SOK-2009 Eksamen

Kandidatnummer 96

Oppgave 1

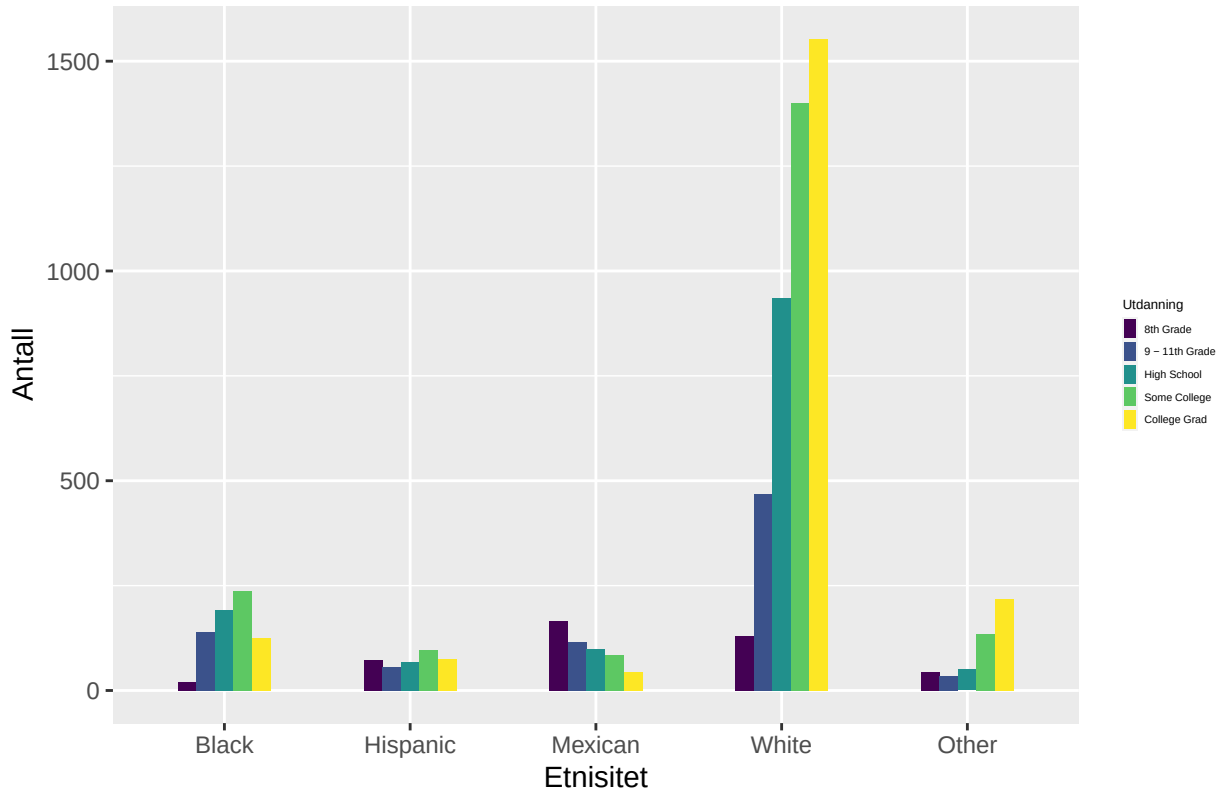
1a

Målenivået for *race1* er nominalnivå og brukes for å kategorisere observasjoner, ikke rangere. Variablene fungerer som kategorier så rekkefølgen på verdiene er irrelevant. Det er ingen nullverdi.

Målenivået for *education* er ordinalnivå og kan rangeres på en skala i forhold til hverandre. De forskjellige nivåene har derimot ingen betydning. I dette tilfellet er det fem nivåer med utdanning registrert, ingen av de er en bestemt størrelse større enn noen andre. Det er derimot et klart rangering der den minste mengden utdanning registrert er 8. klasse, etterfulgt av 9. til 11. klasse, videregående, noe høyere utdanning og til slutt gjennomført høyere utdanning.

1b

Figur 1: Fordeling av utdanningsnivå mellom etnisiteter



Fra figuren kan en se at det er en veldig stor andel av observasjonene som er i gruppen *White*. Dette gjør at *white* har størst antall observasjoner for alle utdanningsnivåer. De fire andre kategoriene er mer lik hverandre i antall.

Gruppen *black* har flest med Some College. Gruppen *hispanic* har omtrent samme antall observasjoner per utdanning. *Mexican* har flest med 8. klasse utdanning, og antallet faller for hvert nivå ekstra med utdanning. *White* har flest College Grad, og antallet øker for hver økning i utdanningsnivå. *Other* har også flest College Grad.

Bare fra grafen kan det virke som de fleste med høy utdanning (Some College og College Grad) er av etnisiteten *white* eller *oter*.

1c

	8th Grade	9 - 11th Grade	High School	Some College	College Grad
Black	21	140	192	237	125
Hispanic	72	57	67	96	74
Mexican	166	115	100	84	44
White	131	468	936	1400	1554
Other	45	34	50	134	219

1d

For å analysere om det er sammenheng mellom utdanningsnivå og etnisitet benyttes en kjei-kvadrat test. En kjei-kvadrat test vil både gi en p-verdi og forventet observasjoner. Deretter kan en se på differansen mellom faktiske observasjoner og forventet.

H_0: Det er ingen sammenheng mellom utdanningsnivå og etnisitet.

H_1: Det er sammenheng mellom utdanningsnivå og etnisitet, utdanning påvirker etnisitet eller etnisitet påvirker utdanning på en eller annen måte.

