

CA07 – STAT100

Fábio Rodrigues Pereira

Exercise 0: Corrections for CA06:

- (1b) ii) $Y_{ij} = \mu_i + \epsilon_{ij} \rightarrow \epsilon_{ij} = Y_{ij} - \mu_i \rightarrow \text{for } \hat{\mu}_i = \bar{Y}_i \rightarrow \epsilon_{ij} = Y_{ij} - \bar{Y}_i := \text{residual where } \epsilon_{ij} \sim N(0, \sigma^2)$ \square
 iii) ϵ_{ij} is unknown because μ_i is unknown \square
- (1H) $FSS = \sum_{i=1}^k \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2 = \text{turns big when there are large differences between groups;}$
 $RSS = \sum_{i=1}^k \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2 = \text{turns big when there are large differences within a group.}$ \square
- (1I) $Y_{ij} = \mu_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$ for $i = 1, \dots, k = 4$ and $j = 1, \dots, m = 10$ \square
- (1I or) Conclusion: Since $F = 13,70 > F_{3, 30, \frac{1-\alpha}{2}} = 3,5$ for $\alpha = 0,05$, then reject H_0 . The averages are different. \square
- (2c) As $m \rightarrow \infty$, $F \rightarrow \infty$ and it becomes easier to reject H_0 . \square
- (2d) As $\sigma \rightarrow 0$, $F \rightarrow 0$ and it becomes harder to reject H_0 . \square
- (3) $FSS := \text{Variance between groups; } RSS := \text{Variance within a group; } TSS := \text{total variance}$
 $\mu := \text{population average; } \bar{x} := \text{sample average; } \sigma := \text{population standard deviation; } SD := \text{sample standard deviation}$
 $\epsilon_{ij} := \text{residual}$ \square

Oppgave 1

Gjør en analyse av dataene der du:

- Setter opp en sannsynlighetsmodell for responsen
- Hva er gruppe i denne oppgaven og forklar hva vi mener med gruppe-effekt

① $\rightarrow Y_{ij} = \mu_i + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$ for $i = 1, 2$ and $j = 1, 2, 3$ \square

② \rightarrow Aa (A), Lager (L) and Strong Lager (S) are the 3 groups of the research where we check which one is most preferred. The effect of groups aim to verify if there are any significant difference between the groups. \square

- Estimerer og gi en tolkning av alle ukjente parametere i modellen

$$\hookrightarrow E[\hat{\mu}_i] = \frac{1}{m_i} \cdot \sum_{j=1}^{m_i} Y_{ij} = \bar{Y}_i \xrightarrow{\text{unbiased estimator}} \text{estimate individual population mean}$$

$$\hookrightarrow E[\hat{\mu}_{ij}] = \frac{1}{k \cdot m} \sum_{i=1}^k \sum_{j=1}^{m_i} Y_{ij} = \bar{Y} = \mu_{ij} \xrightarrow{\text{unbiased estimator}} \text{estimate all population variances.}$$

$$\hookrightarrow E[\hat{\sigma}^2] = \frac{\frac{(m_1-1)}{(m_1-1)} \cdot \sum_{j=1}^{m_1} (Y_{1j} - \bar{Y}_1)^2 + \frac{(m_2-1)}{(m_2-1)} \cdot \sum_{j=1}^{m_2} (Y_{2j} - \bar{Y}_2)^2}{m_1 + m_2 - 2} = \frac{\sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2}{m_1 + m_2 - 2} = \frac{ESS}{m-k} = MSE = \sigma^2 \xrightarrow{\text{unbiased estimator}} \text{estimate within population variances.}$$

$$\hookrightarrow E[\hat{\sigma}^2_B] = \frac{GSS}{k-1} = MSG \xrightarrow{\text{estimate between population variances.}} \square$$

$$\textcircled{c} \text{ for Women: } \bar{Y}_A = \frac{3,00 + 3,56 + 4,44}{3} = 3,66 //$$

$$\bar{Y}_S = \frac{4,22 + 3,67 + 3,71}{3} = 3,88 //$$

$$\bar{Y}_L = \frac{5,31 + 5 + 5,22}{3} = 5,17 //$$

$$\bar{Y} = \frac{3 + 3,56 + 4,44 + 4,22 + 3,67 + 3,71 + 5,31 + 5 + 5,22}{9} = 4,236 //$$

$$\hookrightarrow Y_{ij} = \bar{Y}_i + \varepsilon_{ij} \rightarrow \varepsilon_{ij} = Y_{ij} - \bar{Y}_i$$

$$\hookrightarrow Y_{ij} - \bar{Y} = \bar{Y}_i - \bar{Y} + Y_{ij} - \bar{Y}_i \rightarrow \sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y})^2 = \sum_{i=1}^k \sum_{j=1}^{m_i} (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2$$

