

CA08 - STAT 100

Fábio Rodrigues Pereira

Oppgave 0: CA07's corrections:

1a) $Y_{ij} = \mu_i + \epsilon_{ij}$; $\epsilon_{ij} \sim N(0, 1)$ where Y_{ij} is independent; $i=1, 2, 3$ and $j=1, 2, 3$

'A', 'L', 'S' points

$$\begin{aligned} 1a) \quad Y_1 &= \frac{3+3,56+4,44}{3} \approx 3,67 \\ Y_2 &= \frac{5,31+5+5,22}{3} \approx 5,18 \\ Y_3 &= \frac{4,22+3,67+3,71}{3} \approx 3,87 \end{aligned} \quad \rightarrow \hat{\theta} = \sum_{i=1}^3 \alpha_i \cdot \bar{Y}_i = \sum_{i=1}^3 \alpha_i \cdot \bar{Y}_i = (0 \cdot 3,67) + (1 \cdot 5,18) + (-1 \cdot 3,87) \approx 1,31 //$$

$$SE[\hat{\theta}] = \sqrt{MSE \sum_{i=1}^3 \frac{\alpha_i^2}{m_i}} = \sqrt{\sum_{i=1}^3 \sum_{j=1}^3 (\bar{Y}_{ij} - \bar{Y}_i)^2 \cdot \sum_{i=1}^3 \frac{\alpha_i^2}{m_i}} = \sqrt{\frac{(3-3,67)^2 + (3,56-3,67)^2 + (4,44-3,67)^2 + (5,31-5,18)^2 + (5-5,18)^2 + (5,22-5,18)^2 + (4,22-3,87)^2 + (3,67-3,87)^2 + (3,71-3,87)^2}{6} \cdot \left(\frac{0^2}{3} + \frac{1^2}{3} + \frac{(-1)^2}{3}\right)} \\ \rightarrow = \sqrt{\frac{1,30}{6} \cdot \frac{2}{3}} = 0,38 //$$

1b) $H_0: \theta = 0$ vs $H_1: \theta \neq 0 \rightarrow T^* = \frac{1,31-0}{0,38} \approx 3,45 \rightarrow |T^*| = |3,45| > t_{6, 0,025} \approx 2,45 \Rightarrow \text{reject } H_0$

Conclusion: We are 95% sure that there is significant difference between the expected averages between the been type longer and strong longer.

Oppgave 1:

Priser på bruktbiler av samme type avhenger blant annet av kjørelengde og alder, og vi skal i denne oppgaven prøve å si noe om hvor mye prisen på en bil faller etter hvert som kjørelengden og alderen øker. For å finne ut mer om dette skal vi hente inn data på biler fra <http://finn.no>.

- Gå inn på nettsiden og bruk valgmulighetene til venstre til å begrense søker på bruktbiler til:
 - Volkswagen Golf \rightarrow Mini Cooper 1,6 turbo 3 doors

I listen over søker treff plukk ut de **10 første** og skriv opplysningene inn i et excel-ark som vist i eksemplet nedenfor (finn.no oppdateres kontinuerlig, så du vil få andre data enn det som vises i dette skjermbildet), der «pris» er gitt i 1000 kr, «alder» angir hvor gammel bilen er (årsmodell sammenlignet med det inneværende året) og «km» er kjørelengde i antall 1000. Lagre fila under navnet bildatafil.xlsx ✓