1e

Det er satt et signifikansnivå *(a)* på 1%, det vil si at en p-verdi lavere enn 0,01 gjør at H_0 forkastes og H_1 er gjeldende.

```
##
##  Pearson's Chi-squared test
##
## data:  Etni_utd_tabell
## X-squared = 1094.4, df = 16, p-value < 2.2e-16
```

Kji-kvadrattesten gjennomføres. Den gir en p-verdi lavere enn signifikansnivået, og H_0 forkastes. Det betyr at det er et signifikant sammenheng mellom utdanning og etnisitet.

	8th Grade	9 - 11th Grade	High School	Some College	College Grad
Black	21	140	192	237	125
Hispanic	72	57	67	96	74
Mexican	166	115	100	84	44
White	131	468	936	1400	1554

	8th Grade	9 - 11th Grade	High School	Some College	College Grad
Other	45	34	50	134	219

Fra den originale tabellen kan en observere at *White* utgjør størst andel i alle utdanningsnivåer, utenom *8th Grane* som er dominert av *Mexican*. De andre gruppene har omtrent like mange observasjoner i hvert utdanningsnivå.

	8th Grade	9 - 11th Grade	High School	Some College	College Grad
Black	47.40512	88.70751	146.57446	212.6147	219.6982
Hispanic	24.26612	45.40832	75.02972	108.8349	112.4609
Mexican	33.74714	63.14982	104.34461	151.3579	156.4005
White	297.62460	556.93431	920.24158	1334.8634	1379.3361
Other	31.95702	59.80003	98.80963	143.3291	148.1043

Den forventede tabellen viser hvordan fordelingen forventes å være.

	8th Grade	9 - 11th Grade	High School	Some College	College Grad
Black	-26.40512	51.29249	45.425545	24.385307	-94.69822
Hispanic	47.73388	11.59168	-8.029721	-12.834934	-38.46091
Mexican	132.25286	51.85018	-4.344612	-67.357872	-112.40055
White	-166.62460	-88.93431	15.758421	65.136565	174.66392
Other	13.04298	-25.80003	-48.809633	-9.329066	70.89575

Den siste tabellen viser differansen mellom den observerte tabellen og den forventede tabellen. Her kan en se at det er *White* ser ut til å påvirke resultatet mest. Den største differansen er fra *White* og *College Grad*, og for alle utdanningsnivåer er det *White* som står for den største differansen fra forventet tabell, utenom *High School* der *Other* påvirker mest og *Some College* der *Mexican* såvidt påvirker mer enn *White*. Alt i alt er store differanser på noen av observasjonene fra den forventede til den faktiske tabellen.

Oppgave 2

2a

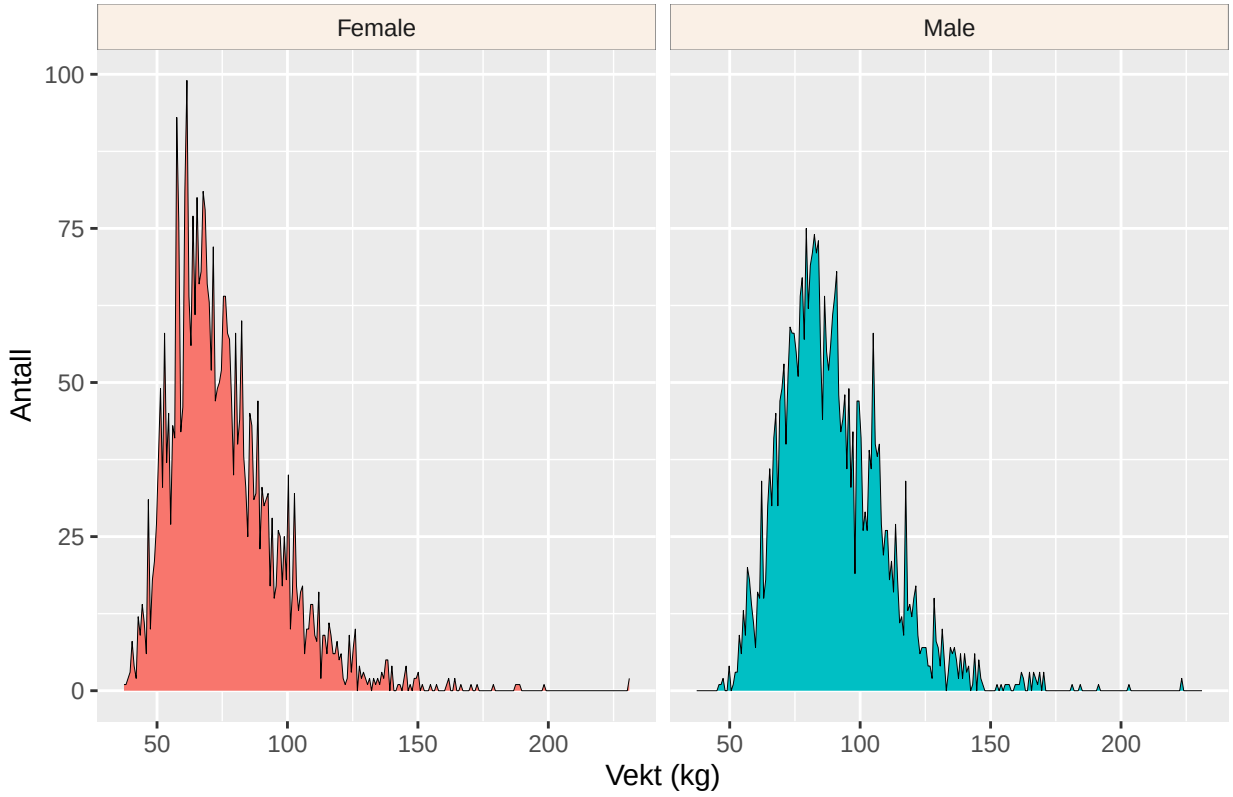
Målenivået for *gender* er nominalnivå. Det er dikotome variabler, med bare to alternativer. Det brukes for å kategorisere, ikke rangere.