$\underbrace{\phantom{\sum_{i=1}^k \sum_{j=1}^{m_i}}}_{TSS}$ $\underbrace{\phantom{\sum_{i=1}^k \sum_{j=1}^{m_i}}}_{GSS}$ $\underbrace{\phantom{\sum_{i=1}^k \sum_{j=1}^{m_i}}}_{ESS}$

$$\hookrightarrow GSS = \sum_{i=1}^k \sum_{j=1}^{m_i} (\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^k m_i (\bar{Y}_i - \bar{Y})^2 = 3 \cdot (3,66 - 4,236)^2 + 3 \cdot (3,88 - 4,236)^2 + 3 \cdot (5,17 - 4,236)^2 \underset{4}{=} 1,29 \rightarrow MSG = \frac{4}{k-1} = \frac{4}{2} = 2 //$$

$$\hookrightarrow ESS = (5,73 - 6,06)^2 + (6,33 - 6,06)^2 + (6,12 - 6,06)^2 + (4,33 - 4,62)^2 + \dots + (5,44 - 4,55)^2 = 1,29 \rightarrow MSE = \frac{1,29}{m-k} = \frac{1,29}{6} = 0,215 //$$

$$\hookrightarrow SE[\sigma^2] = \sqrt{MSE} \approx 0,46 //$$

d. Beregner R^2 og forklar hva den sier oss

$$\hookrightarrow R^2 = \frac{GSS}{TSS} := \text{proportion of explained variation (between groups).}$$

$$\hookrightarrow \frac{4}{4+1,29} = \frac{4}{5,29} = 0,75 = 75\% \rightarrow \text{The variance between groups explains 75\% of the total variance.} //$$

e. Definerer hypoteser for om det er forskjell mellom øltyper.

$$\hookrightarrow H_0: \mu_A = \mu_L = \mu_S \quad vs \quad H_1: \text{not all averages are equal} //$$

f. Velg 5 % nivå, gjør en hypotesetest for om det er forskjell mellom øltypene og konkluder. Finn p-verdien og gi en forklaring av hva denne betyr til en person som er interessert i øl, men som ikke har noe særlig statistisk kunnskap.

$$\hookrightarrow \alpha = 0,05$$

\hookrightarrow Statistic test := $F = \frac{MSG}{MSE} := \text{ratio of the variance between groups estimator and the variance within groups estimator.}$

$$\hookrightarrow \frac{2}{0,215} = 9,3$$

\hookrightarrow Since the test statistic $F = 9,3 > F_{\text{critical}} = 5,14$, then reject H_0 .

$$\hookrightarrow p\text{-value} = P(F > F_{0,05, 2, 6} \mid H_0 \text{ is true}) = 0,014$$

\hookrightarrow Since $p\text{-value} = 0,014 < \alpha = 0,05$, then reject H_0 .

\hookrightarrow There are significant differences between the beer types.

\hookrightarrow This means that the women really felt significant differences in the beers types. //

- g. Hva er en kontrast i forventninger generelt? Les fila kontraster.pdf som du finner på Canvas.

↳ *Contrast is a linear combination of averages. This aims to get some info if there are any difference between groups.*

- h. Konstruer en kontrast for å se om det er forskjell mellom Lager typene

↳ We know that $\sum_{i=1}^k c_i \cdot \mu_i$ where $\sum_{i=1}^k c_i = 0$, then $\theta = 0 \cdot \mu_A + 1 \cdot \mu_S - 1 \cdot \mu_L$ for $i = A, S, L$.

