SOK-2009-EksamenH2022

Kandidatnummer 1

Eksamen Sok-2009 Høst 2022

Oppgave 1

I datasettet NHANES skal vi først se på utdanningsnivå og etsinitet om det er noen forskjeller. Vi har et signifikant nivå på 1%

Oppgave 1a

"Hvilken type målenivå er variablene race1 og education? og hva er spesielt med disse to variablene?"

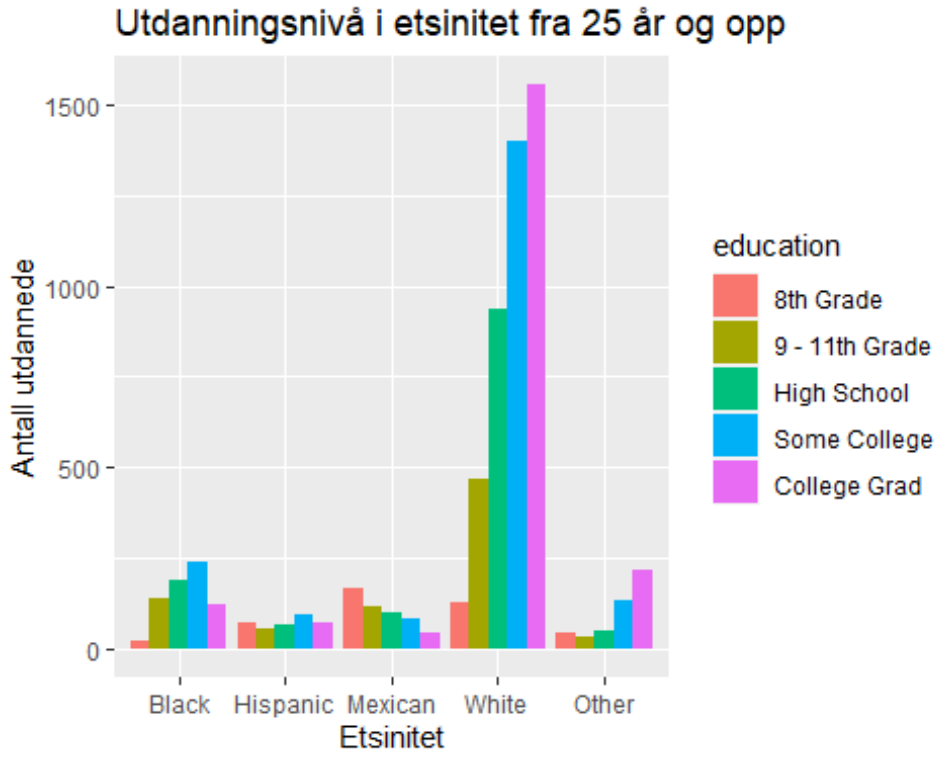
```
 [1] "id"                 "survey_yr"          "gender"
 [4] "age"                "age_decade"         "age_months"
 [7] "race1"              "race3"              "education"
[10] "marital_status"     "hh_income"          "hh_income_mid"
[13] "poverty"            "home_rooms"         "home_own"
[16] "work"               "weight"             "length"
[19] "head_circ"          "height"             "bmi"
[22] "bmi_cat_under20yrs" "bmi_who"            "pulse"
[25] "bp_sys_ave"         "bp_dia_ave"         "bp_sys1"
[28] "bp_dia1"            "bp_sys2"            "bp_dia2"
[31] "bp_sys3"            "bp_dia3"            "testosterone"
[34] "direct_chol"        "tot_chol"           "urine_vol1"
[37] "urine_flow1"        "urine_vol2"         "urine_flow2"
[40] "diabetes"           "diabetes_age"       "health_gen"
[43] "days_phys_hlth_bad" "days_ment_hlth_bad" "little_interest"
[46] "depressed"          "n_pregnancies"      "n_babies"
[49] "age1st_baby"        "sleep_hrs_night"    "sleep_trouble"
[52] "phys_active"        "phys_active_days"   "tv_hrs_day"
[55] "comp_hrs_day"       "tv_hrs_day_child"   "comp_hrs_day_child"
[58] "alcohol12plus_yr"   "alcohol_day"        "alcohol_year"
[61] "smoke_now"          "smoke100"           "smoke100n"
[64] "smoke_age"          "marijuana"          "age_first_marij"
[67] "regular_marij"      "age_reg_marij"      "hard_drugs"
[70] "sex_ever"           "sex_age"            "sex_num_partn_life"
[73] "sex_num_part_year"  "same_sex"           "sex_orientation"
[76] "pregnant_now"

Rows: 10,000
Columns: 4
$ age        <int> 34, 34, 34, 4, 49, 9, 8, 45, 45, 45, 66, 58, 54, 10, 58,
50…
$ age_decade <fct>  30-39,  30-39,  30-39,  0-9,  40-49,  0-9,  0-9,  40-49,
```

```
…
$ race1      <fct> White, White, White, Other, White, White, White, White,
Whi…
$ education  <fct> High School, High School, High School, NA, Some College,
NA…
```

Etsiniteten(race1) er kategorisk og regnes som nominelt målenivå fordi det ikke er fornuftig i å kunne rangere disse. Denne type variabel er generelt ikke mulig å rangere da ingen er større eller mindre en den andre kategorien, unntaket er diktome variabler og er kun 2 kategorier som for eksempel kan settes til 0 og 1, ja eller nei eller de kan være true eller false. Utdanningsnivå(education) er ordinal nivå fordi den kan rangeres.