Målenivået for *height* er forholdstallnivå. Det eksisterer et nullpunkt og verdiene kan rangeres. Det er mulig å si noe om avstanden mellom hver verdi. Det er også enkelt å konvertere enheter. En person som er 180 cm høy er også 1,8 m eller 0,0018 km.

Målenivået for *weight* er også forholdstall. Samme gjelder her, det eksisterer et nullpunkt og verdiene kan enkelt rangeres. Hver verdi har en objektiv avstand mellom hverandre. Samme som med høyde så kan enheter lett konverteres. En person som veier 80 kg veier også 80000eg eller 176,4 lb.

2b

Figur 2: Forskjell i vekt mellom kjønn

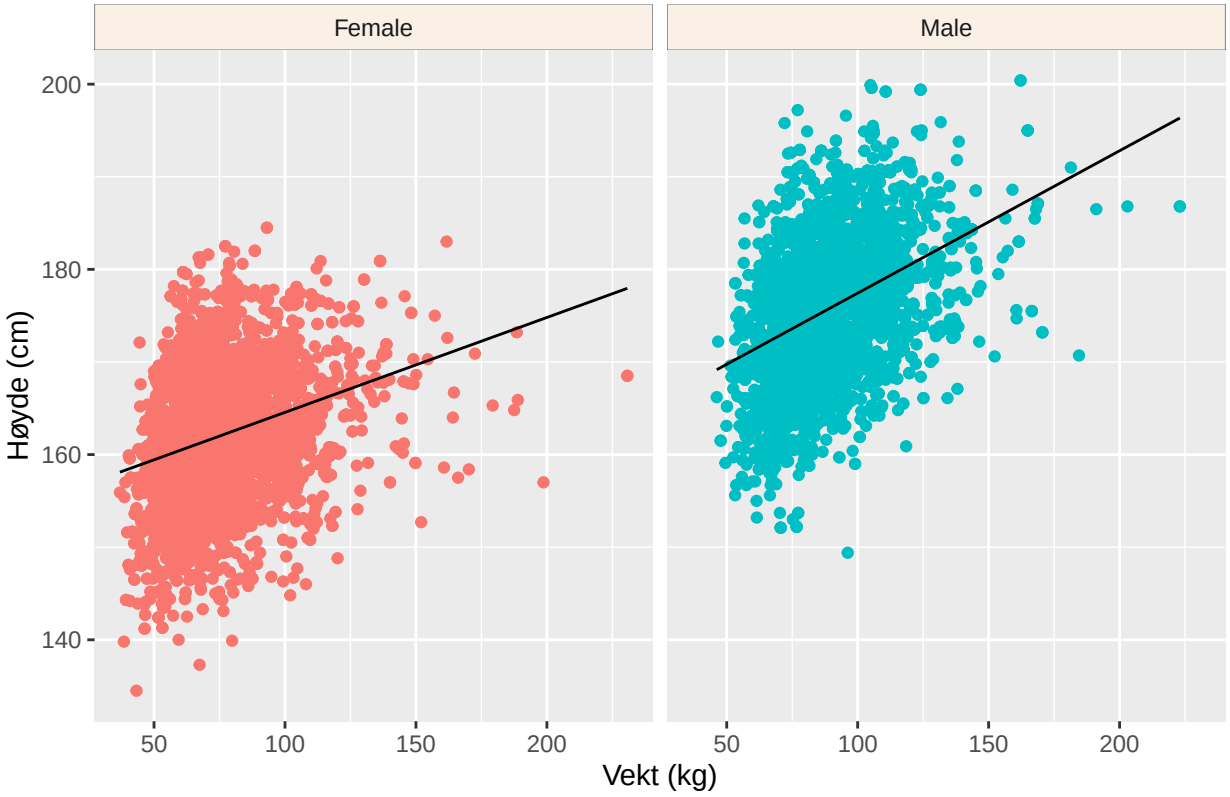


Blant begge kjønn er de fleste observasjonene i vekt mellom 50 kg og 100 kg. Veldig få av begge kjønn veier mindre enn 50 kg. Det er flere menn som veier mer enn 100 kg.

Grafene har samme struktur rundt de samme verdiene, men det er så klart små forskjeller. Veldig få kvinner veier mer enn 150 kg, mens et større antall menn veier over 150 kg.

2c

Figur 3: Forskjell i vekt og høyde mellom kjønn



Fra figurene kan en se at det fremdeles er flest observasjoner i mellom 50 kg og 100 kg, men i dette tilfellet når høyde er lagt inn kan en observere at observasjonene av menn er noe høyere enn kvinner. Trendlinjen til menn er brattere enn kvinner, noe som tyder på at ved høyere vekt øker også høyden for menn mer.