- i. Estimer kontrasten og finn dens standardfeil.

$$\rightarrow E[\hat{\theta}] = \sum_{i=1}^k c_i \bar{Y}_i = (0 \cdot Y_A) + (1 \cdot 4,55) + (-1 \cdot 4,62) = -0,07 = \mu_\theta$$

$$\rightarrow S_p^2 = MSE = \frac{ESS}{m-k} = \frac{\sum_{i=1}^k \sum_{j=1}^{m_i} (Y_{ij} - \bar{Y}_i)^2}{m-k} = \frac{1,29}{4} = 0,21547 \text{ for women}$$

$$\rightarrow SE[\hat{\theta}] = \sqrt{S_p^2 \cdot \sum_{i=1}^k \frac{c_i^2}{m_i}} = \sqrt{0,216 \cdot \frac{2}{3}} \approx 0,38 \text{ for women}$$

- j. Test ved bruk av kontrasten om det er forskjell mellom de to øltypene (bruk 5 % nivå).

$$\rightarrow H_0 := \theta = 0 \quad v.s. \quad H_1 := \theta \neq 0$$

$$\rightarrow \alpha = 0,05$$

$$\rightarrow \text{Test statistic } T = \frac{\hat{\theta} - \mu_\theta}{SE[\hat{\theta}]} = \frac{-0,07}{0,38} = -0,18 \text{ for women}$$

$$\rightarrow t_{6;0,025} = 2,447$$

↳ Since $|T| = 0,18 < t_{6;0,025} = 2,447$, then hold H_0 . → *Not significant differences between Lager beers.*

- k. Etter at du har utført h-j sjekk resultatene ved Rstudio.

↳ OK!

Extra - Exercise 2:

```
17 - ***{r}
18  anova(lm(beerf$Poeng~beerf$type))
19  ```

Analysis of Variance Table

Response: beerf$Poeng
          Df Sum Sq Mean Sq F value Pr(>F)
beerf$type  2  0.0362  0.01810  9.3662 0.01428 *
Residuals   6  1.2928  0.21547
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

20 - ***{r}
21  anova(lm(beerm$Poeng~beerm$type))
22  ```

Analysis of Variance Table

Response: beerm$Poeng
          Df Sum Sq Mean Sq F value Pr(>F)
beerm$type  2  0.3692  0.18458  6.1016 0.03581 *
Residuals   6  0.21482 0.35803
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

23 - ***{r}
24  qf(0.95, 2, 6)
25  ```

[1] 5.143253
```

for men: Test statistic $F = 6,10$ | $p\text{-value} = 0,035$

$$F\text{-critical} = 5,14$$

↳ Since test statistic $F = 6,10 > F\text{-critical} \approx 5,14$, then reject H_0 ;

↳ Since $p\text{-value} = 0,03 < \alpha = 0,05$, then reject H_0 .

↳ There are significant differences between the beer types.

↳ This means that the men really felt significant differences in the beer types, but less than the women.

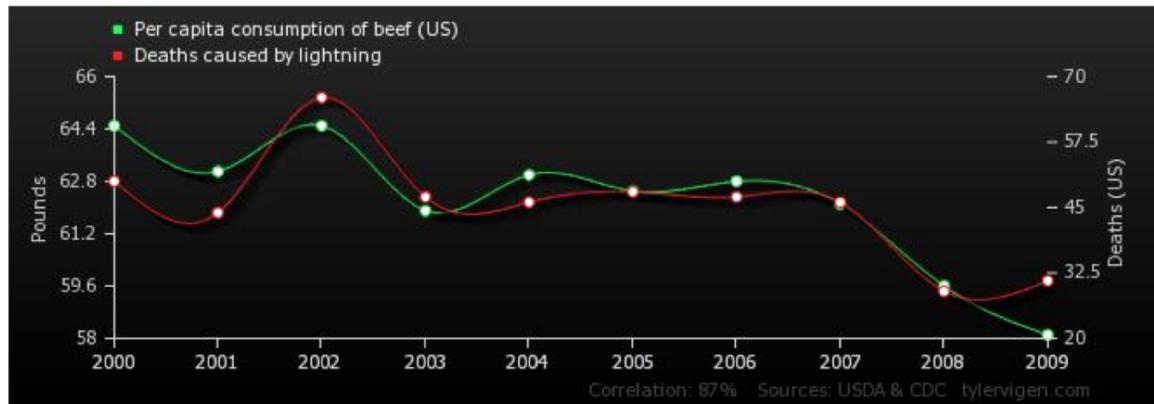
Korrelasjon:

Oppgave 3

Vi skal se litt nærmere på et mye brukte statistisk mål på samvariasjon, nemlig korrelasjon. Vi hører mye om korrelasjon i media, men vi skal være klar over at høy korrelasjon ikke nødvendigvis betyr at en variabel er årsak til den andre.

