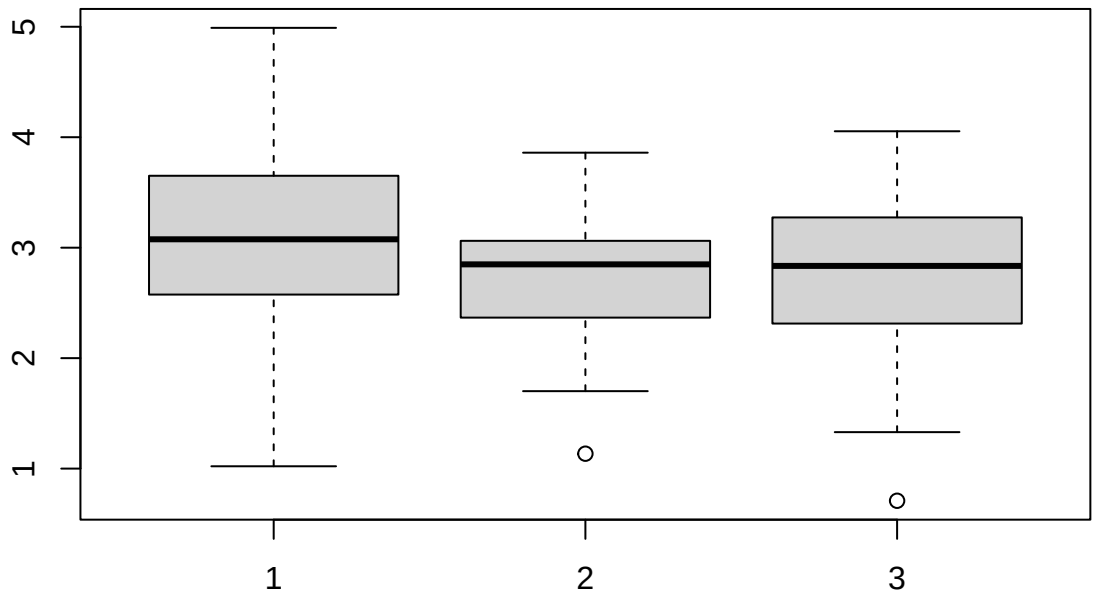
nullhypotesen, og heller gå ut fra at det er en forskjell i gjennomsnittlig babyvekt mellom mødre som røyker og ikke. Vi leser også fra utskriften at konfidensintervallet er $[-0.48695979, -0.07646677]$.

c)

Lager boxplots for de tre gruppene:

```
# Sorterer ut gruppene
yethnic0 = y[x6 == 1]
yethnic1 = y[x4 == 1]
yethnic2 = y[x5 == 1]

# Lager boxplots
boxplot(yethnic0, yethnic1, yethnic2)
```



Gjennomfører ANOVA ved å først lage en vektor som inkluderer alle de 3 kategoriske variablene, og deretter bruke `aov`. Vi antar at data fra de tre kategoriene er samplet for normalfordelinger med ukjente og potensielt ulike middelverdier.

```
# Lager kategorisk vektor som kombinerer de tre gruppene
x_ethnic = ifelse(x4 == 1, 1, 0)
x_ethnic = ifelse(x5 == 1, 2, x_ethnic)
x_ethnic = factor(x_ethnic)

# Gjennomfører ANOVA
```

```r
aov_ethnic = aov(y~x_ethnic)
summary(aov_ethnic)
```