2d

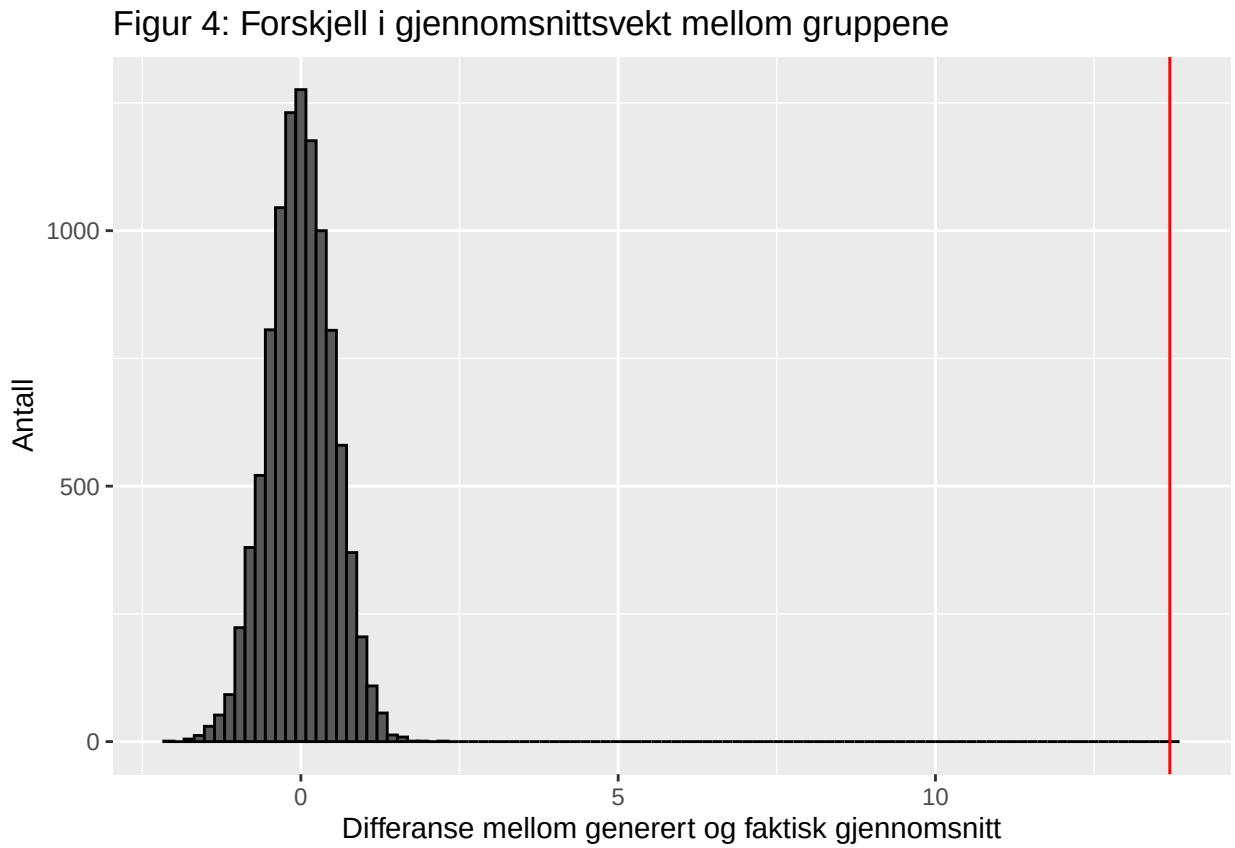
gender	me_height	me_weight	st.dev_height	st.dev_weight	st.err_height	st.err_weight	antall
female	162.0497	75.51715	7.294981	20.43694	0.1206648	0.3380431	3655
male	175.7777	89.20989	7.481865	19.72010	0.1261606	0.3325239	3517

2e

i

H_0: det er ingen sammenheng mellom gjennomsnittsvekt og kjønn H_1: det er sammenheng mellom gjennomsnittsvekt og kjønn. De to variablene påvirker hverandre på en eller annen måte.

ii



Dette er en permutasjonstest via "just one test" fremgangsmåten. Dette kalles også for en todimensjonal eller bivariat analyse. Det brukes for å se på hvordan to variabler forholder seg til hverandre. Fra grafen kan en se at differansen i generert gjennomsnitt er mye lavere enn det faktiske gjennomsnittet.

iii

p_value
0

P-verdien fra testen er 0. Den er tilnærmet null men i virkeligheten er den veldig lav og avrundet. Den egentlige verdien er regnet ut fra 3/ antall reps, i dette tilfellet er det 3/10000, som gir et veldig lavt tall.

Siden p-verdien er mindre enn signifikansnivået (a) på 1% eller 0,01 betyr det at H_0 forkastes og alternativhypotesen gjelder. Det er forskjell i gjennomsnittsvekt mellom menn og kvinner.

2f

term	estimate	std.error	statistic	p.value
(Intercept)	75.51715	0.3322816	227.26857	0
gendermale	13.69274	0.4745043	28.85693	0

(Intercept) er gjennomsnittsvekten til kvinner med en verdi på 75. *gendermale* har en gjennomsnittsvekt som er 13 større. Dette er statistisk signifikant. Det er også en veldig lav *R-squared* verdi på 0,104, noe som tyder på at denne modellen ikke er svært representativ.

2g

term	estimate	std.error	statistic	p.value
(Intercept)	-76.3000446	4.8959424	-15.584343	0.0000000
gendermale	0.8316146	0.6080989	1.367565	0.1714912
height	0.9368559	0.0301512	31.071920	0.0000000

Regresjonsmodellen viser at det å bruke høyde som forklarende variabel ikke gir veldig tydelige resultater. Her er *(Intercept)* gjennomsnittlig vekt for kvinner med en høyde på 0, som gir -76. Det samme gjelder for *gendermale*, her er gjennomsnittlig vekt for en mann med høyde på 0 lik -76.3 + 0.83 = -75.47. Dette er ikke statistisk signifikant.