Bruk passende deskriptiv statistikk til å beskrive variablene i datasettet. ✓

Du finner det du trenger for å gjøre analysene i Rstudio i R-script-fila ✓

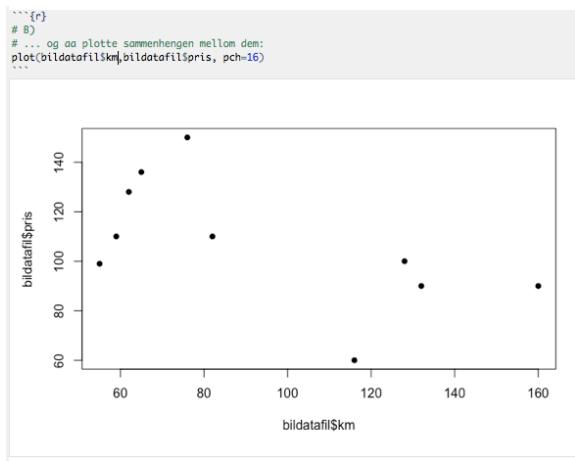
Modul11_Kollokvieoppgaver.R på Canvas. Jobb deg gjennom kommandoene og kommentarene i fila (det er noen spørsmål der også) parallelt med at du løser oppgavene under. ✓

A	B	C	D	E
1	alder	km	pris	
2	3	65	165	
3	12	147	34	
4	1	35	531	
5	14	201	17	
6				

→ Explanatory variables: alder and km
 → response variable: pris

b. ① Lag et plott som viser sammenhengen mellom pris og kjørelengde. Synes du en linearitetsantakelse er en rimelig antakelse for sammenhengen?

② Skriv opp en enkel lineær regresjonsmodell med pris som responsvariabel og km som forklaringsvariabel. Beskriv de 4 antagelsene dere gjør når dere spesifiserer denne modellen.



① → Not possible to say that it has a linearity assumption from the plot, because we have a very sparse dataset. However, that is reasonable to think that there is a correlation between the km and price of the car. //

② → $Y_i = \alpha + \beta X_i + \epsilon_i$; $\epsilon_i \sim N(0, 1)$, independent and $i = 1, \dots, 10$; Where $Y_{i,i}$:= pris and $X_{i,i}$:= km //

③ ↑ pris when ↑ km; ④ ↓ pris when ↑ km;

⑤ ↑ pris when ↓ km; ⑥ ↓ pris when ↓ km

◻

c. Tilpass den lineære modellen i Rstudio.

Finn estimatorer **for alle** modellparametrene, og tolk disse i lys av problemet.

estimaterer :=

$$\hat{Y} = \frac{110+136+100+90+150+99+150+99+110+90+60+128}{10} = 107,3 // \rightarrow \text{pris mean}$$

$$\bar{X} = \frac{59+65+128+160+76+55+82+132+116+62}{10} = 93,5 // \rightarrow \text{km mean}$$

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{(110-107,3)(59-93,5) + (136-107,3)(65-93,5) + \dots + (128-107,3)(128-93,5)}{(59-93,5)^2 + (65-93,5)^2 + \dots + (62-93,5)^2} = -0,4118164 //$$

→ The expected change of the response variable (pris) when the explanatory variable (km) increases 1 unit.

$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} = 107,3 - (-0,4118164 \cdot 93,5) = 145,8048$ → The expected value of the response variable (pris) when the value of the explanatory variable (km) is equal to 0.

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i$$

$$RSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{10} (\hat{\alpha} + \hat{\beta} x_i - \bar{Y})^2 = 2122,708 //$$

→ Regression sum of squares.

$$ESS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{10} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = 3925,392 //$$

→ Error/Residual sum of squares.

$$TSS = RSS + ESS = 6048,1 // \rightarrow \text{Total sum of squares}$$

$$R^2 = \frac{2122,708}{6048,1} = 0,35091 // \rightarrow \text{This is how good is the regression.}$$

$$\hat{\sigma}^2 = S^2 = \frac{ESS}{n-2} = MSE = \frac{3925,392}{8} \approx 390,54$$

→ Mean square error

$$SE[\hat{\beta}] = \sqrt{\frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{390,54}{\sum_{i=1}^{10} (x_i - \bar{x})^2} \approx 0,1766$$

more precise than above.