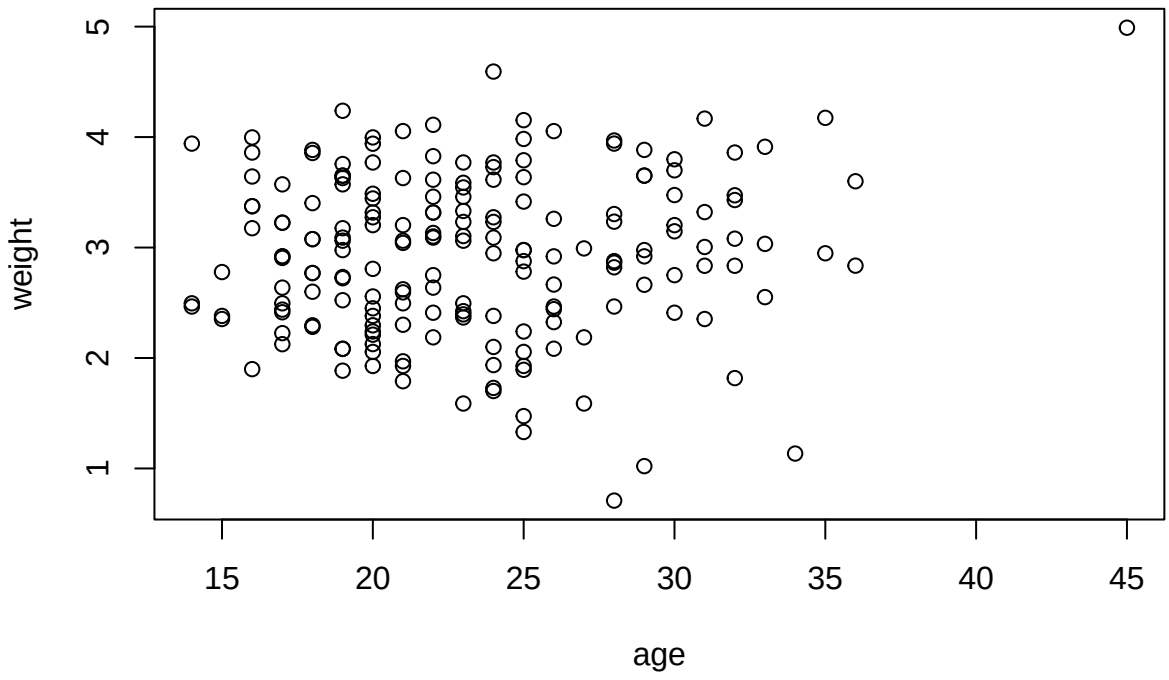
```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## x_ethnic      2   5.07  2.5353   4.972 0.00788 **
## Residuals   186  94.85  0.5099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vi finner at P-verdien for den funnede F-scoren er så lav at vi kan forkaste nullhypotesen ($P = 0.00788$), og heller konkludere med den alternative hypotesen, altså at den gjennomsnittlige fødevekten for de tre gruppene er forskjellige.

d)

Plotter alder på moren mot vekt på barnet:

```r
# Plotter alder vs. fødselsvekt
plot(x2, y, xlab="age", ylab="weight")
```



Regner korrelasjon mellom de to variablene, og regner ut konfidensintervall:

```r
# Regner ut korrelasjon
cor.test(x2, y)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x2 and y
## t = 1.2339, df = 187, p-value = 0.2188
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.05355077  0.22965101
## sample estimates:
##       cor
## 0.08986639
```

Fra utskriften leser vi at korrelasjonen er på ca. 9%, altså er en eventuell korrelasjon svak. Siden 95%-konfidensintervallet $[-0.05355077, 0.22965101]$ inneholder 0, kan vi ikke si med høy konfidens at det er en korrelasjon mellom de to variablene i det hele tatt, slik også den relativt høye P-verdien forteller oss.

e)

Lager regresjonsmodellen:

```r
# Lager regresjonsmodell
fit = lm(y ~ x1 + x2 + x3)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.06992 -0.43322  0.01365  0.51641  1.81376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.362714   0.300690   7.858 3.11e-13 ***
## x1           0.008860   0.003791   2.337   0.0205 *
## x2           0.007094   0.009925   0.715   0.4757
## x3          -0.267215   0.105802  -2.526   0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7088 on 185 degrees of freedom
## Multiple R-squared:  0.06988,    Adjusted R-squared:  0.05479
## F-statistic: 4.633 on 3 and 185 DF,  p-value: 0.003781
```

Siden vi bruker mer enn 1 variabel, leser jeg av for den justerte R-verdien: `Adjusted R-squared:  0.05479`. Den er lav, og kan tolkes som at kun ca. 5,5% av variasjonen i y (fødselsvekt) forklares av x_1, x_2 og x_3 (mors vekt før fødselen, alder og røyking).

Beregner konfidensintervaller for β_1, β_2 og β_3:

```r
# Fyller inn formelverdier
beta1 = 0.008860
beta2 = 0.007094
beta3 = -0.267215
se_beta1 = 0.003791
se_beta2 = 0.009925
se_beta3 = 0.105802

cval = qt(0.975, df=length(y) - 4)

# Beregner konfidensintervaller
beta1_conf = c(beta1 - cval*se_beta1, beta1 + cval*se_beta1)
beta2_conf = c(beta2 - cval*se_beta2, beta2 + cval*se_beta2)
beta3_conf = c(beta3 - cval*se_beta3, beta3 + cval*se_beta3)

cat("95% konfidensintervall for beta1: (", beta1_conf, ")\n")
```

```
## 95% konfidensintervall for beta1: ( 0.00138085 0.01633915 )
```

```r
cat("95% konfidensintervall for beta2: (", beta2_conf, ")\n")
```

```
## 95% konfidensintervall for beta2: ( -0.01248673 0.02667473 )
```

```r
cat("95% konfidensintervall for beta3: (", beta3_conf, ")\n")
```

```
## 95% konfidensintervall for beta3: ( -0.4759486 -0.05848141 )
```

Som vi også kunne tolke fra å se på P-verdiene i oppsummeringen over, ser vi her at b_2 er den eneste koeffisienten med et konfidensintervall som inneholder 0. Det vil si at vi ikke kan fastslå at alder har noen virkning på fødselsvekten, slik vi også så i d).

f)

Lager regresjonsmodellen:

```r
# Lager regresjonsmodell
fit = lm(y ~ x1 + x2 + x3 + x4 + x5)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28181 -0.44736  0.02215  0.47229  1.74780
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.841953   0.321337   8.844 7.67e-16 ***
```

```
## x1           0.008816   0.003830   2.302 0.022480 *
## x2          -0.002036   0.009817  -0.207 0.835920
## x3          -0.400328   0.109207  -3.666 0.000323 ***
## x4          -0.511537   0.157028  -3.258 0.001339 **
## x5          -0.400608   0.119542  -3.351 0.000977 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6819 on 183 degrees of freedom
## Multiple R-squared:  0.1484, Adjusted R-squared:  0.1251
## F-statistic: 6.377 on 5 and 183 DF,  p-value: 1.744e-05
```

Fra å legge til x_4 og x_5 som variabler, ser vi at R-verdien har økt, som vil si at etnisiteten til mødrene også bidrar til å forklare variasjonen i barnets fødselsvekt. Vi ser også at de to koeffisientene β_4 og β_5 har lave P-verdier, ergo kan man med høy konfidens fastslå at etnisiteten til moren påvirker fødselsvekten til barnet, slik vi kunne forvente fra resultatene i c).

g)

Resultater for hvit etnisitet:

```
# Lager regresjonsmodell for hvit etnisitet
fit = lm(yethnic0 ~ x3[x6 == 1])
summary(fit)
```

```
##
## Call:
## lm(formula = yethnic0 ~ x3[x6 == 1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.40775 -0.41573  0.03775  0.44425  1.56125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4288     0.1004  34.135  < 2e-16 ***
## x3[x6 == 1]  -0.6000     0.1365  -4.396 2.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6663 on 94 degrees of freedom
## Multiple R-squared:  0.1706, Adjusted R-squared:  0.1617
## F-statistic: 19.33 on 1 and 94 DF,  p-value: 2.895e-05
```

Resultater for svart etnisitet:

```
# Lager regresjonsmodell for svart etnisitet
fit = lm(yethnic1 ~ x3[x4 == 1])
summary(fit)
```

```
##
## Call:
```

```
## lm(formula = yethnic1 ~ x3[x4 == 1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3690 -0.3525 -0.0055  0.4579  1.0055
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.8545     0.1568  18.204  1.5e-15 ***
## x3[x4 == 1]  -0.3505     0.2528  -1.386    0.178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6272 on 24 degrees of freedom
## Multiple R-squared:  0.07413,    Adjusted R-squared:  0.03556
## F-statistic: 1.922 on 1 and 24 DF,  p-value: 0.1784
```

Resultater for verken hvit eller svart etnisitet:

```
# Lager regresjonsmodell for verken hvit eller svart etnisitet
fit = lm(yethnic2 ~ x3[x5 == 1])
summary(fit)
```

```
##
## Call:
## lm(formula = yethnic2 ~ x3[x5 == 1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.04817 -0.50124  0.02076  0.48130  1.23976
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.81424    0.09796  28.729   <2e-16 ***
## x3[x5 == 1] -0.05707    0.23147  -0.247    0.806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7265 on 65 degrees of freedom
## Multiple R-squared:  0.0009343,  Adjusted R-squared:  -0.01444
## F-statistic: 0.06079 on 1 and 65 DF,  p-value: 0.806
```

Fra utskriftene og resultatene fra hypotesetestene for β_3 i hvert spesifikke tilfelle, ser man at det kun er for kvinner av hvit etnisitet at man har et statistisk signifikant resultat. Det vil si at man kan si med høy konfidens at røyking påvirker fødselsvekten for hvite mødre som røyker (den synker, som vi kan lese av den estimerte koeffisienten). Også svarte og verken svarte eller hvite mødre har β_3-koeffisienter som er estimerte til å være negative, som altså tilsier at røyking senker fødselsvekten, men fordi disse koeffisientene har høye tilhørende P-verdier kan man ikke konkludere med at denne effekten nødvendigvis er reell for disse etnisitetene.

h)

Predikerer på den fulle regresjonsmodellen (som inkluderer alle prediktorene):

```r
# Lager regresjonsmodell og predikerer på den
fit = lm(y ~ x1 + x2 + x3 + x4 + x5)
pred = predict(fit, newdata=data.frame(x1=60, x2=25, x3=1, x4=0, x5=0))
cat("Forventet fødselsvekt for babyen til Mrs. Jones:", pred, "kg")
```

```
## Forventet fødselsvekt for babyen til Mrs. Jones: 2.919708 kg
```

```r
# Lager regresjonsmodell og predikerer på den
fit = lm(y ~ x1 + x2 + x3 + x4 + x5)
pred = predict(fit, newdata=data.frame(x1=60, x2=25, x3=0, x4=0, x5=0))
cat("Forventet fødselsvekt for babyen til Mrs. Smith:", pred, "kg")
```

```
## Forventet fødselsvekt for babyen til Mrs. Smith: 3.320036 kg
```