Oppgave 1b

"Lag figur for å vise sammenhengen mellom utdanningsnivå og etsinitet. Kommenter dine funn. ser det ut som om det er forskjell mellom utdanningsnivå og etsinitet"



Det jeg finner er at hvite mennesker har mye mer utdanning enn de andre etsinitetene, de ligger generelt ganske lavt, det ser ut som at det er en forskjell mellom utdanningsnivå og etsinitet, især den ene kategorien mot de andre, men så kan man spørre seg om hvor stor andel av disse tallene er hvite og at dette kanskje kan ha en påvirkning av resulatet i plottet. Men la oss se på tabellen:

Oppgave 1c

En tabell for å vise tallene.

```
            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
```

Og i tabellen ser man at andelen hvite tilsammen er mye større en de andre og på et plot vil dette bare vise antall og ikke generelt om der er en sammenheng, dette må vi finne ut av ved å kjøre en statististisk test.

Oppgave 1d

For å finne ut om det har en sammenheng kan vi kjøre en statistisk test for å analyse dette. For våre kategorsike variabler bruker jeg en chisq test, fordi denne er best å bruke på disse variablene.

Nullhypotesen(H0) er at det ikke er en sammenheng mellom etsinitet og utdanningsnivå.

Alternativ hypotesen(H1) er at det er en sammenheng mellom etsinitet og utdanningsnivå.

Oppgave 1e

Gjennomfører testen.

```
    Pearson's Chi-squared test

data:  nhanesfra25$race1 and nhanesfra25$education
X-squared = 1094, df = 16, p-value <2e-16
```

Siden P-verdien er lavere enn signifikantnivået på 1 % ,vi kan da forkaste nullhypotesen og gå for den alternative som er at det er en sammenheng mellom etsinitet og utdanningsnivå.

Oppgave 2

Vi fortsetter å bruke NHANES datasettet men nå skal vi se på Vekt, høyde og kjønn og om vi finner ut om det er en forskjell i vekt mellom menn og kvinner.

Oppgave 2a

“Hvilke type variabler er height, weight og gender? og hva er spesielt med dem?”

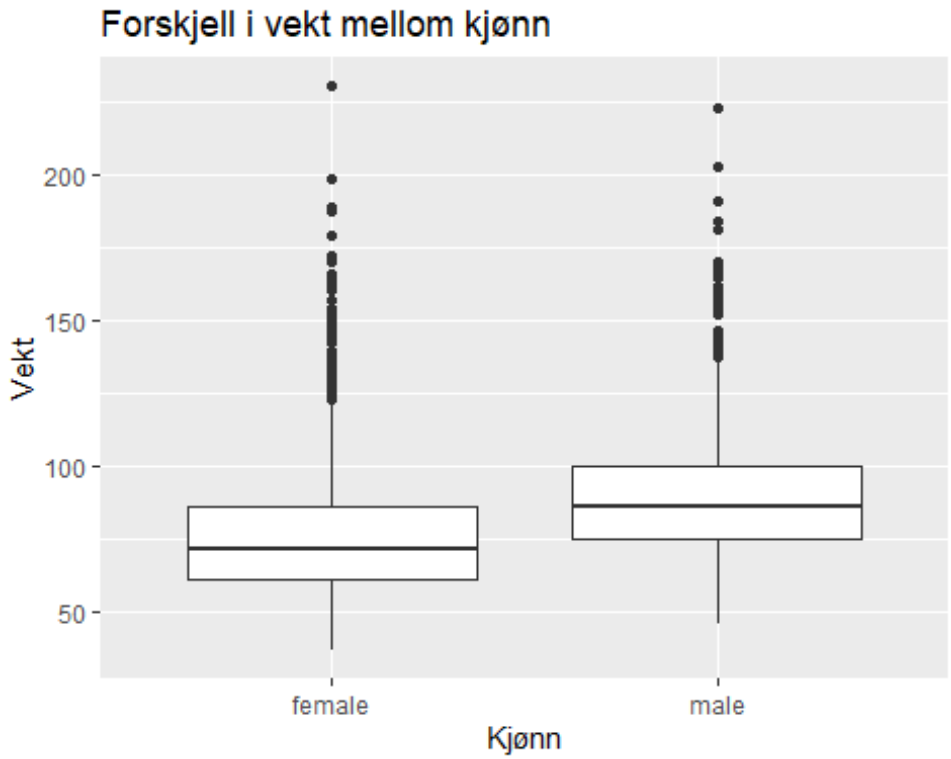
```
Rows: 7,172
Columns: 3
$ weight <dbl> 87.4, 87.4, 87.4, 86.7, 75.7, 75.7, 75.7, 68.0, 78.4, 74.7,
57.…
$ height <dbl> 164.7, 164.7, 164.7, 168.4, 166.7, 166.7, 166.7, 169.5, 181.9,
…
$ gender <fct> male, male, male, female, female, female, female, male, male,
m…
```

Kjønn er en faktor, en diktome variabel i nominalnivå og kan være 0 og 1, ja eller nei eller true og false.

Vekt og høyde er tall med desimaler, disse er forholdstall som kan regnes om til relative tall og man kan finne for eksempel gjennomsnittet eller medianen..

Oppgave 2b

Lager en figur for å vise forskjell i vekt mellom menn og kvinner.