$$\hat{\sigma} = \sqrt{MSE} \approx 19,76207 //$$

→ Residual standard error

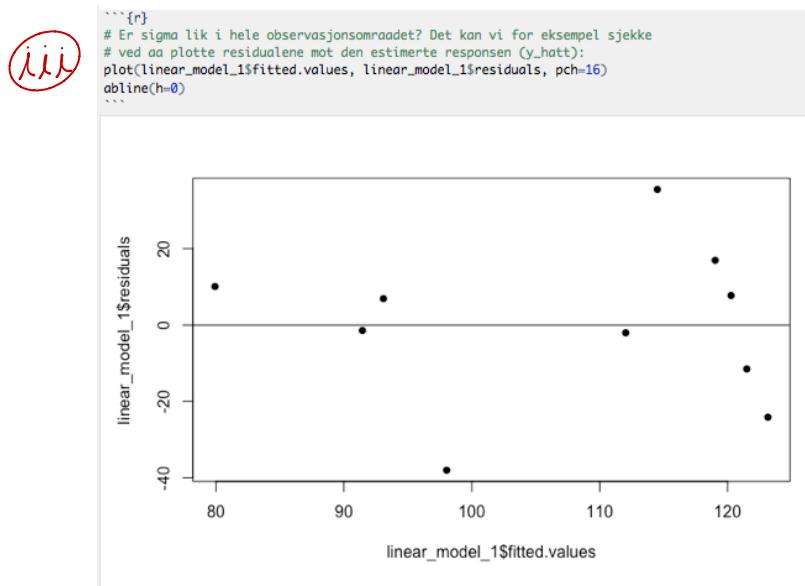
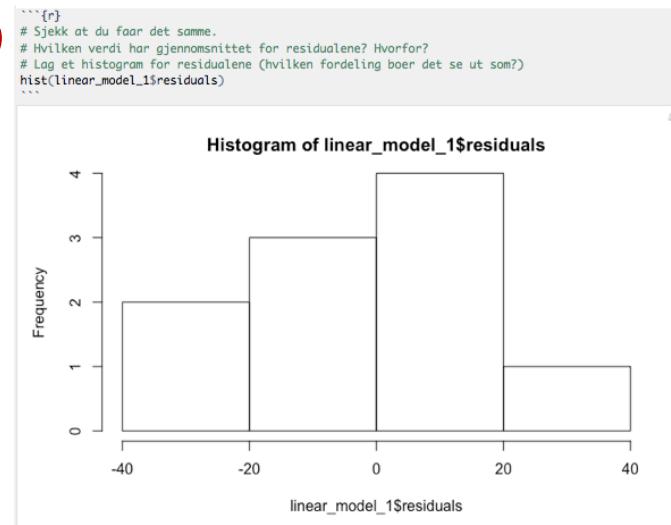
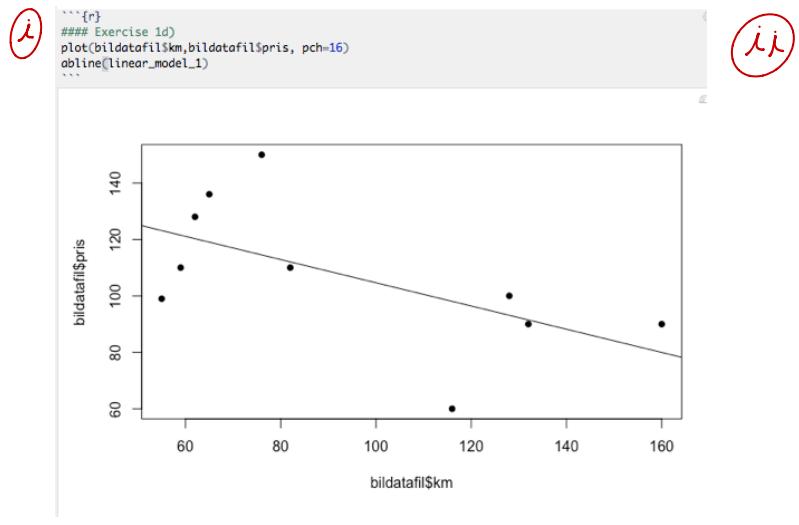
◻

d. ① Lag et plott av pris mot kilometerstand, som også har den estimerte regresjonslinja.

