Forklaring av (univariat) regresjonsmodell

Modellen

I eksemplet bruker vi assosiasjonen mellom VO_{2max} og høyde i `cyclingstudy`.

Modellen er:

$$\text{VO}_{2 \text{ max } [i]} \sim \text{Normal } (\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 \times \text{høyde}_i$$

```
# In R:

library(tidyverse)
```

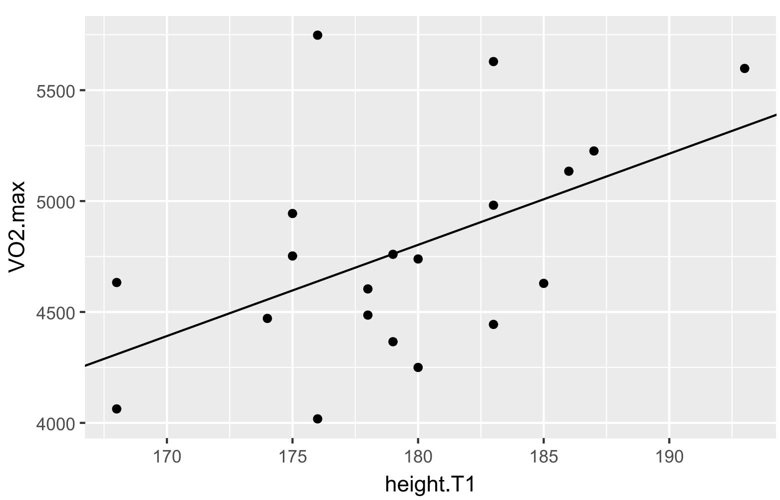
```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.0.4
── Conflicts ──────────────────────────────────────────── tidyverse_conflicts()
──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(gt)

# Load data
dat <- exscidata::cyclingstudy |>

      filter(timepoint == "pre") |>
      select(height.T1, VO2.max)



# Summary statistics
dat |>
      pivot_longer(everything()) |>
```

```r
    summarise(.by = name,
              m = mean(value),
              s = sd(value),
              n = n()) |>
    gt()
```

name	m	s	n
height.T1	179.300	6.122435	20
VO2.max	4773.831	494.074815	20

```r
# Fit model
m <- lm(VO2.max ~ height.T1, data = dat)

# Intercept
intercept <- coef(m)[1]

# Slope
slope <- coef(m)[2]

# Antall obs
n <- 20


# A plot
dat |>
    ggplot(aes(height.T1, VO2.max)) + geom_point() +

    geom_abline(intercept = coef(m)[1],
                slope = coef(m)[2])
```



Residualene

Vi kan bruke modellformuleringen for å beregne predikerte verdier

$$\hat{y}_i = \beta_0 + \beta_1 \times \text{høyde}_i$$

Residualene er beregnet som

$$\text{Residual} = y_i - \hat{y}_i$$

```
# Predikerte verdier
ypred <- intercept + slope * dat$height.T1


# Residualer
res <- dat$VO2.max - ypred


# A plot of residuals
```

I JASP

- R: Korrelasjonskoeffisienten
- R^2: "Proportion of variance explained"
- Adjusted R^2: "Adjusted variance explained" ⟶ in the population
- RMSE (Root mean squared error):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \hat{\sigma}$$

```
r <- cor(dat$height.T1, dat$VO2.max)
r2 <- r^2

rmse <- sqrt(sum(res^2) / (n-2))
```

ANOVA

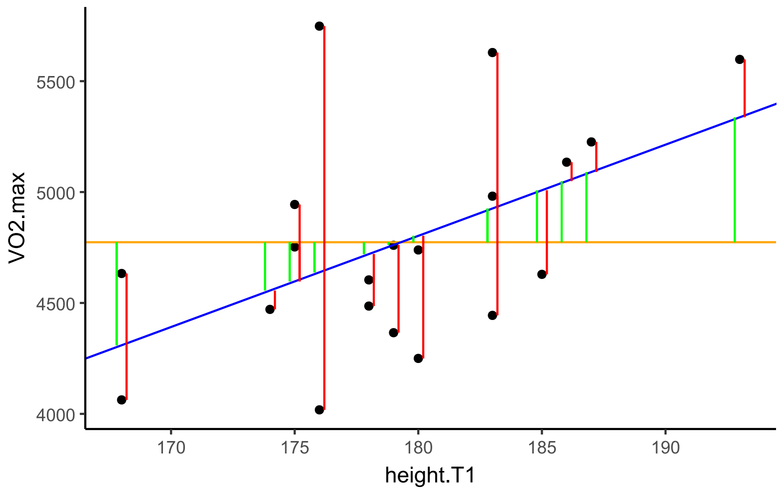
- ANOVA-tabellen gir et ratio av varians for regresionslinjen mot varians i dataene, hvor mye forklarer modellen?

```
# Sum of square regression model
ssr <- sum((ypred - mean(dat$VO2.max))^2)

# Resid sum of squares
sse <- sum( (dat$VO2.max - ypred )^2 )
```

```r
F.ratio <- (ssr/1) / (sse/(n-2))

# F is also
r2 * (n-2) / (1-r2)
```