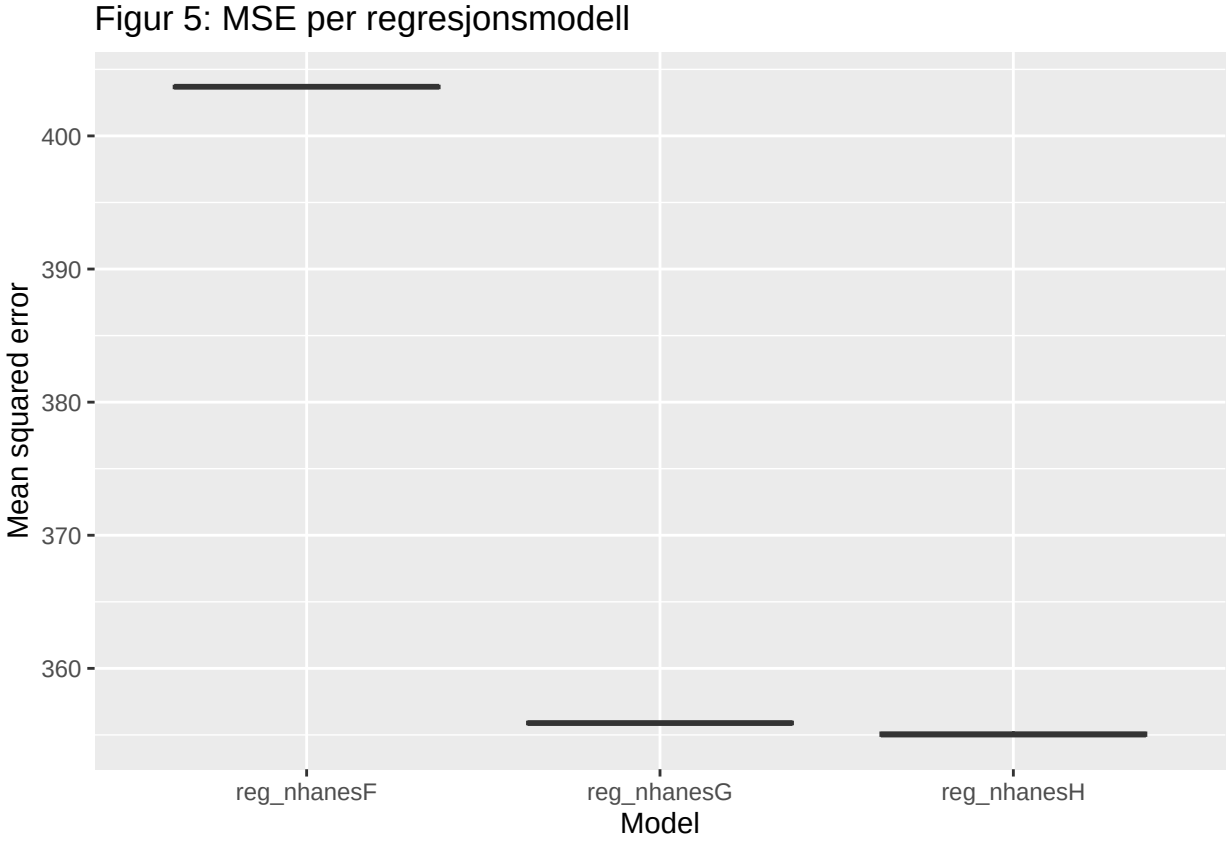
Til slutt er det også en variabel for *height* som indikerer at for hver ekstra enhet med høyde øker vekten med 0.93. Dette er statistisk signifikant og viser at vekt og høyde har et sammenheng, noe som gir mening i og med at høyde og vekt ofte har en korrelasjon, men vanligvis påvirker høyden vekta og ikke motsatt.

Det er også en større *R-squared* på 0,21, enn i $2f$ noe som tyder på at etterhvert som det blir lagt til flere forklarende variabler så blir modellen mer og mer representativ for den uavhengige variabelen.

2h

term	estimate	std.error	statistic	p.value
(Intercept)	-54.8076444	6.9290668	-7.909816	0.00e+00
gendermale	-43.6936867	10.1888246	-4.288393	1.82e-05
height	0.8042274	0.0427157	18.827463	0.00e+00
gendermale:height	0.2636628	0.0602273	4.377798	1.22e-05

2i




```
## # A tibble: 3 x 3
##   group1      group2        p.value
##   <chr>       <chr>           <dbl>
## 1 reg_nhanesG reg_nhanesF 8.14e-133
## 2 reg_nhanesH reg_nhanesF 4.47e-133
## 3 reg_nhanesH reg_nhanesG 2.63e- 34
```

Fra de tre modellene og de tre testene på *mean squared error* er det modellen i $2h$, med en MSE på 354.61 som er best til å predikere vekten til en person. Den har litt lavere MSE enn $2g$ med 355.56 og mye lavere MSE enn $2f$ med 403.44. Fra t-testen kan en også observere at det er signifikant p-verdi mellom alle tre modellene, noe som bekrefter at det er model $2h$ som er best til å predikere vekt.

2j

Det er forskjell i vekt mellom kvinner og menn. Fra de tre lineære regresjonsmodellene kan en observere at det er en signifikant forskjell mellom vekt og kjønn. Den første modellen ser kun på vekt og kjønn og det er registrert ulike verdier for kvinner og menn. Når høyde blir lagt til er det ikke signifikant p-verdi for *gendermale*, men det er det i $2h$ modellen. Siden denne har lavest MSE og kjønn har signifikant p-verdi kan en konkludere at det er forskjell i vekt mellom kvinner og menn.

2k

Predikert vekt fra modellen i $2g$ er 82.97 kg for kvinner og 83.8 kg for menn. Predikert vekt fra modellen i $2h$ er 81.91 kg for kvinner og 83.04 kg for menn.