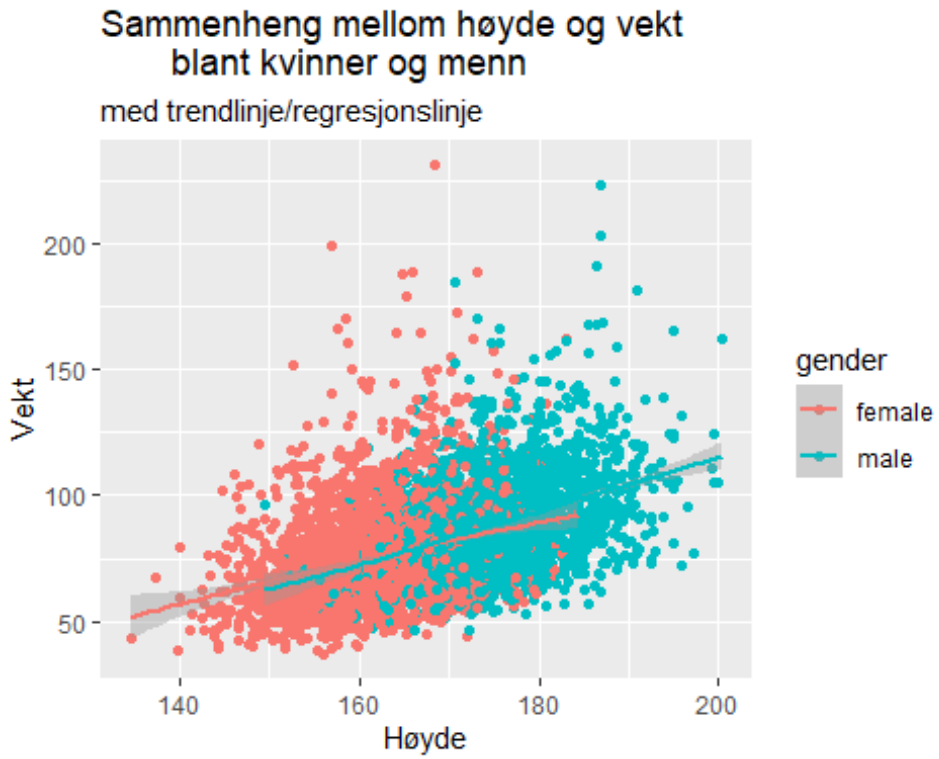


Kvinner veier mindre enn menn og har også et lavere gjennomsnitt enn menn, det ser vi på boksene at mennene sin er høyere med linjen også høyere enn kvinnenes. Menn er også mer samlet rundt gjennomsnittet og færre utenfor boksen som er større avstand mellom gjennomsnittet og den faktiske vekten enn kvinner.

Oppgave 2c

Her skal vi se om det er noe sammenheng mellom vekt og høyde hos menn og kvinner.

```
`geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Ikke den beste figuren, men vi klarer å se at de røde/kvinner er lavere generelt i høyde men en sammenheng med jo høyere jo mer veier dem hvis man ser den røde linjen som viser trenden blandt kvinnene. Samme med de blå/mennene bare at de generelt er høyere enn kvinnene samt at de også er tyngre jo høyere de er. Den blå linjen viser en stigende trend også blandt mennene.

Oppgave 2d

En tabell med gjennomsnittet til høyde, vekt, standardavviket, standardfeil og antall i hver gruppe.

```
 Descriptive statistics by group
gender: female
        vars    n   mean    sd median trimmed   mad   min   max range  skew
weight     1 3655  75.52 20.44   71.5   73.44 17.20  37.0 230.7 193.7  1.39
height     2 3655 162.05  7.29  162.1  162.06  7.26 134.5 184.5  50.0 -0.04
gender*    3 3655   1.00  0.00    1.0    1.00  0.00   1.0   1.0   0.0   NaN
        kurtosis   se
weight      4.03 0.34
height     -0.02 0.12
gender*      NaN 0.00
------------------------------------------------------------
gender: male
        vars    n   mean    sd median trimmed   mad   min   max range skew
```