② Se på den deskriptive statistikken og histogrammet for residualene.

③ Plott residualene mot de tilpassede verdiene \hat{y}_i .

④ Kommenter de fire modellantakelsene på bakgrunn av det du vet om om dataene og det du ser i utskriftene.



→ From ① we can see that the price has a negative correlation against the km. This means that as km↑ as pris↓

- e. Hvor stor er R^2 og hvordan tolker du denne? → done in ①c $R^2 = 0,35$ → measures how good is the regression. □
- f. En student bor i Moss og kjører hver dag sin Golf til NMBU. Hvor stort daglig prisfall vil du anslå at han har på sin bil hvis vi antar at det er ca. 3 mil hver vei.

Since $\hat{\beta} = -0,41$ which is the decrease of the car price for each km, then for 60 km it will be a expected decrease of $\hat{\beta} \cdot x_i = -0,41 \cdot 0,0060 = -0,00246$ or 24 kr/day. □

- g. Lag en ny modell, men nå med alder som forklaringsvariabel for pris.
 Gjøre deloppgavene b - e om igjen men med alder som forklaringsvariabel
 Sammenlign R^2 med tilsvarende for modellen med kjørelengde som
 forklaringsvariabel.
 Hvilken modell av de to du har tilpasset forklarer mest av variasjonen i pris?

```
```{r}
Exercise 1g)
linear_model_2 = lm(bildatafil$pris~bildatafil$alder)
summary(linear_model_2)
```
```

```
Call:
lm(formula = bildatafil$pris ~ bildatafil$alder)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.355 -14.898 -2.111  7.570 42.645 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 166.632    29.396   5.669 0.000471 ***
bildatafil$alder -8.128     3.811  -2.078 0.071305 .  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.16 on 8 degrees of freedom
Multiple R-squared:  0.3506   Adjusted R-squared:  0.2694 
F-statistic: 4.319 on 1 and 8 DF,  p-value: 0.07131
```

→ This time $R^2 = 0.3506$ which is almost the same as the previous regression when we got $R^2 = 0.351$.

Then, both explanatory variables has quite the same impact on the car's price. ☐

- h. Hvilke andre faktorer som kunne påvirke prisen på en bil? Hva tror du skjer med standardavviket (σ) til støyleddet (ε) hvis du «forfiner» søket i finn.no ved å legge til flere forklaringsvariable?

→ There are many other factors, like the mechanical conditions, accessories and etc...

→ If we add more data or/and explanatory variables, the regression will be better with higher R^2 . Also, the residual will be more symmetrical getting closer to the standard normal distribution and the standard error will be reduced. ☐

Extra:

```
```{r}
Exercise 1g)
linear_model_3 = lm(bildatafil$pris~bildatafil$alder+bildatafil$km)
summary(linear_model_3)
```
```

```
Call:
lm(formula = bildatafil$pris ~ bildatafil$alder + bildatafil$km)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.477 -13.812  1.347  7.936 27.543 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 181.9809    28.1216   6.471 0.000343 ***
bildatafil$alder -6.2029     3.7189  -1.668 0.139262  
bildatafil$km   -0.3144     0.1883  -1.670 0.138933  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.03 on 7 degrees of freedom
Multiple R-squared:  0.5356   Adjusted R-squared:  0.4029 
F-statistic: 4.036 on 2 and 7 DF,  p-value: 0.06828
```

→ less than before!

→ Increased as expected! ☐

Oppgave 2

Om å kunne være kritisk til modeller.

