Exercises solutions

E. Pastucha

4/21/2020

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

AD1. Investigating populations

The data of interest (Europe and 1992) is extracted from the gapminder dataset. Then a linear model is fitted using the *lm* command, and the results are examined.

```
library(gapminder)
gapminder_european_data_from_1992 <- gapminder %>%
  filter(year == 1992, continent == 'Europe')
fit <- lm(lifeExp ~ gdpPercap, data = gapminder_european_data_from_1992)
summary(fit)
```

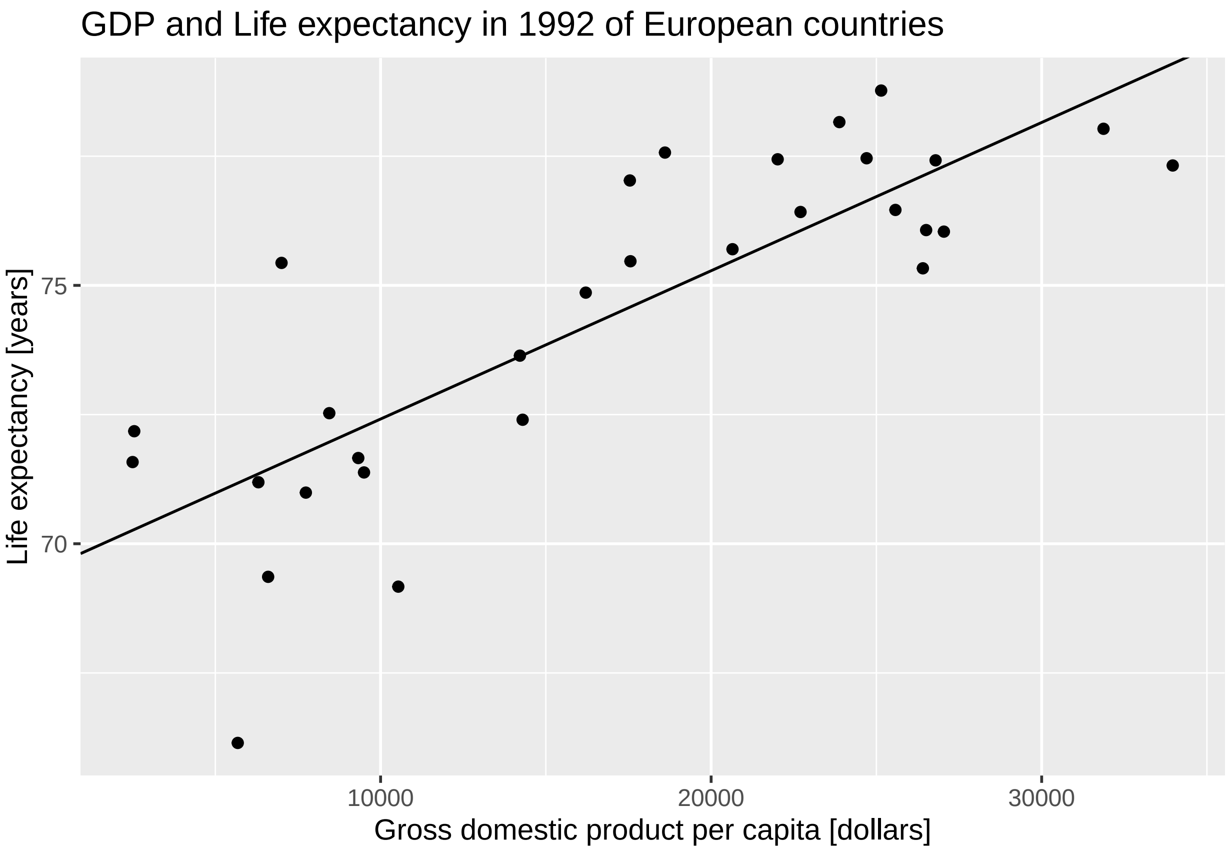
```
##
## Call:
## lm(formula = lifeExp ~ gdpPercap, data = gapminder_european_data_from_1992)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0280 -1.0323  0.1025  1.2109  3.8809
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.954e+01  7.447e-01  93.386  < 2e-16 ***
## gdpPercap   2.869e-04  3.865e-05   7.424 4.37e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.896 on 28 degrees of freedom
## Multiple R-squared:  0.6631, Adjusted R-squared:  0.6511
## F-statistic: 55.12 on 1 and 28 DF,  p-value: 4.373e-08
```