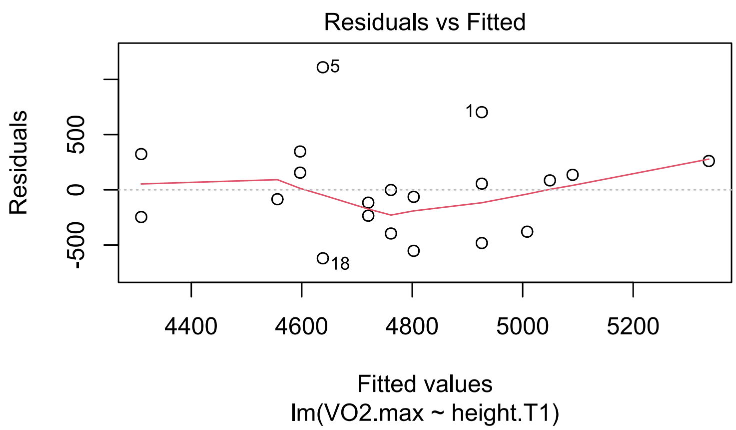
```
[1] 6.306151
```

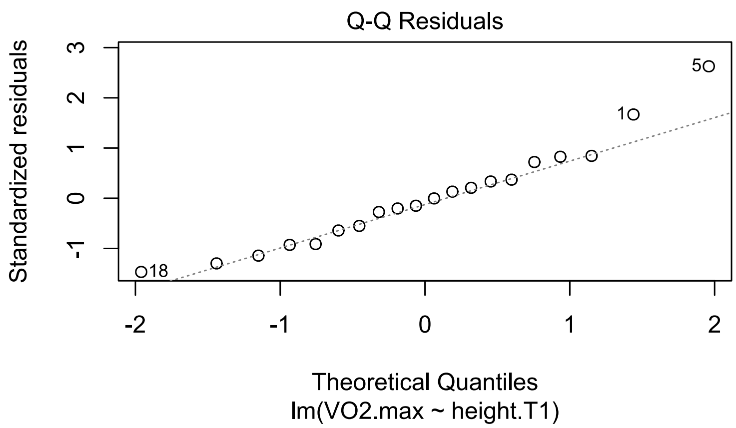
```r
dat |>
      ggplot(aes(height.T1, VO2.max)) +
      theme_classic() +
      # Data punkter
      geom_point() +
      # Gjennomsnitt y
      geom_hline(yintercept = mean(dat$VO2.max),
                 color = "orange") +
      # Modellen
      geom_abline(intercept = intercept,
                  slope = slope,
                  color = "blue") +
      # SSR
      geom_segment(aes(x = height.T1 - 0.2,
                       xend = height.T1 - 0.2,
                       y = mean(VO2.max),
                       yend = ypred),
                   color = "green") +

      # SSE
            geom_segment(aes(x = height.T1 + 0.2,
                       xend = height.T1 + 0.2,
                       y = VO2.max,
                       yend = VO2.max - res),
                       color = "red")
```

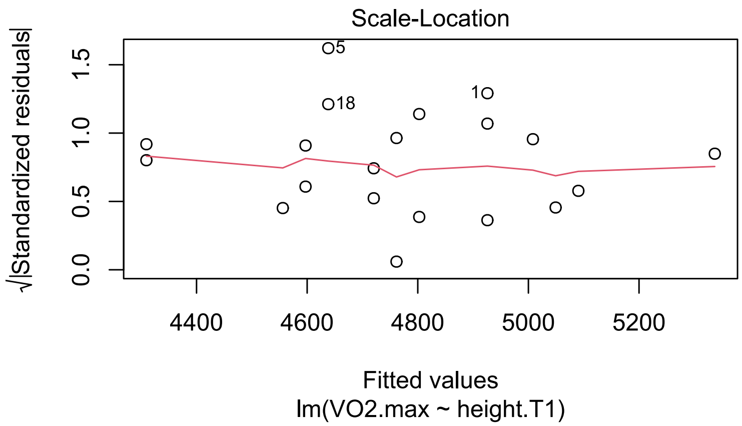


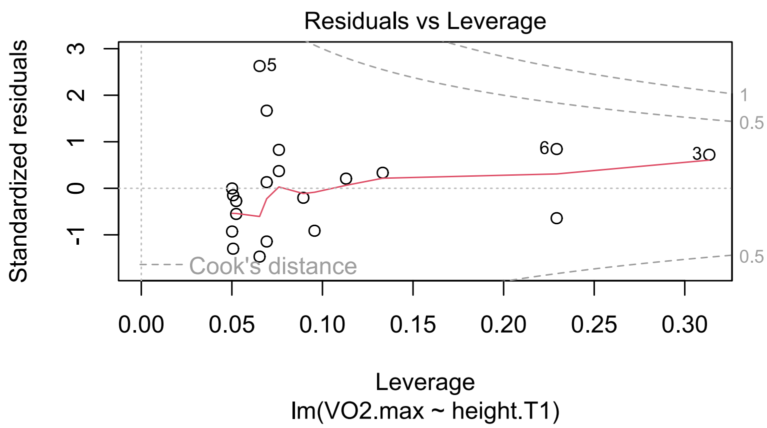
Diagnostikk

```
plot(m)
```




Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(VO2.max ~ height.T1)
5
1
18


Scale-Location
√|Standardized residuals|
Fitted values
lm(VO2.max ~ height.T1)
5
18
1


Residuals vs Leverage
Standardized residuals
Leverage
lm(VO2.max ~ height.T1)
Cook's distance
5
6
3
1
0.5
0.5