Hvis du tilpasser en lineær regresjonsmodell til data, kan du beregne residualer.

Residualene har et gjennomsnitt lik null (uavhengig av hvor godt eller dårlig modellen passer til data).

Et residualplot er gjerne et scatterplot med residualene på y-aksen.

På x-aksen kan du enten ha x eller tilpasset Y-verdi.

I tillegg kan vi ha histogram over residualene.

a. Generelt for lineær regresjon

- ① Skriv ned modellen for lineær regresjon. $\hat{Y}_i = \alpha + \beta X_i + \varepsilon_i$; $\varepsilon_i \sim N(0, \sigma)$, independent and $i = 1, \dots, n$
- ② Hva er tilpasset (Y) verdi for en gitt x-verdi?
- ③ Hva er et residual?

④ Hvilke krav setter vi til feilreddene (ε -ene) i en regresjonsmodell?

⑤ Hvorfor ser vi på residualene (e_i) og ikke direkte på feilreddene (ε_i)?

⑥ Y_i is a response variable which is dependent on X_i . X_i is an independent explanatory variable.

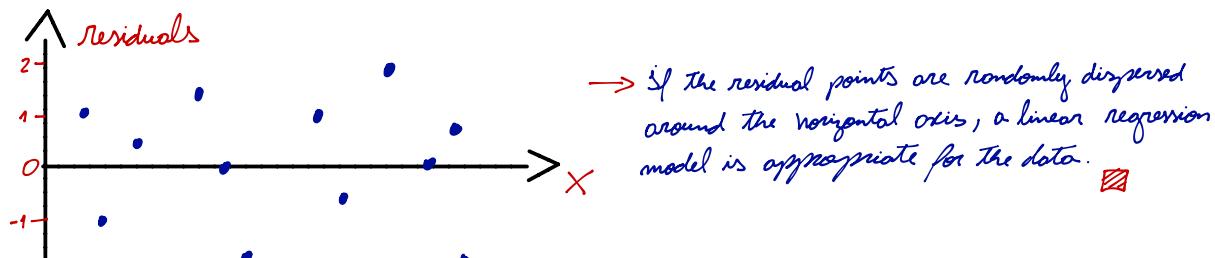
⑦ Residual is the difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}), such that $E_i = Y_i - \hat{Y}_i$.

⑧ There are 2 requirements: 1) ε_i are independent; 2) ε_i are normal distributed with mean equal to 0 and O .

⑨ ?

b.

Lag en skisse av et residualplot uten modellproblemer.



c.

Det var utført en undersøkelse om helse og levealder der vi har plukket ut 13 land.

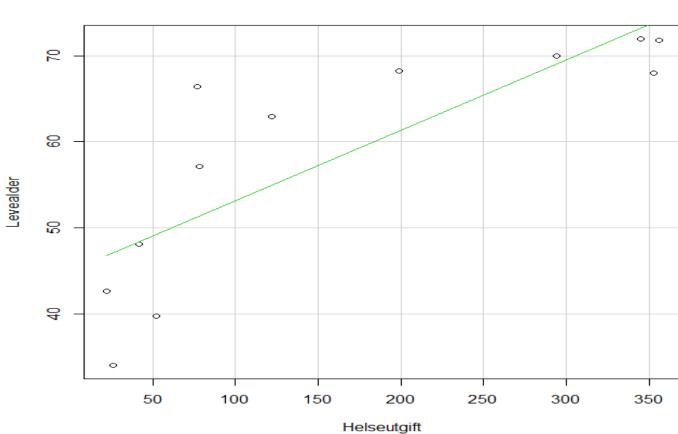
La responsvariabel (Y) være gjennomsnittlig levealder og forklaringsvariabel (x) være helseutgift, målt i innsats pr. person, der enheten er US \$.