The fitted model has the form

$$lifeExp = 69.5 + 2.87 \times 10^{-4} \cdot gdp$$

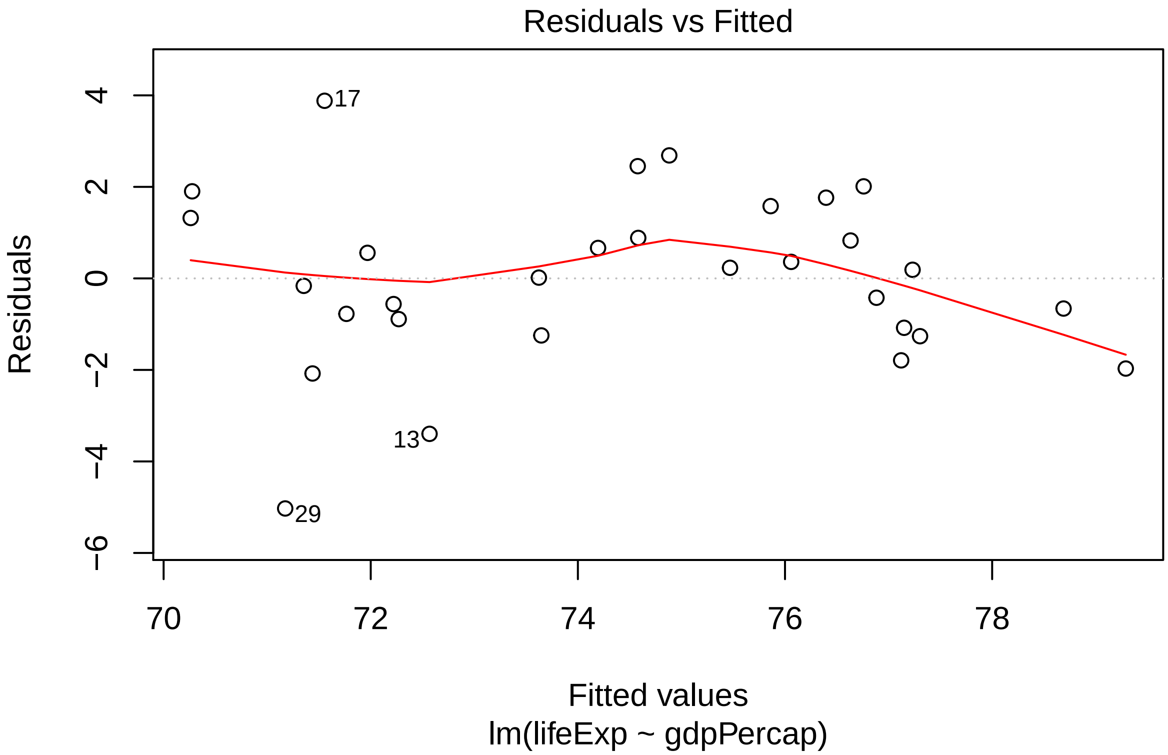
and is shown in the plot below.

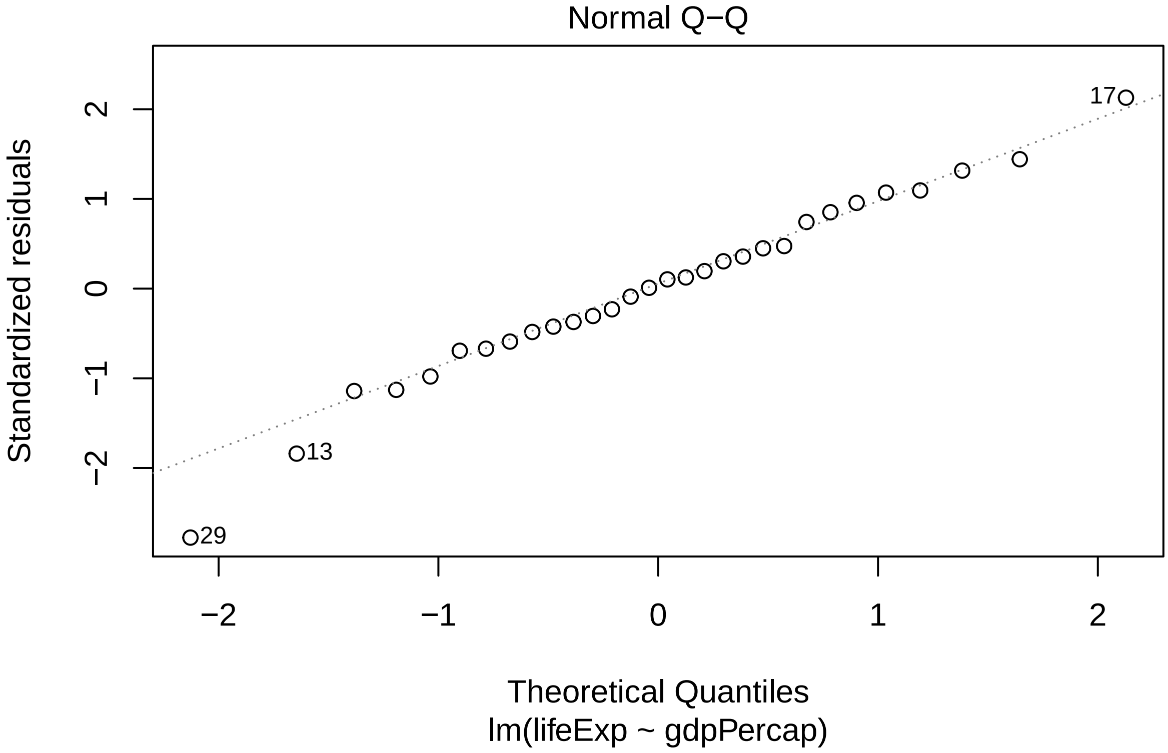
```
ggplot(gapminder_european_data_from_1992) +
  geom_point(aes(x=gdpPercap, y=lifeExp)) +
  labs(x = 'Gross domestic product per capita [dollars]',
       y = 'Life expectancy [years]',
       title = 'GDP and Life expectancy in 1992 of European countries') +
  geom_abline(intercept = fit$coefficients[[1]], slope=fit$coefficients[[2]])
```