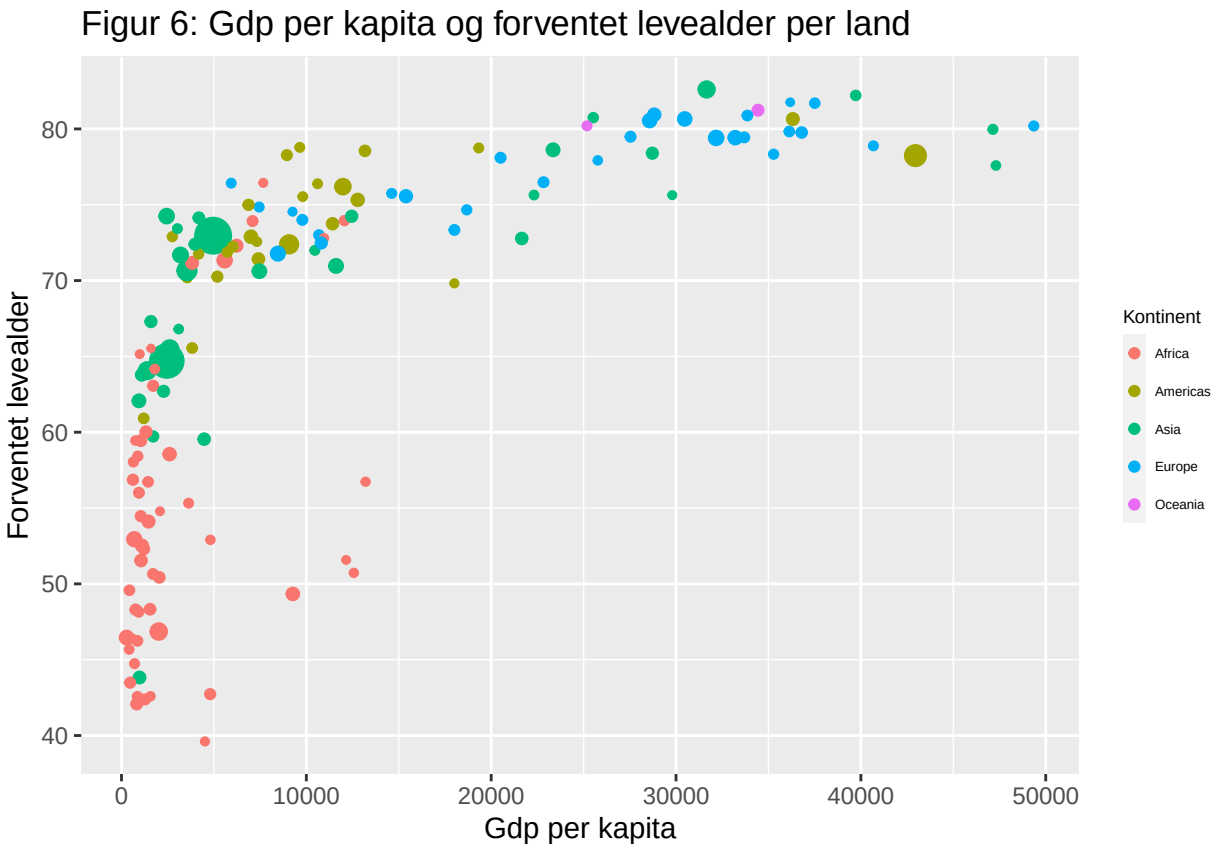
Oppgave 3

3a

continent	life_exp	pop	gdp_percap
Africa	54.80604	17875763	3089.033
Americas	73.60812	35954847	11003.032
Asia	70.72848	115513752	12473.027
Europe	77.64860	19536618	25054.482
Oceania	80.71950	12274974	29810.188

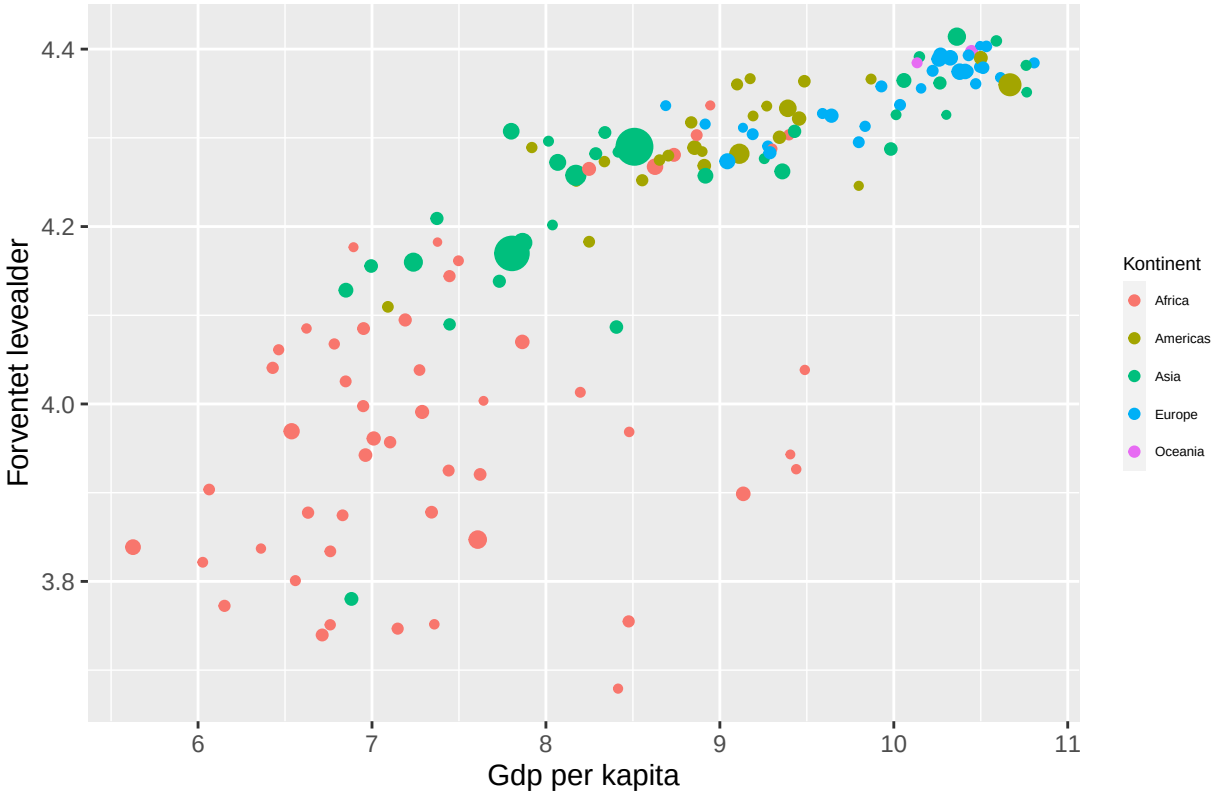
Fra tabellen kan en se at *Afrika* er kontinentet med gjennomsnittlig lavest forventet levealder og gdp per kapita. *Oceania* har både høyest forventet levealder og høyest gdp per kapita og lavest befolkning. Asia har størst gjennomsnittlig befolkning.