| Land | Levealder (\bar{Y}_i) | Helseutgift (X_i) |
|--------------|---------------------------|-----------------------|
| Cameroon | 48,1 | 42 |
| Colombia | 71,8 | 356 |
| DominicanRep | 68,0 | 353 |
| Gambia | 57,1 | 78 |
| Indonesia | 66,4 | 77 |
| Malaysia | 72,0 | 345 |
| Mongolia | 62,9 | 122 |
| Niger | 42,6 | 22 |
| Samoa | 68,2 | 199 |
| Sierra-leone | 34,0 | 26 |
| Turkye | 70,0 | 294 |
| Zambia | 39,7 | 52 |

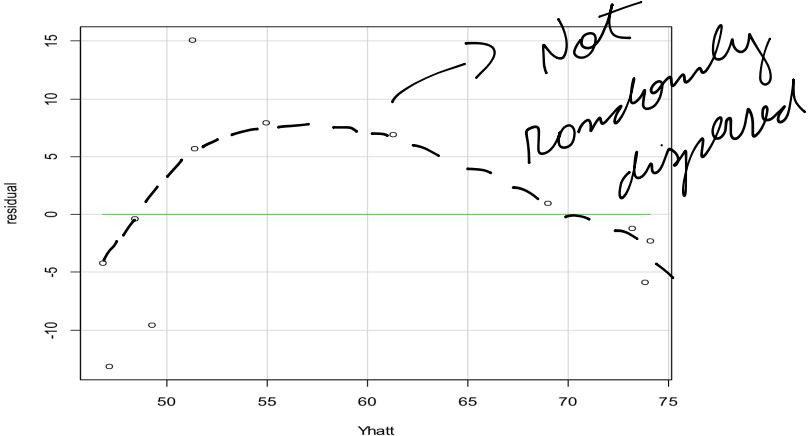
Data er plottet i figur 1a. Vi har kjørt en lineær regresjons modell, det ga residualplot vist i figur 1b.

- (i) Sett opp modellen og gi en tolkning av parametrene i modellen i lys av problemet.
- (ii) Bruk Figur 1 og kunnskap du måtte innha til å diskutere hvorfor påstanden 'En dollar økning i helsebudsjettet gir samme effekt uansett hvor rikt landet er' neppe er riktig men identisk med å bruke en lineær modell?
- (iii) Hvilke feil kan du lese ut av residualplottet (figur 1b)?

Har dere noen forslag på hva som burde vært gjort?



Figur 1a



Figur 1b: Residualer mot tilpassede verdier

$$i) Y_i = \alpha + \beta X_i + \epsilon_i; \epsilon_i \sim N(0, \sigma^2), \text{ independent}, \text{ for } i=1, \dots, 12 \rightarrow \text{The text says 13 countries, however the data table only has 12!}$$

response variable
explanatory variable
Residual
Expected value of Y_i when X_i increases 1 unit.

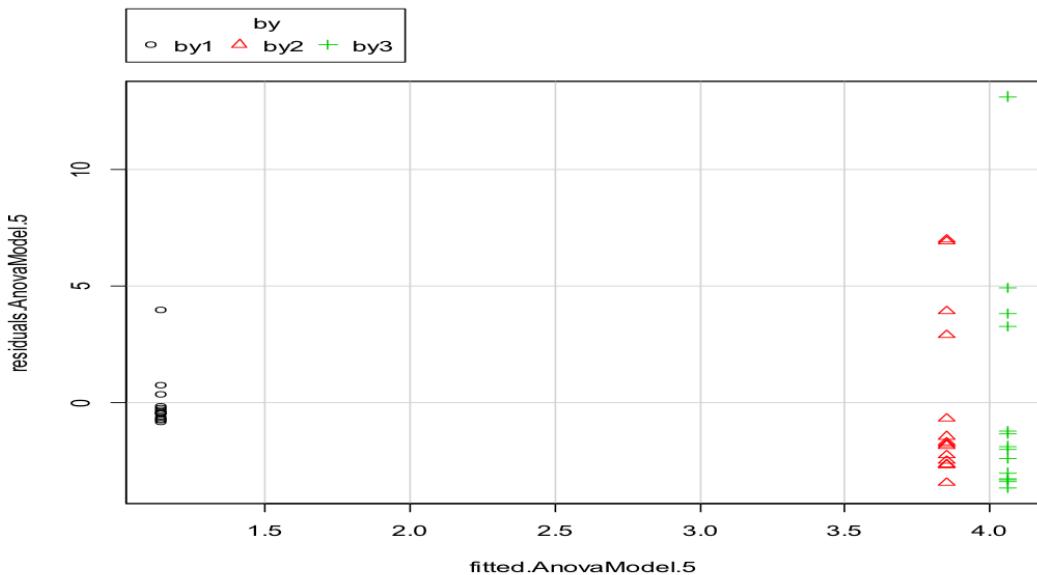
ii) The assumption is correct only if you disregards other relevant factors. As our regression model here only has 1 explanatory variable, so the assumption is correct for this case. //