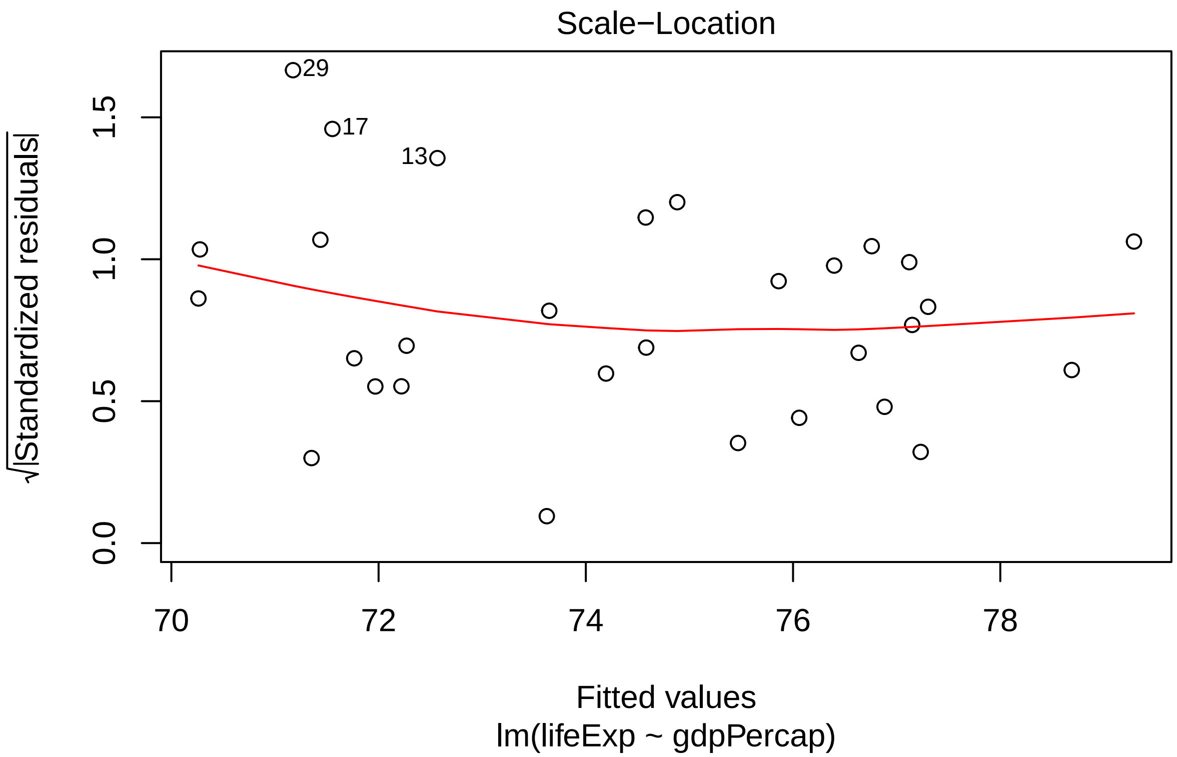


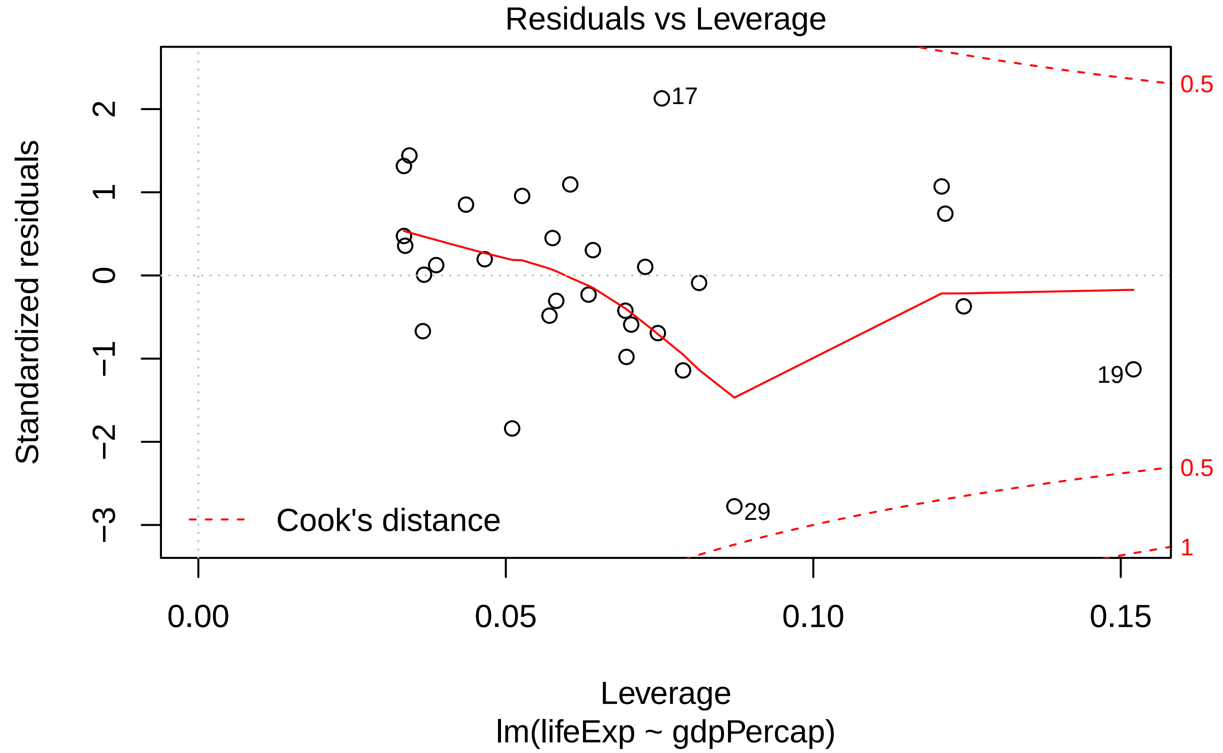
To check the assumptions for making a linear fit, the residuals are inspected. The following things are observed:

```
plot(fit)
```


Residuals vs Fitted
Residuals
4
2
0
−2
−4
−6
17
13
29
70
72
74
76
78
Fitted values
lm(lifeExp ~ gdpPercap)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
17
13
29
−2
−1
0
1
2
Theoretical Quantiles
lm(lifeExp ~ gdpPercap)