```
weight     1 3517  89.21 19.72   86.2   87.66 18.24  46.2 223.0 176.8 1.09
height     2 3517 175.78  7.48  175.6  175.75  7.12 149.4 200.4  51.0 0.03
gender*    3 3517   2.00  0.00    2.0    2.00  0.00   2.0   2.0   0.0  NaN
        kurtosis   se
weight      2.75 0.33
height      0.14 0.13
gender*      NaN 0.00
```

Oppgave 2e

Gjør en permutasjonstest for å se om det er forskjell i gjennomsnittsvekten mellom kjønnene, bruker likemange gjennomsnitt som observasjoner i datasettet.

Oppgave 2e i

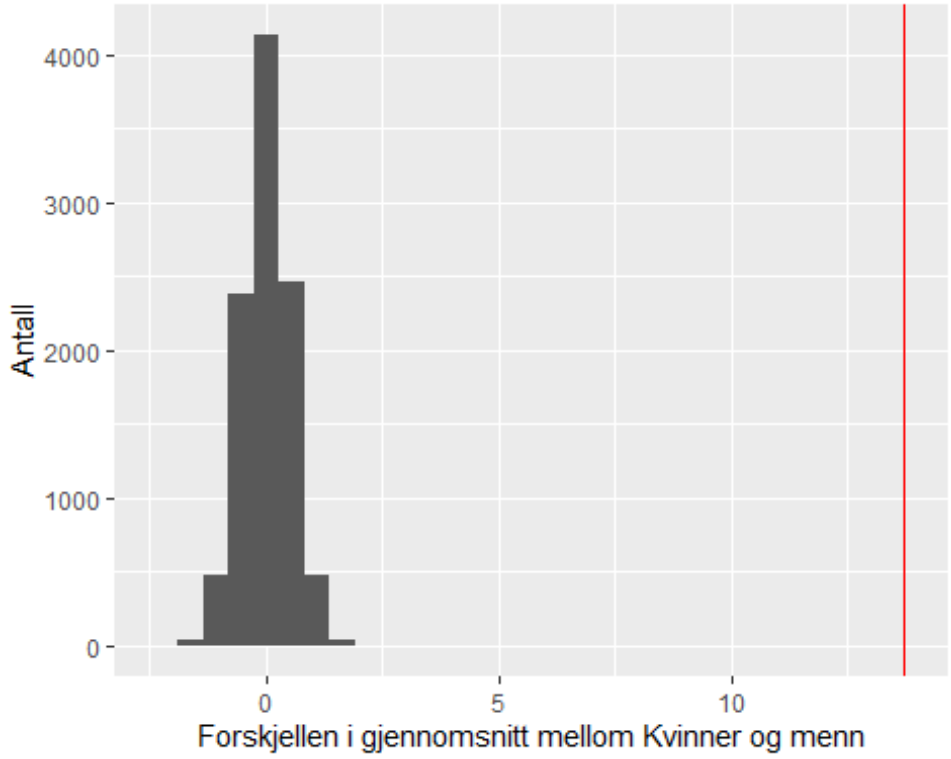
Nullhypotesen(H0) : Det er ingen forskjell mellom kjønnene

Alternativhypotesen(H1): det er en forskjell mellom kjønnene

Oppgave 2e ii

Plotter testen i et histogram.

"Hvilken type tekst er dette i klassisk statistikk?"



Dette er en permutasjonstest. Permutasjon er en bestemt ordning av elementene i en mengde i dette tilfellet dataene vi har.(matematikk.net,20.09.11,Permutasjon – Matematikk.net)

Den tester alle mulige kombinasjoner.

Figuren viser resultatet med det faktiske forskjellen som den røde streken og vi ser på resultatet at resultatet avviker fra den faktiske forskjellen.

Oppgave 2e iii

Finne P-verdi.

```
# A tibble: 1 × 1
  p_value
    <dbl>
1       0

[1] 0
```

P-verdien er lik ren 0 og er da mindre en 0.01 og er derfor lavere enn signifikantnivået som betyr at vi kan forkaste nullhypotesen.

Oppgave 2f

Nå skal vi gjøre noen regresjonsmodeller og da har vi satt et signifikantnivå på 5%.

Regresjonsmodell 1

```

Call:
lm(formula = weight ~ gender, data = nhanes20p)

Residuals:
   Min     1Q Median     3Q    Max
-43.01 -14.02  -3.61  10.88 155.18

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   75.517      0.332   227.3   <2e-16 ***
gendermale    13.693      0.475    28.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.1 on 7170 degrees of freedom
Multiple R-squared:  0.104, Adjusted R-squared:  0.104
F-statistic:  833 on 1 and 7170 DF,  p-value: <2e-16
```

Oppgave 2g

Regresjonsmodell 2

```

Call:
lm(formula = weight ~ gender + height, data = nhanes20p)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-41.52 -13.01  -2.93   9.63 149.14

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -76.3000     4.8959  -15.58   <2e-16 ***
gendermale    0.8316     0.6081    1.37     0.17
height        0.9369     0.0302   31.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.9 on 7169 degrees of freedom
Multiple R-squared:  0.21,  Adjusted R-squared:  0.21
F-statistic:  955 on 2 and 7169 DF,  p-value: <2e-16
```

Oppgave 2h

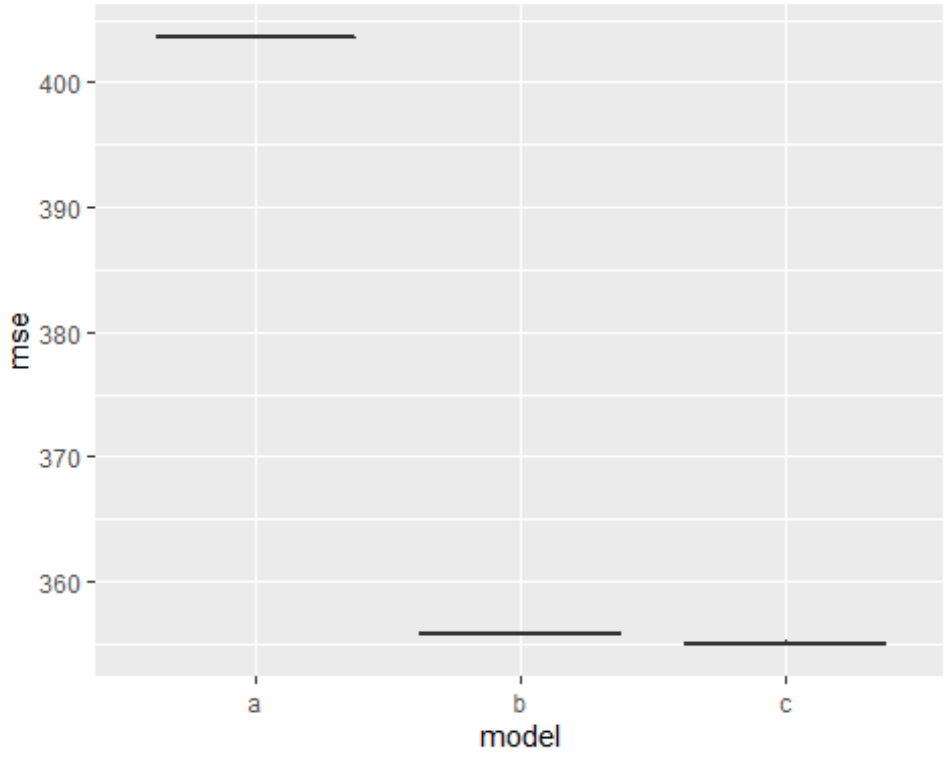
regresjonsmodell 3