3b



3c

Figur 7: Gdp per kapita og levealder per land



3d

term	estimate	std.error	statistic	p.value
(Intercept)	3.4560849	0.0713370	48.4473147	0.0000000
log(gdp_percap)	0.0711688	0.0093186	7.6372550	0.0000000
continentAmericas	0.1980831	0.0292409	6.7741878	0.0000000
continentAsia	0.1714562	0.0269095	6.3715830	0.0000000
continentEurope	0.1843052	0.0334214	5.5145917	0.0000002
continentOceania	0.2022714	0.0797067	2.5376965	0.0122948
pop	0.0000000	0.0000000	0.3234099	0.7468853

I den lineære regresjonsmodellen er det verdien for log av forventet levealder som er verdien for *(Intercept)*, det er denne de isolerte endringene måles mot. *Estimate*-kolonnen viser verdien av de isolerte endringene, målt opp mot *(Intercept)*. *Pr(>|t|)*-kolonnen er p-verdiene for de isolerte endringene.

Her er det to hypoteser: - H_0: det er ingen sammenheng mellom forventet levealder og de valgte variablene. - H_1: det er sammenheng mellom forventet levealder og de valgte variablene. Det er også et signifikansnivå *(a)* på 1% eller 0,01.

log(gdp_percap) viser at gdp per kapita fører til en økning i forventet levealder, og at det har en p-verdi lavere enn signifikansnivået på 1%. Det betyr at H_0 hypotesen forkastes, og at det er et statistisk sammenheng mellom forventet levealder og økt gdp per kapita.

continentAmericas, *continentAsia* og *continentEurope* har alle signifikante p-verdier. Det betyr at forventet levealder økes av de variablene, i ulike størrelse. Her forkastes også H_0 hypotesen, og det er et statistisk sammenheng mellom forventet levealder og kontinent.

continentOceania er har en p-verdi >1% eller >0.01. Det betyr at H_0 beholdes, og at det ikke er statistisk sammenheng mellom variablene. Grunnen til at Oceania ikke er statistisk signifikant kan komme av at det kun er to observasjoner fra dette kontinentet, Australia og New Zealand. Antallet observasjoner er for få, det er ikke nok observasjoner til å kunne med sikkerhet si at Oceania øker forventet levealder, eller om det er andre faktorer som spiller inn.

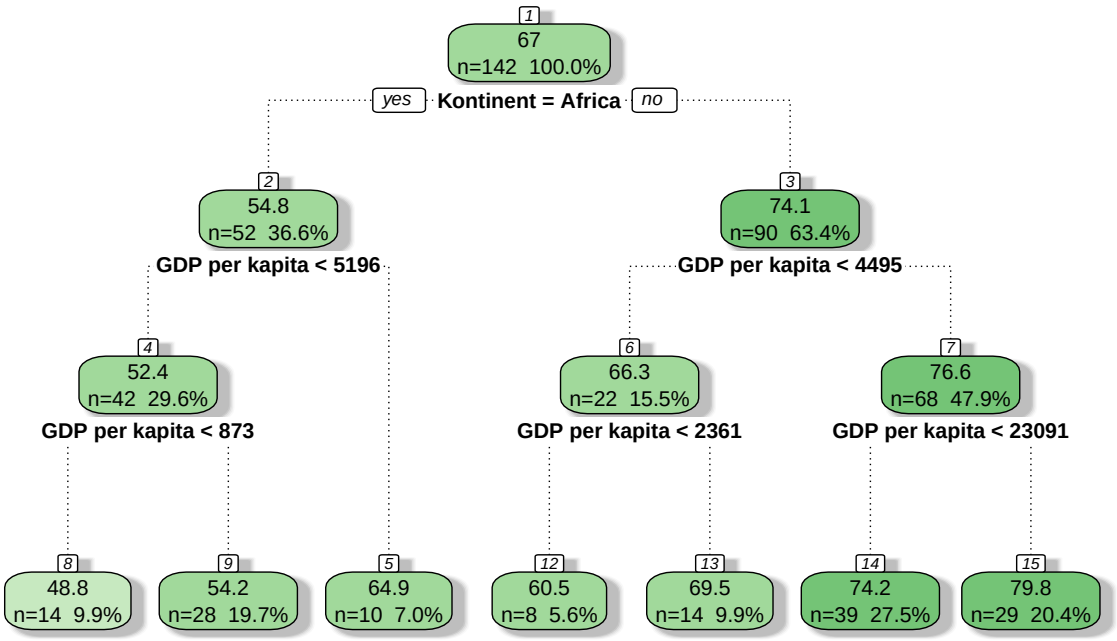
Den siste variabelen, *pop*, har ingen statistisk sammenheng med levealder, H_0 beholdes. Dette kan komme av at den totale mengden mennesker i et kontinent ikke nødvendigvis påvirker hvor lenge individer gjennomsnittlig lever.