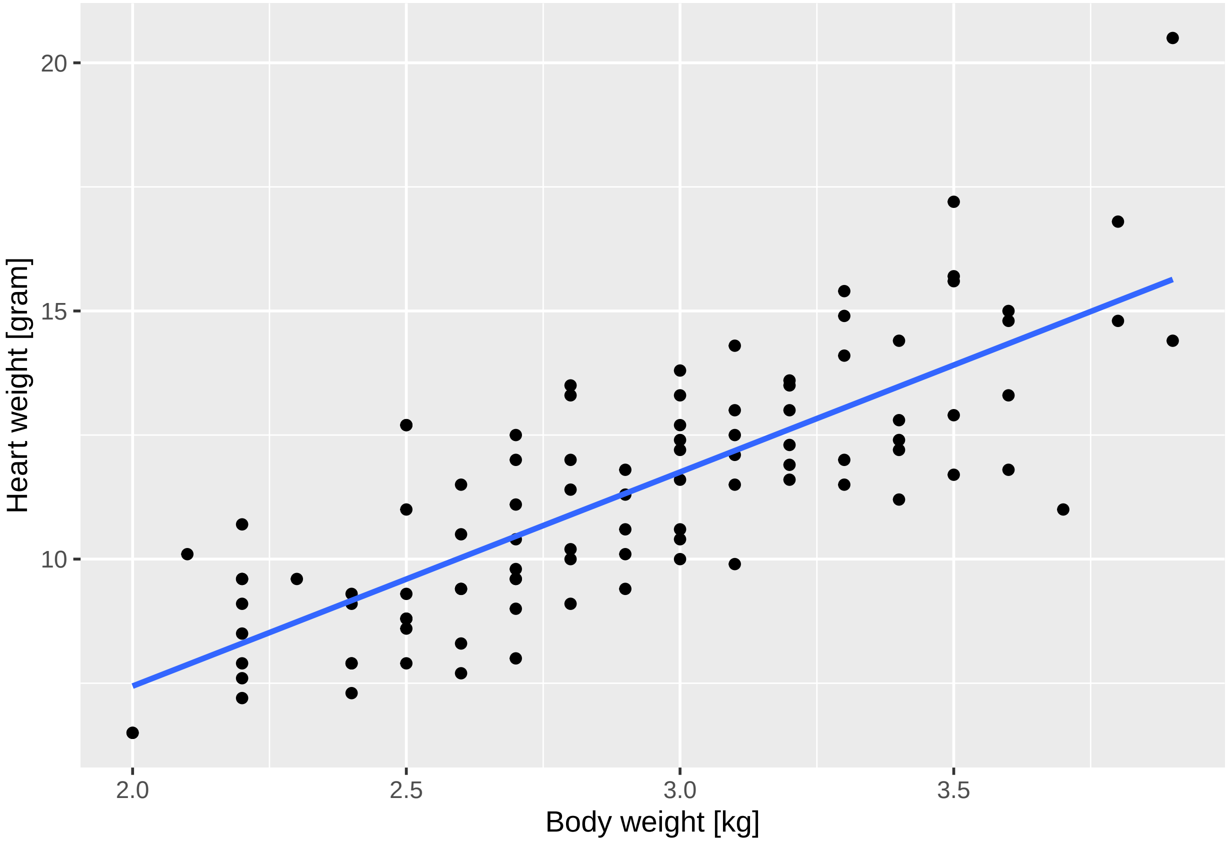
- There is a linera relationship visible in the data
- That the variance in the residuals decreases when the gross domestic product increases.
- The residuals seem to almost follow normal distribution.
- I assume the data is independent.

One of the underlying assumptions for the linear model is that the variance in residuals are uniform over the input data. This assumption is not met with the given data.

AD2. Cats body and heart weights

In this section the *catsM* dataset from the *boot* package is examined. The dataset contains measures of body and heart weight of a set of male cats.

```
library(boot)
catsM %>%
  ggplot() +
  geom_point(aes(Bwt, Hwt)) +
  stat_smooth(aes(Bwt, Hwt), method='lm', fullrange=TRUE, se=FALSE) +
  labs(x = 'Body weight [kg]',
       y = 'Heart weight [gram]')
```



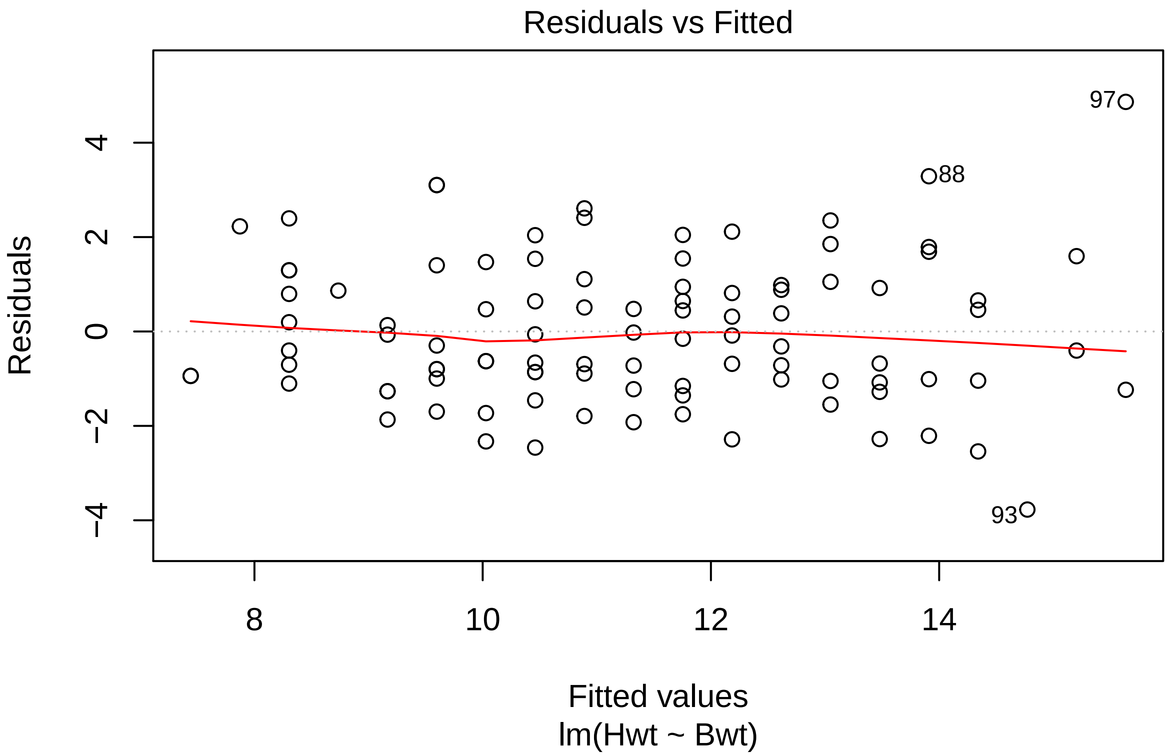
```
fit <- lm(Hwt ~ Bwt, data = catsM)
fit %>%
  summary()
```