```
Call:
lm(formula = weight ~ gender + height + gender * height, data = nhanes20p)

Residuals:
   Min     1Q Median     3Q    Max
-42.79 -13.05  -3.03   9.57 150.00

Coefficients:
                  Estimate Std. Error t value          Pr(>|t|)
(Intercept)       -54.8076     6.9291   -7.91 0.000000000000003 ***
gendermale        -43.6937    10.1888   -4.29 0.000018233107397 ***
height              0.8042     0.0427   18.83           < 2e-16 ***
gendermale:height   0.2637     0.0602    4.38 0.000012158668337 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.8 on 7168 degrees of freedom
Multiple R-squared:  0.212, Adjusted R-squared:  0.212
F-statistic:  645 on 3 and 7168 DF,  p-value: <2e-16
```

Oppgave 2i

“Hvilken modell er best til å predikere vekten til en person?”

```
[1] 403.44

[1] 355.56

[1] 354.61
```




```
     a      b      c
403.65 355.85 355.03
```

Modellen som er best for å predikere vekten til en person er regresjonsmodell 3. dette fordi den har den laveste verdien. Muligehetn for feil er lavere enn de andre. det ser du også på boxplottet at nr c/regresjonsmodell er den med lavest snitt.

Oppgave 2j

Ja det er forskjell i vekt mellom kvinner og menn, kvinner er generelt mindre og da også lettere enn menn, kvinner har også fra naturens siden mer fett på kroppen enn menn men samtidig mindre muskler, og muskler veier mer enn fett også, så dette spiller inn på om det er forskjell mellom menn og kvinner. selv om forskjellen er liten så er den likevel der biologisk sett tenker jeg. Ved å predikere vekten utfra de modellene vi har kan vi finne dette ut.

Oppgave 2k

å predikere vekten til en kvinne og en mann som er like høye, i dette tilfeller 170cm høy.

```
Call:
lm(formula = weight ~ height + gender, data = nhanes20p)

Residuals:
   Min     1Q Median     3Q    Max
-41.52 -13.01  -2.93   9.63 149.14
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -76.3000     4.8959  -15.58   <2e-16 ***
height        0.9369     0.0302   31.07   <2e-16 ***
gendermale    0.8316     0.6081    1.37     0.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.9 on 7169 degrees of freedom
Multiple R-squared:  0.21,  Adjusted R-squared:  0.21
F-statistic:  955 on 2 and 7169 DF,  p-value: <2e-16

     1
82.965

     1
83.797
```

En kvinne på 170 vil veie 82,9 kg mens en mann på 170 vil veie 83,7 kg. Altså veier faktisk mannen mer en kvinnen.

Oppgave 3

I denne oppgaven bruker vi datasettet gapminder.

```
[1] "country"    "continent"  "year"       "life_exp"   "pop"
[6] "gdp_percap"

Rows: 1,704
Columns: 6
$ country    <fct> "Afghanistan", "Afghanistan", "Afghanistan",
"Afghanistan",…
$ continent  <fct> Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia, Asia,
Asia,…
$ year       <int> 1952, 1957, 1962, 1967, 1972, 1977, 1982, 1987, 1992,
1997,…
$ life_exp   <dbl> 28.801, 30.332, 31.997, 34.020, 36.088, 38.438, 39.854,
40.…
$ pop        <int> 8425333, 9240934, 10267083, 11537966, 13079460, 14880372,
1…
$ gdp_percap <dbl> 779.45, 820.85, 853.10, 836.20, 739.98, 786.11, 978.01,
852…