På nettsiden snl.no (Store Norske Leksikon) finner dere et oppslag om korrelasjon, <https://snl.no/korrelasjon> Les dette og oppklar for hverandre eventuelle vanskeligheter.

Oftest kan det oppstå merkelige, fåpelige eller morsomme eksempler på samvariasjon, såkalte «spurious correlations» og en morsom side med eksempler på slike kan dere finne her: <http://tylervigen.com/spurious-correlations>



Korrelasjon henger sammen med hvor godt punkter passer til en rett linje.

Skriv inn danske data for antall hekkende storker og barnefødsler (i 1000) for årene 1946, 1947, 1948, 1949 og 1950, so gitt i R-script-fila.

- ① Plott barnefødsler mot antall hekkende storker. Ser det ut som om punktene ligger på en rett linje?
- ② Finn korrelasjonen mellom x og y.
- ③ Er det årsak-virkning? (Altså: (1) Øker barnetallet fordi storketallet øker?, (2) Øker storketallet fordi barnetallet øker? Eller (3) Er det andre forklaringer? Hvilke?)

④ Yes, the plot between these groups show a positive correlation which the points lies on a straight line //

$$\bar{x} = \frac{51+54+58+61+62}{5} = 57 \rightarrow \text{barn}$$

$$\bar{y} = \frac{100+130+151+170+180}{5} = 146,2 \rightarrow \text{stork}$$

⑤ The correlation is very strong 0,99 //

$$r = \frac{\text{covariance between } X \text{ and } Y}{S_x \cdot S_y} = \frac{\cancel{1} \cdot \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\cancel{1} \sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\cancel{1} \sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{[(51-57) \cdot (100-146,2) + \dots + (62-57) \cdot (180-146,2)]}{\sqrt{(51-57)^2 + \dots + (62-57)^2} \cdot \sqrt{(100-146,2)^2 + \dots + (180-146,2)^2}} = 0,99 //$$

⑥ Despite the numbers show strong correlation, this is another case that both groups have not rational relationships, then it is not possible to say that they are correlated! //

Oppgave 4

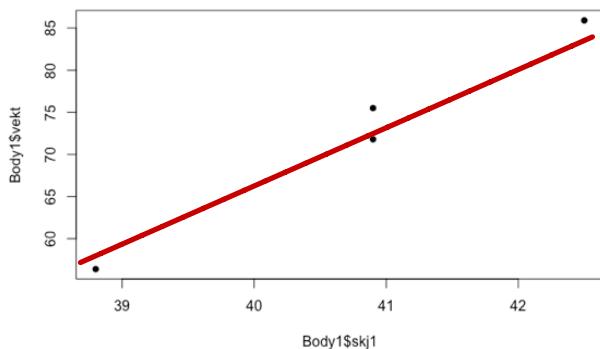
Vi skal se på et datasett «Body1.Rdata» som inneholder diverse kroppsmaål, såkalt «antropometriske mål», blant annet diverse lengder av skjelettdeler i kroppen (9 variabler) og diverse omkretser, f.eks livvidde og omkrets rundt håndledd (12 variabler). Dessuten er det målt høyde og vekt, og vi har også en variabel som viser alder på personene. I første omgang har man bare målt

$n = 4$ personer.

se under moduler > data) som et eksempel på korrelasjoner mellom

- Last ned datasettet fra Canvas (Moduler – Data). Åpne Rstudio (hvis det ikke allerede er åpent og åpne fila Body1.Rdata fra filmenyen øverst til venstre). ✓
- Vi er ute etter å finne gode variabler som kan brukes til å si noe om vekten til personer. En kandidat er skjelett-variabelen «skj1». Lag et plott av «skj1» mot «vekt». Beskriv sammenhengen mellom de to variablene og gjett på korrelasjonen mellom dem (dvs et tall mellom -1 og 1)

	Person 1	Person 2	Person 3	Person 4
Skj1	38.8	42.5	40.9	40.9
Vekt	56.4	85.9	71.8	75.5



→ It seems a positive correlation. I would guess correlation = 80%. ☐

- Kontrollér gjetningen din ved å beregne korrelasjonen i Rstudio. = 0,992 //

Det finnes en enkel nettside som heter guessthecorrelation.com der du kan «spille» denne gjetteleken og trenere opp din egen magefølelse. Det kan faktisk være ganske nyttig. Søk opp siden, lag deg et brukernavn og prøv det ut (det kan være nødvendig å skru av lyden på mobilen eller PC-en mens du prøver dette.) ✓

- Beregn korrelasjonen mellom vekt og alle skjelettvariablene. Du skal da få ut en stor tabell med korrelasjoner. I første kolonne og første rad finner du korrelasjonene mellom vekt og alle de andre variablene.