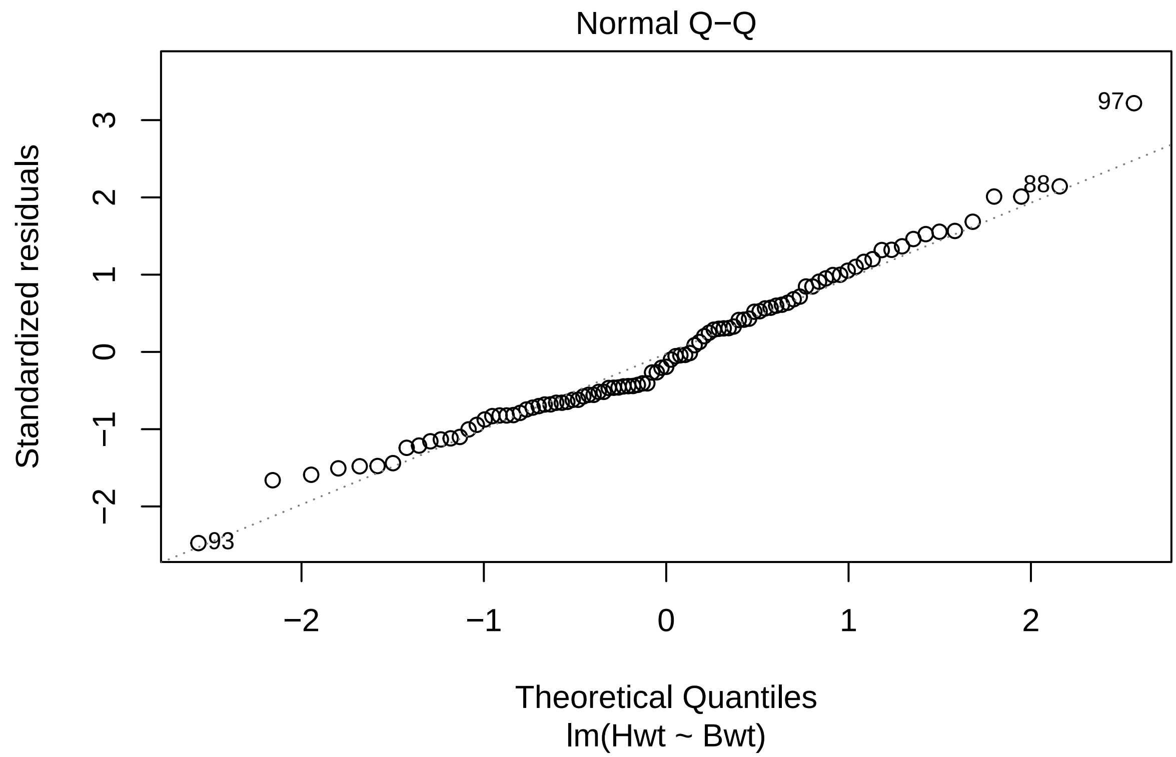
```
##
## Call:
## lm(formula = Hwt ~ Bwt, data = catsM)
##
## Residuals:
```

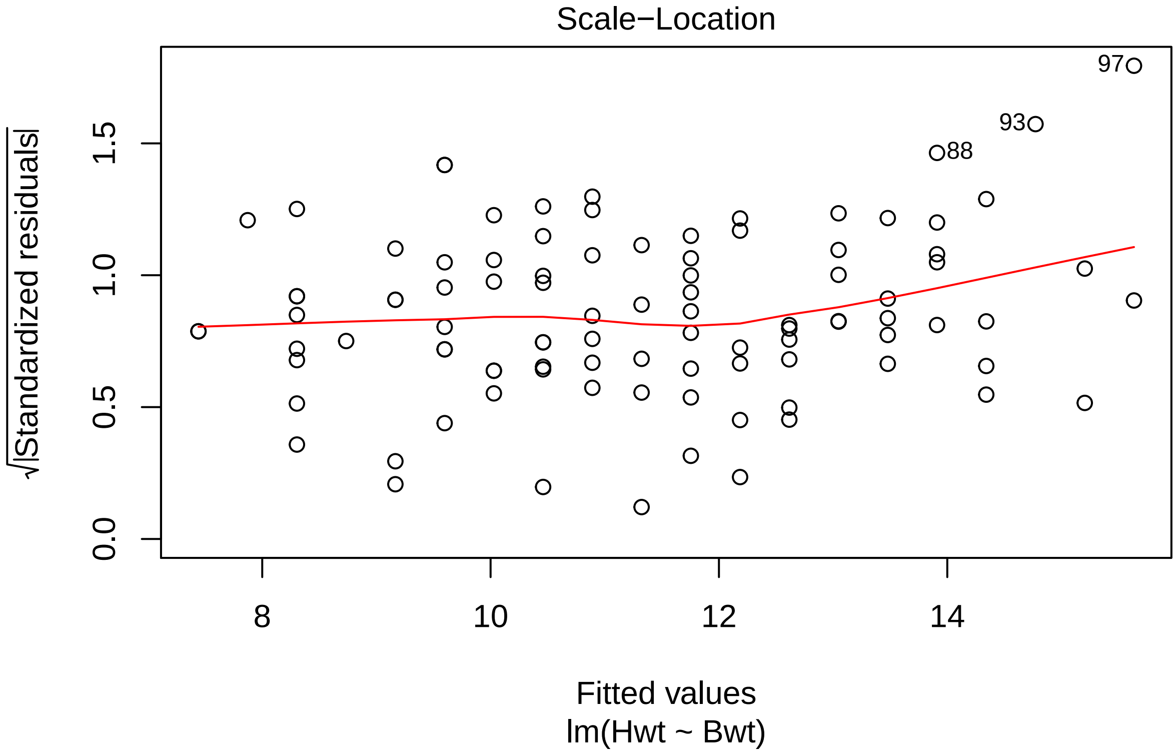
```
##     Min      1Q  Median      3Q     Max
## -3.7728 -1.0478 -0.2976  0.9835  4.8646
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.1841     0.9983  -1.186    0.239
## Bwt           4.3127     0.3399  12.688   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.557 on 95 degrees of freedom
## Multiple R-squared:  0.6289, Adjusted R-squared:  0.625
## F-statistic:   161 on 1 and 95 DF,  p-value: < 2.2e-16
```