[1] "country"    "continent"  "year"       "life_exp"   "pop"
[6] "gdp_percap"
```

Oppgave 3a

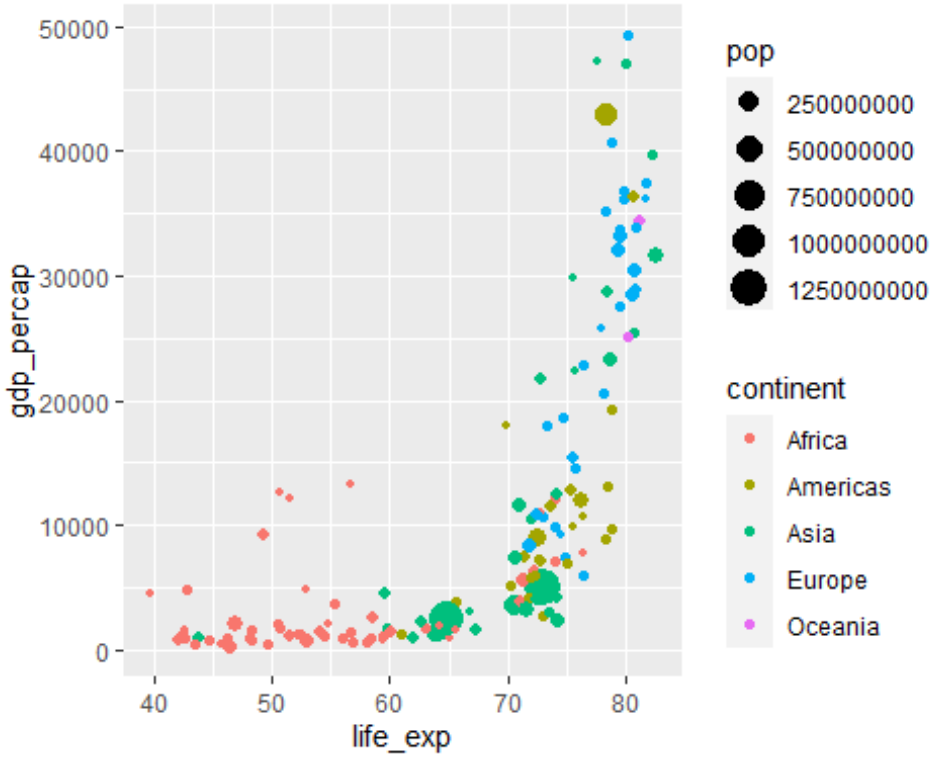
Lager en tabell og skal kommentere denne

continent	bnp_prcap	antall_mennesker	forventet_liv
Africa	3089	17875763	54.806
Americas	11003	35954847	73.608
Asia	12473	115513752	70.728
Europe	25054	19536618	77.649
Oceania	29810	12274974	80.719

I tabellen viser det for hvert kontinent gjennomsnittelig bnp pr capita, gjennomsnittlig antall mennesker, og gjennomsnittelig forventet levealder.

Oppgave 3b

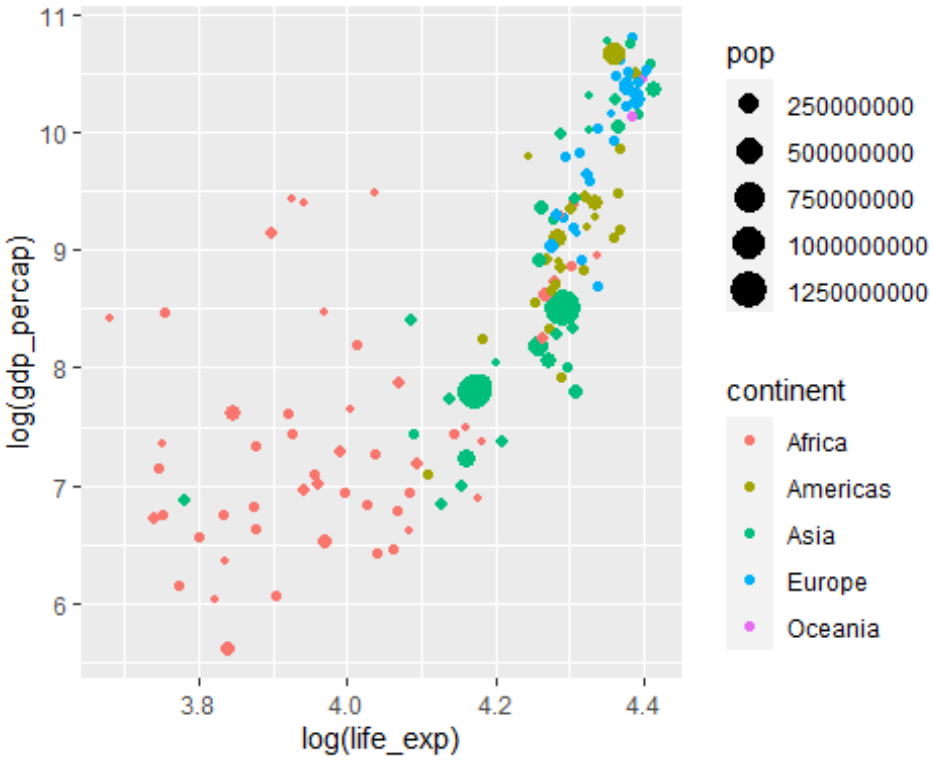
Et plot som vi kan bruke for å forøke å forklare en forventet levealder basert på brutto nasjonalprodukt pr capita.



Her ser vi resultatet men det er ikke lett å se om det er en eventuell sammenheng.

Oppgave 3c

Siden plottet i oppgave 3b ikke er linær så skal vi prøve på å bruke logmaritmen og kjøre denne i et nytt plott for å se hva vi får og om det er lettere å se en eventuell sammenheng.



Her ser man at det er mere linært og lettere å tolke en eventuell sammenheng.

Oppgave 3d

Estimer en regrasjon i log, hvordan tolker vi estimatene?, er funnene av hva vi kan forvente?, hvorfor er ikke Oceania statistisk signifikant når dette kontinentet har den høyeste forventet levealder?