iii) It is forming a kind of U-shape which indicates that the model is not linear. //

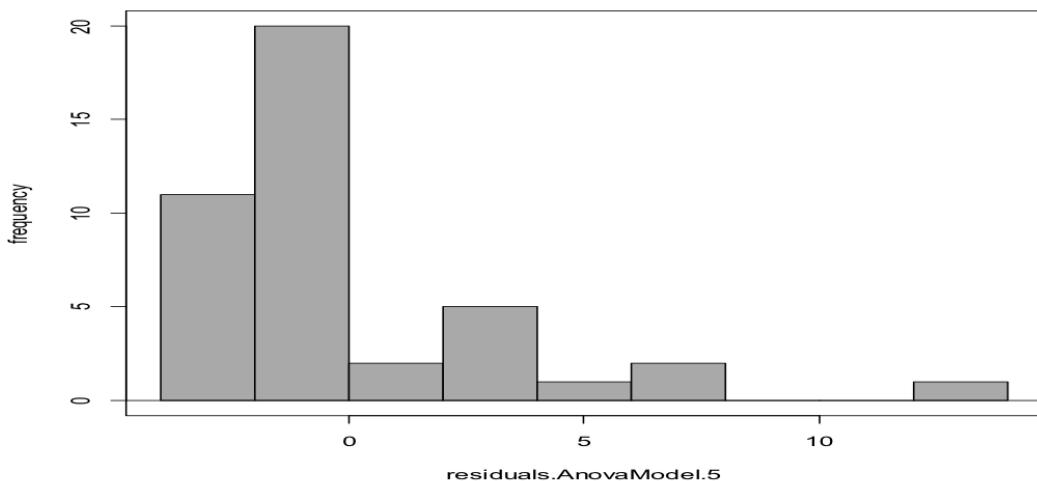
d.

Klorinnhold i tre forskjellige byer (målt i p.p.m., parts per million), på forskjellige vilkårige steder og tider i tre byer. Deretter ble det kjørt en enveis variansanalyse. Se figur 3 og figur 4.

Hvilke problemer har vi møtt på her. Hvordan kan du løse disse?



Figur 3. Residualer mot tilpassede verdier spørsmål d.



Figur 4. Histogram over residualer. d.

- The residuals are not normally distributed, because fig.4 plot is not symmetrical. Also, the residuals are not randomly dispersed on fig.3, which suggests that the residuals are not independent and non-linear.
- There are 2 way to transform the variables using a mathematical operation to change its measurement scale and achieve linearity, in other words, we try to increase the linear relationships between 2 variables. The first technique is called linear transformation where, for example, we multiply the variable by a constant and divide by the same constant. Here, the correlation between variables is preserved. The second technique is called nonlinear transformation where, for example, we take the square root of the variable or its reciprocal. Here, there correlation between variables is changed, but we can increase the linearity. ☐