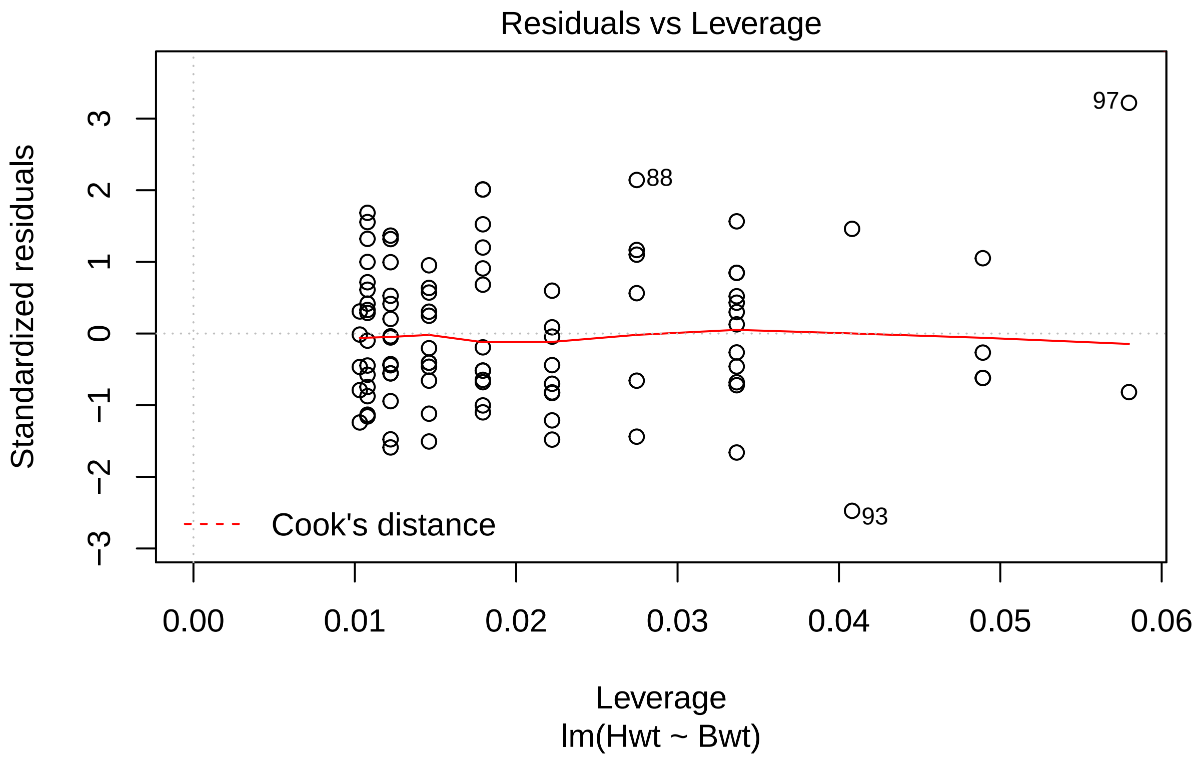
$$hwt = 4.31 \cdot bwt - 1.18$$

```
plot(fit)
```




Normal Q−Q
Standardized residuals
−2
−1
0
1
2
3
−2
−1
0
1
2
97
88
93
Theoretical Quantiles
lm(Hwt ~ Bwt)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
8
10
12
14
97
93
88
Fitted values
lm(Hwt ~ Bwt)



By investigating the residual plots, no violations of the following assumptions is detected

- There is a linear trend in the data
- The residuals are normal distributed (qq plot below)
- Constant variability of the residuals (plot from 3a)
- I assume data points are independent