```
Call:
lm(formula = life_exp ~ log(gdp_percap) + continent + pop, data = gapminder2)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.485  -2.325  -0.011   2.413  14.884 

Coefficients:
                         Estimate      Std. Error t value Pr(>|t|)    
(Intercept)       20.049877770721  4.060944298035    4.94  2.3e-06 ***
log(gdp_percap)    4.640280064903  0.530474731843    8.75  7.7e-15 ***
continentAmericas 11.662903964054  1.664571814607    7.01  1.1e-10 ***
continentAsia     10.012260544378  1.531856414227    6.54  1.2e-09 ***
continentEurope   11.242929993820  1.902552368881    5.91  2.7e-08 ***
continentOceania  12.907837641248  4.537400237294    2.84   0.0051 ** 
pop                0.000000000924  0.000000003534    0.26   0.7941    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.95 on 135 degrees of freedom
Multiple R-squared:  0.767, Adjusted R-squared:  0.757
F-statistic: 74.3 on 6 and 135 DF,  p-value: <2e-16
```

I estimatene finner vi at oceania avviker mest fra de andre. I og med at vi finner at oseania har høyeste og laveste verdi som avviker veldig fra de andre så er ikke denne statistisk signifikant,

Oppgave 3e

Gjennomfører en statistisk test for å se om det er forskjell i forventet levealder mellom Amerika og Asia.

```
# A tibble: 25 × 1
   life_exp
      <dbl>
 1     75.3
 2     65.6
 3     72.4
 4     80.7
 5     78.6
 6     72.9
 7     78.8
 8     78.3
 9     72.2
10     75.0
# … with 15 more rows

# A tibble: 33 × 1
   life_exp
      <dbl>
 1     43.8
 2     75.6
 3     64.1
 4     59.7
 5     73.0
 6     82.2
 7     64.7
 8     70.6
 9     71.0
10     59.5
# … with 23 more rows


    Welch Two Sample t-test

data:  Americas and Asia
t = 1.75, df = 52, p-value = 0.086
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
```

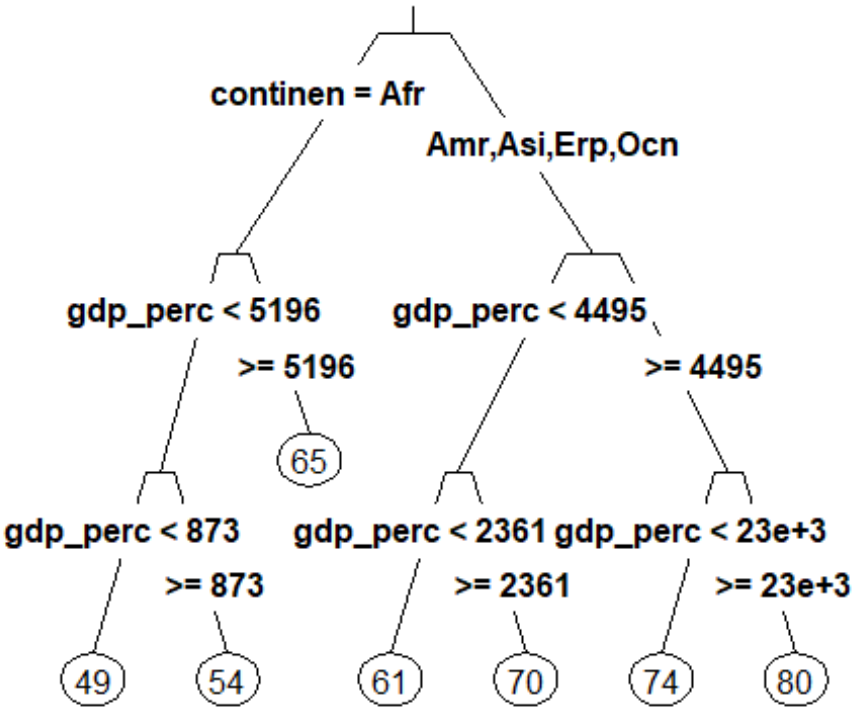
```
 -0.42419  6.18346
sample estimates:
mean of x mean of y
   73.608    70.728
```

Nullhupotese(H0) = at forskjell ikke er tilstede. Og alternativ hypotesen er at det er forskjell så indikerer p verdien som er over signifikantnivået på 5% at det er en forskjell på forventet levealder mellom Amerika og Asia.

Oppgave 3f

Gjør en rpart analyse og lager figur samt skal tolke denne.