3e

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
136	1.484325	NA	NA	NA	NA
135	1.474613	1	0.0097121	0.8891371	0.3473957

En linearHypothesis test mellom *Americas* og *Asia* gir en p-verdi >1%. Dette betyr at H_0 beholdes, det er ingen statistisk sammenheng mellom forventet levealder i *Americas* og *Asia*. Ikke veldig overaskende at det ikke er noe statistisk sammenheng mellom de to kontinentene, ettersom gjennomsnittlig levealder i et kontintent ikke blir påvirket av gjennomsnittlig levelader i et annet.

3f



Figuren viser hvordan forventet levealder påvirkes av de uavhengige variablene GDP per kapita, kontinent og populasjon.

Treet tar utgangspunkt i en forventet levealder på 67.01, dette er gjennomsnittet av alle landene. Øverst deler treet seg i to, en retning for Afrika og den andre er de resterende kontinentene. Om en følger kontinentet Afrika deles det først etter GDP per kapita mindre eller større enn 5196. Om landet har en GDP per kapita større enn 5196 vil forventet levealder være 65 år. Her er det 10 observasjoner eller 7% av totalen.

For landene med mindre enn 5196 i GDP per kapita deles det en gang til, denne gangen om GDP per kapita er mer eller mindre enn 873. Om GDP per kapita er mer enn 873 er forventet levealder 54 år, det er 28 observasjoner eller omtrent 20% av totalen. Den siste alternativet er land med GDP per kapita mindre enn 873, disse landene har en forventet levealder på 49 og det er 14 observasjoner eller 10% av totalen.

For de resterende fire kontinentene så er det tilbake til starten. Forventet levealder blir delt opp etter land med større eller mindre GDP per kapita enn 4495. Om den er mindre blir den delt opp etter om GDP per kapita er mindre enn 2361 er forventet levealder 61, her er det 8 observasjoner eller 6% av totalen. Er GDP per kapita mer enn 2361 men mindre enn 4495 er forventet levealder 70 år, med 14 observasjoner, 10% av totalen.

For de landene med GDP per kapita større enn 4495 deles de igjen i to, de landene som har GDP per kapita mellom 4495 og 23091 er forventet levealder 74, 39 observasjoner eller 27.5% av totalen. De siste observasjonene er land med GDP per kapita over 23091, disse har en forventet levealder på 80, det er 29 observasjoner og 20% av totalen.

Oppgave 4

4a

For å definere hva standardavvik er, må en vite hva stokastisk variabel er. Dette er en variabel av et utfall med tilfeldig hendelse. Flere variabler av utfall med tilfeldig hendelse i lag gjør opp et datasett. Et slikt datasett er nødvendig for definere standardavvik.

I et datasett vil standardavviket være et mål for gjennomsnittlig avstand fra gjennomsnittet. Det vil si at i et datasett med 100 observasjoner og et gjennomsnitt på 50 vil standardavviket være den et mål for den gjennomsnittlige avstanden fra det målte gjennomsnittet på 50.

Standardavvik er et viktig mål for å kunne bekrefte om et datasett er normalfordelt eller ikke. Ofte måler man standardavvik i 3 eller 4 nivåer. I et normalfordelt datasett vil ett standardavvik fra gjennomsnittet inneholde $\approx$ 68,3% av alle observasjonene. $\approx$ 95,5% av observasjonene vil være to standardavvik fra gjennomsnittet. Det betyr at i et normalfordelt datasett vil kun 5% av observasjonene ikke inntreffe innenfor to standardavvik. Tre standardavvik inneholder 99,7%.

4b

Standardfeil er et mål på feilmarginen i et estimat med data. Standardfeil gir et estimat på hvor nøyaktig dataene representerer populasjonen som den er hentet fra. Størrelsen på standardfeilen er avgjørende for hvor nøyaktig dataene er. En større standardfeil tyder på at innhentet data ikke gir en god forklaring på hele befolkningen, mens en liten standardfeil viser at dataene representer hele befolkningen bedre.

4c

En hypotesetest er en statistisk prosedyre for å teste teorier om data er korrekt eller ikke. En hypotese baseres på innsamlet data der en forsøker å forklare om dataene kan ha et sammenheng eller ikke.

Fremgangsmåten innebærer det å sette en nullhypotese, H_0, og en alternativhypotese, H_1. Nullhypotesen er en teori om at innsamlet data ikke har en direkte statistisk sammenheng, mens alternativhypotesen er at det eksisterer ett sammenheng. Et signifikansnivå settes, ofte er dette på 95%. Signifikansnivået er avgjørende for om nullhypotesen eller alternativhypotesen er gjeldende.

P-verdien måles mot nullhypotesen. P-verdien avgjør om nullhypotesen er forkastes eller ikke. En lav p-verdi indikerer at dataene er lav sannsynlighet for at utfallet er tilfeldig, så nullhypotesen kan forkastes, og alternativshypotesen er gjeldende. P-verdien er tett knyttet sammen med signifikansnivået. Om signifikansnivået er på 95% vil p-verdien være 5% eller 0,05.

4d

Konfidensintervall er et mål for hvor godt et estimat er. Konfidensintervall er ytterpunktene for et intervall, og størrelsen avgjør hvor sikker et estimat er. I konfidensintervaller er det både en øvre og en nedre grense.

Konfidensintervallet kan regnes ut fra gjennomsnittet av et estimat $\pm$ variasjonen i det estimatet. Ofte benyttes et konfidensintervall på 95%, som vil si at i 19 av 20 tilfeller vil estimatet havne mellom den øvre og nedre grensen på konfidensintervallet.