Gjør tilsvarende med omkretsvariablene.

Finn de 3 skjelett- og omkretsvariablene som har høyest korrelasjon med vekt.

Body1\$vekt	skj1	skj2	skj3	skj4	skj5	skj6	skj7	skj8	skj9
1.00	0.99	0.67	0.71	0.67	0.92	0.95	0.82	0.88	0.93
skj1	0.99	1.00	0.62	0.66	0.58	0.92	0.95	0.80	0.84
skj2	0.67	0.62	1.00	1.00	0.44	0.86	0.41	0.24	0.95
skj3	0.71	0.66	1.00	1.00	0.48	0.88	0.46	0.30	0.96
skj4	0.67	0.58	0.44	0.48	1.00	0.46	0.71	0.84	0.59
skj5	0.92	0.92	0.86	0.88	0.46	1.00	0.75	0.55	0.96
skj6	0.95	0.95	0.41	0.46	0.71	0.75	1.00	0.95	0.68
skj7	0.82	0.80	0.24	0.30	0.84	0.55	0.95	1.00	0.52
skj8	0.88	0.84	0.95	0.96	0.59	0.96	0.68	0.52	1.00
skj9	0.93	0.89	0.52	0.57	0.88	0.74	0.95	0.95	0.75

→ Highest correlation between vekt and skj9 with 0.93; vekt and skj6 with 0.95 and vekt and skj1 with 0.99. ☐

Body1\$vekt	omkr1	omkr2	omkr3	omkr4	omkr5	omkr6	omkr7	omkr8	omkr9	omkr10	omkr11	omkr12		
1.00	0.96	0.93	0.89	0.85	0.83	0.55	0.83	0.84	0.82	0.87	0.58	0.86		
omkr1	0.96	1.00	1.00	0.76	0.67	0.65	0.48	0.96	0.96	0.85	0.95	0.72	0.93	
omkr2	0.93	1.00	1.00	0.70	0.61	0.62	0.52	0.97	0.98	0.81	0.97	0.70	0.90	
omkr3	0.89	0.76	0.70	1.00	0.98	0.84	0.28	0.55	0.54	0.83	0.55	0.52	0.76	
omkr4	0.85	0.67	0.61	0.98	1.00	0.91	0.38	0.43	0.44	0.70	0.48	0.34	0.63	
omkr5	0.83	0.65	0.62	0.84	0.91	1.00	0.71	0.41	0.46	0.46	0.58	0.06	0.45	
omkr6	0.55	0.48	0.52	0.28	0.38	0.71	1.00	0.40	0.48	-0.01	0.65	-0.25	0.12	
omkr7	0.83	0.96	0.97	0.55	0.43	0.41	0.40	1.00	1.00	0.79	0.95	0.78	0.90	
omkr8	0.84	0.96	0.98	0.54	0.44	0.46	0.48	1.00	1.00	0.75	0.98	0.72	0.87	
omkr9	0.82	0.85	0.81	0.83	0.70	0.46	-0.01	0.79	0.75	1.00	0.65	0.91	0.98	
omkr10	0.87	0.95	0.97	0.55	0.56	0.48	0.58	0.65	0.95	0.98	0.65	1.00	0.56	0.79
omkr11	0.58	0.72	0.70	0.52	0.34	0.06	-0.25	0.78	0.72	0.91	0.56	1.00	0.91	
omkr12	0.86	0.93	0.90	0.76	0.63	0.45	0.12	0.90	0.87	0.98	0.79	0.91	1.00	

→ Highest correlation between vekt and omkr1 with 0,96; vekt and omkr2 with 0,93 and vekt and omkr3 with 0,89. //

- e) Last inn datasettet Body2.Rdata som inneholder samme variablene målt på n = 503 andre personer.

Gjenta beregningen av korrelasjonene i oppgave d, men nå med dette store datasettet.