AD3. Average temperatures

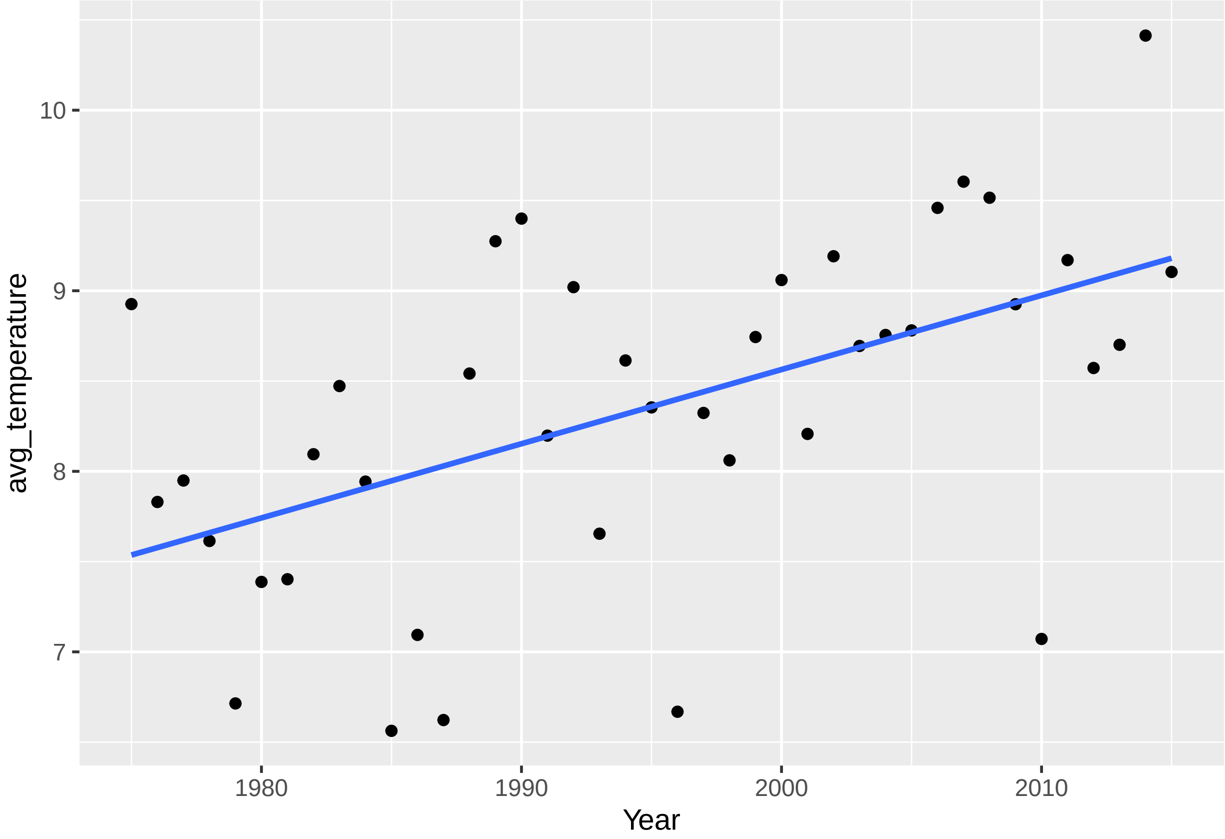
Data about temperatures in different countries are loaded from the file *temperatures.csv*.

```
temperatures <- readr::read_delim('temperatures.csv',
                                  delim = '\t')
```

```
## Parsed with column specification:
## cols(
##   tas = col_double(),
##   Year = col_double(),
##   Month = col_double(),
##   Country = col_character(),
##   ISO3 = col_logical(),
##   ISO2 = col_logical()
## )
```