```
n= 142

node), split, n, deviance, yval
      * denotes terminal node

 1) root 142 20552.000 67.007
   2) continent=Africa 52  4730.300 54.806
     4) gdp_percap< 5196.1 42  2323.100 52.399
       8) gdp_percap< 873.03 14   438.120 48.774 *
       9) gdp_percap>=873.03 28  1608.900 54.212 *
     5) gdp_percap>=5196.1 10  1142.300 64.914 *
   3) continent=Americas,Asia,Europe,Oceania 90  3607.200 74.057
     6) gdp_percap< 4495.3 22  1019.100 66.266
      12) gdp_percap< 2361.2 8   355.690 60.547 *
      13) gdp_percap>=2361.2 14   252.230 69.534 *
     7) gdp_percap>=4495.3 68   820.790 76.578
      14) gdp_percap< 23091 39   227.180 74.172 *
      15) gdp_percap>=23091 29    64.173 79.814 *
```



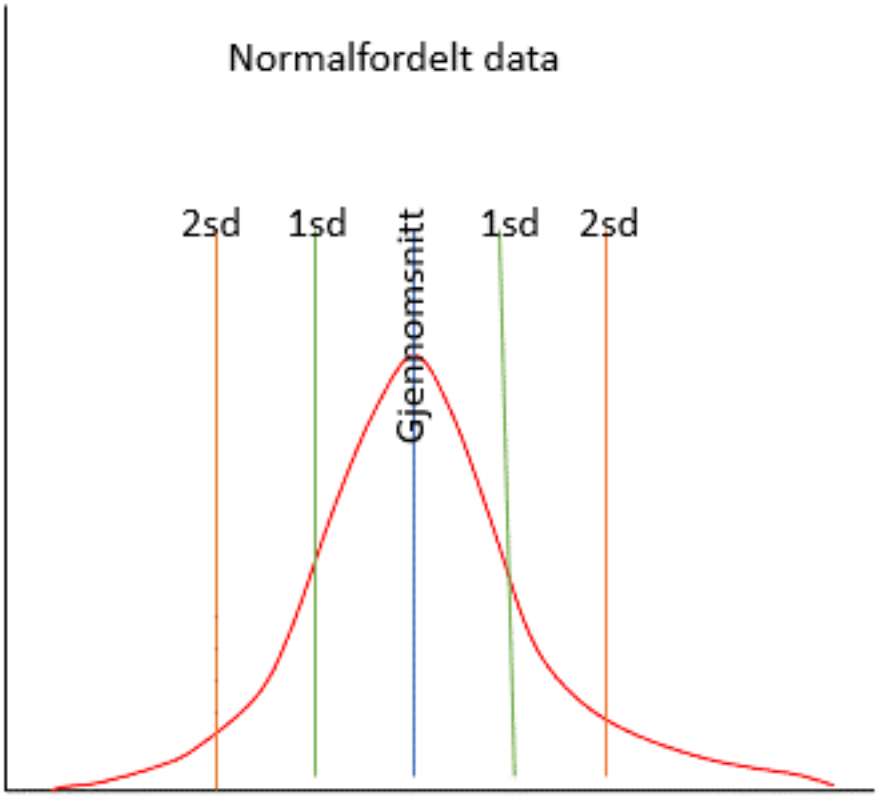
Rpart setter datasettet i ja eller nei kategorier og vi kan få de opp i et tre, starter med at afrika skiller seg ut til venstre og måler videre på størrelsen på bnp capita om hvor mange av landene som har bnp lavere enn 5196, hvis ja så går den til venstre, er den høyere så går den til høyre og teller andel som har bnp samme eller høyere enn 5196. og slik fortsetter den nedover, samme prisippet på andre siden hvor du finner de andre kontinentene. Bnp pr capita i kontinentene blir på en måte rangert etter hvor stor andel under har hva.

Oppgave 4

Oppgave 4a

"Hva er standardavvik?"

Standardavvik er et mål som sier noe om spredning i dataene du har, kalles også for spredningsmål, det er kvadratroten av variansen.(Helbæk,M,Statistikk Kort og godt, 3utg,4opplag 2019). 1 standardavvik kan være et avvik mot høyre eller venstre, ved 2 så hopper den bort et hakk til. Se tegnet bilde.



Standardavvik i normalfordelt data, tegnet eksempel

Oppgave 4b

"Hva er standard error eller standard feil?"

En Standard feil er forskjellen i gjennomsnittet i et utvalg kontra gjennomsnittet for hele poulasjonen. Den gir deg en indikasjon på hvor godt utvalg du har av populasjonen. (My Race,blog and life, Standardfeil | hva det Er, Hvorfor Det Betyr noe, Og Hvordan Man Beregner | My Race (racem.org))

Oppgave 4c

"Hva er en hypotesetest og dens tilhørende p-verdi?"

En hypotesetest er en statistisk test for å teste en hypotese eller en påstand, p-verdien forteller om hvor sikkert resultatet er. Du setter et signifikantnivå om hvor mye feil av type 1 du tillater for å beholde eller forkaste nullhypotesen/påstanden.

Altså p verdien forteller om testen/modellen er eller ikke er statistisk signifikant.

Oppgave 4d

"Hva er konfidensintervall"

Et mål på kvaliteten av estimatene i en test eller ukjente størrelser. Jo kortere konfidensintevall jo sikrere er resultatene. Den gir en øvre og nedre grense for størrelsen som estimeres.(Bjørnstad,J, Store Norske Leksikon, 26.06.18, konfidensintervall – Store norske leksikon (snl.no))

Hvis du kaster terningen 40 ganger så har du lange konfidensintervaller og et usikkert gjennomsnitt. kaster du den derimot 40 000 ganger så får du kortere konfidensintervall og vi blir sikrere på hvor gjennomsnittet faktisk ligger. om du e 30 % sikker på om det er gjennomsnittet eller om du er 96 % sikker.

Kilder:

Inspirasjon av koder er fra seminaroppgaver og forelesninger, datacamp, pensumboken, samt egen kunnskap. Der direkte kopiert måte er dette skrevet før koden brukes.

Helbæk, Morten, Statistikk Kort og godt, 3 utg- 4 opplag 2019)

Matematikk.net, 20.09.11, lest 12.12.22, Permutasjon – Matematikk.net

Bjørnstad,Jan, Store Norske Leksikon, 26.06.18, lest 12.12.22, konfidensintervall – Store norske leksikon (snl.no))

My Race,blog and life, Standardfeil | hva det Er, Hvorfor Det Betyr noe, Og Hvordan Man Beregner | My Race (racem.org) lest 12.12.22.

Spørsmålene under oppgavene er skrevet inn fra eksamenoppgaven for bedre oversikt i dokumentet.