- i) Hva er korrelasjonen mellom skj1 og vekt nå?
- ii) Hvilke 3 variabler er nå mest korrelert med vekt?
- iii) Som vi ser var ikke resultatene fra Body1 så pålitelige. Hvorfor?

i) 0,725 //

Body2\$vekt	skj1	skj2	skj3	skj4	skj5	skj6	skj7	skj8	skj9	
1.00	0.73	0.50	0.67	0.80	0.83	0.80	0.76	0.77	0.73	
skj1	0.73	1.00	0.31	0.49	0.58	0.77	0.77	0.72	0.64	0.66
skj2	0.50	0.31	1.00	0.67	0.36	0.33	0.32	0.28	0.43	0.37
skj3	0.67	0.49	0.67	1.00	0.47	0.53	0.53	0.47	0.61	0.50
skj4	0.80	0.58	0.36	0.47	1.00	0.67	0.67	0.61	0.55	0.60
skj5	0.83	0.77	0.33	0.53	0.67	1.00	0.76	0.73	0.66	0.67
skj6	0.80	0.77	0.32	0.53	0.67	0.76	1.00	0.84	0.73	0.82
skj7	0.76	0.72	0.28	0.47	0.61	0.73	0.84	1.00	0.71	0.77
skj8	0.77	0.64	0.43	0.61	0.55	0.66	0.73	0.71	1.00	0.72
skj9	0.73	0.66	0.37	0.50	0.60	0.67	0.82	0.77	0.72	1.00

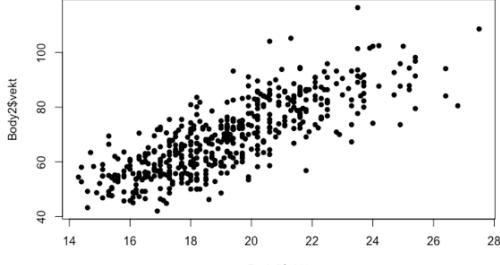
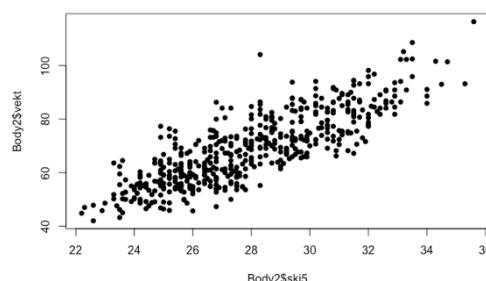
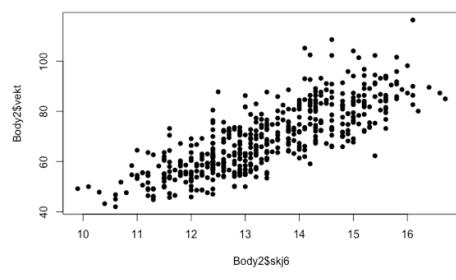
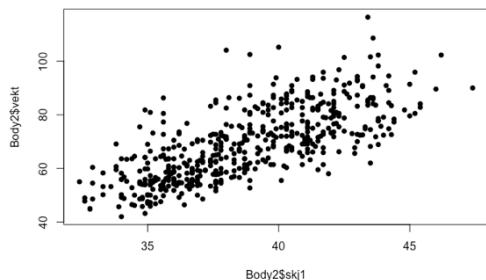
→ skj4 with vekt; skj5 with vekt; skj6 with vekt. //

- iii) The results on Body1 is not reliable, because there are very little number of samples. //

- f) Lag et nytt plott (som i oppgave a) av vekt mot skj1 for det store datasettet og et annet plott av vekt mot den skelett/omkretsvariabelen som er mest korrelert med vekt. Bruk R-koden

plot(Body2\$XXX, Body2\$vekt, pch=16)

der XXX byttes ut med variabelnavnet til den variabelen som er høyest korrelert med vekt. Ser det ut til at det er en lineær sammenheng?



→ all plots have quite the same shape where shows a positive and strong correlation! //

- g) Korrelasjoner vil selvsagt ofte gjenspeile årsaks-/virkningsforhold. Diskuter dere fram til et par eksempler på kontinuerlige variabler som forventes å være positivt (eller negativt) korrelerte som følge av et årsaks-/virkningsforhold. (Tips: Du finner noen eksempler i artikkelen på snl.no, men finn på noen flere selv.)

↳ One example is the color of the eyes of the parents and the color of the eyes of the children. ↘