Make a linear fit to the average temperature in Denmark from 1975 to 2015.

```
avg_DK <- temperatures %>%
  group_by(Country, Year) %>%
  summarise(avg_temperature = mean(tas)) %>%
  filter(Year >= 1975, Country == "DNK")
ggplot(avg_DK) +
  geom_point(aes(x = Year, y = avg_temperature)) +
  geom_smooth(aes(x = Year, y = avg_temperature), method = lm, se = FALSE)
```



```
fit <- lm(avg_temperature ~ Year, data = avg_DK)

fit %>%  summary()
```

```
##
## Call:
## lm(formula = avg_temperature ~ Year, data = avg_DK)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.90300 -0.39722  0.01128  0.49561  1.38983
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -73.62673   20.52183  -3.588 0.000919 ***
## Year          0.04110    0.01029   3.995 0.000278 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

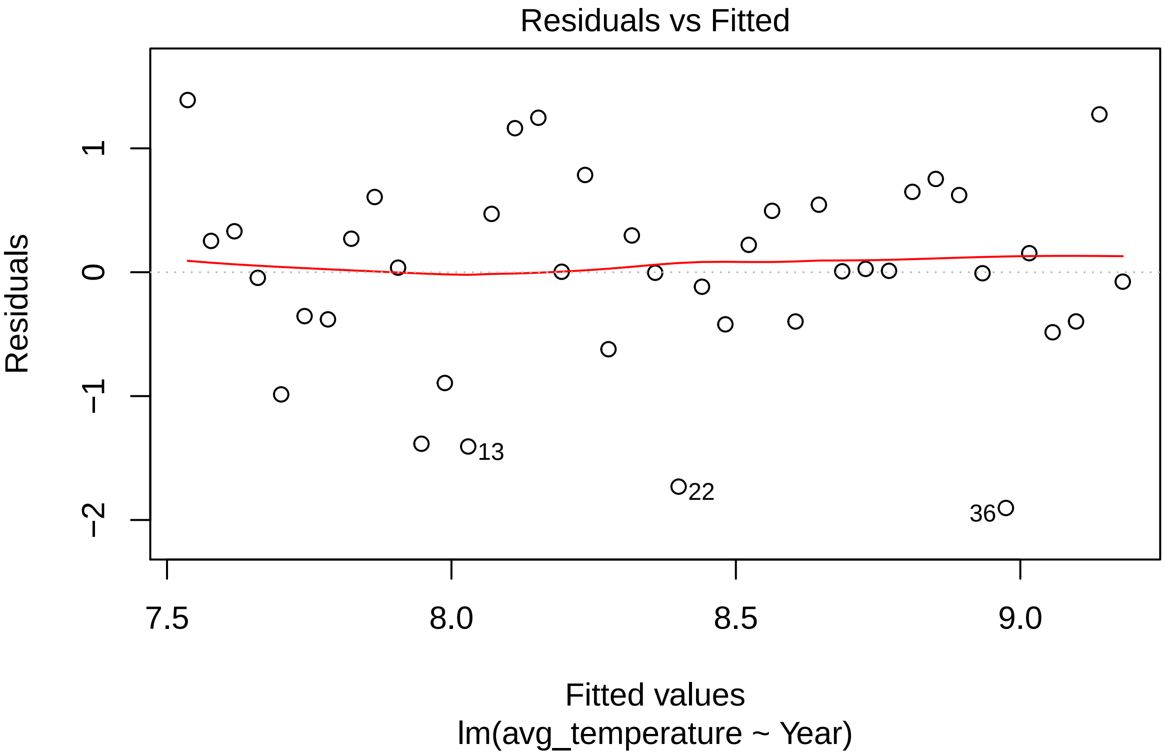
```
##
## Residual standard error: 0.7793 on 39 degrees of freedom
## Multiple R-squared:  0.2904, Adjusted R-squared:  0.2722
## F-statistic: 15.96 on 1 and 39 DF,  p-value: 0.0002779
```

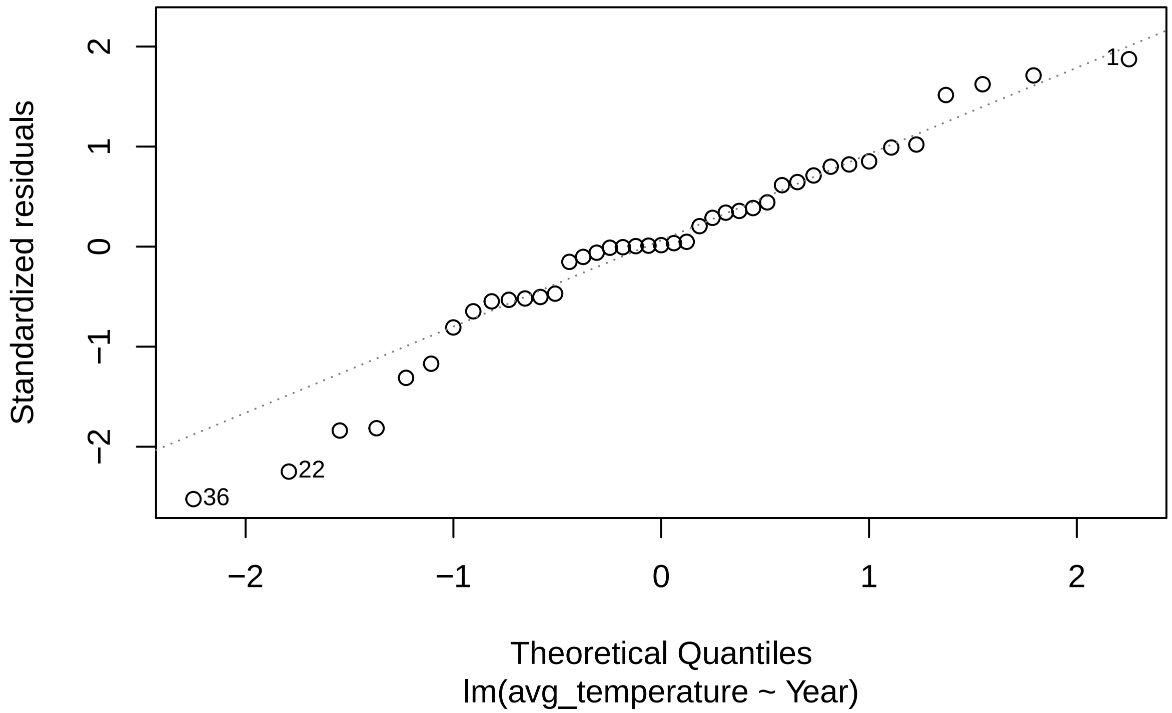
The fitted model has the form

$$\text{temperature} = -73.62673 + 0.04110 \cdot \text{year}$$

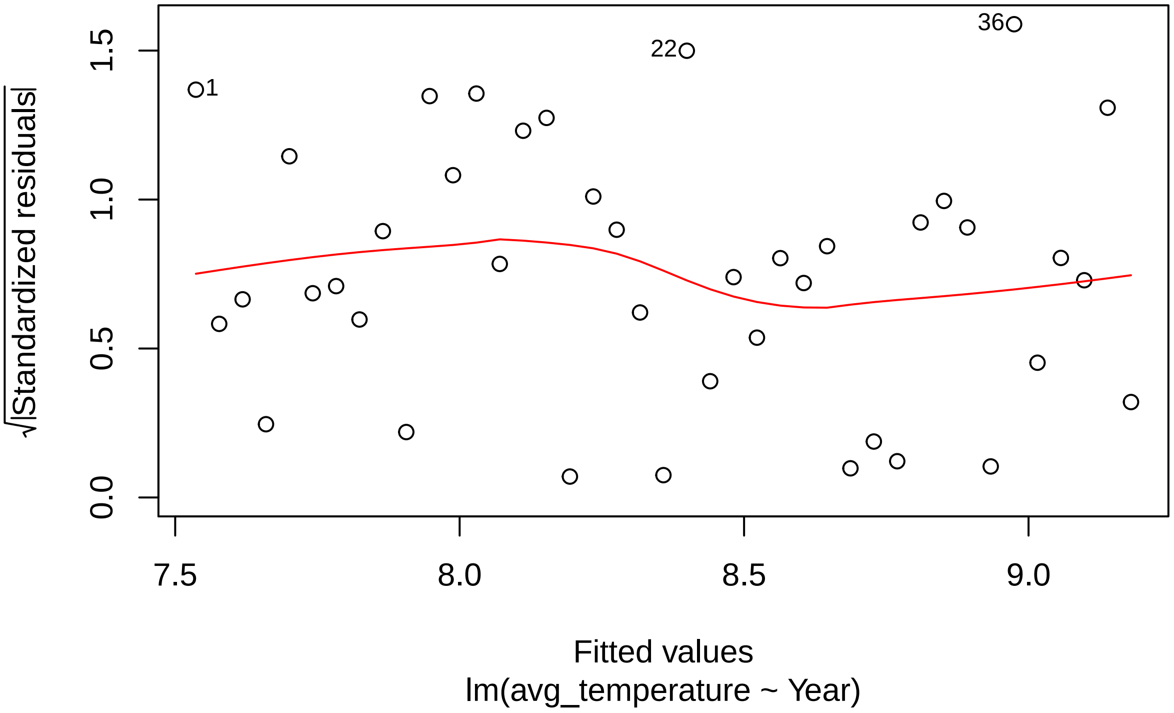
Where the slope of the line (0.04644) has a standard error of 0.01, with associated p value $6.21 \cdot 10^{-5}$.

```
plot(fit)
```

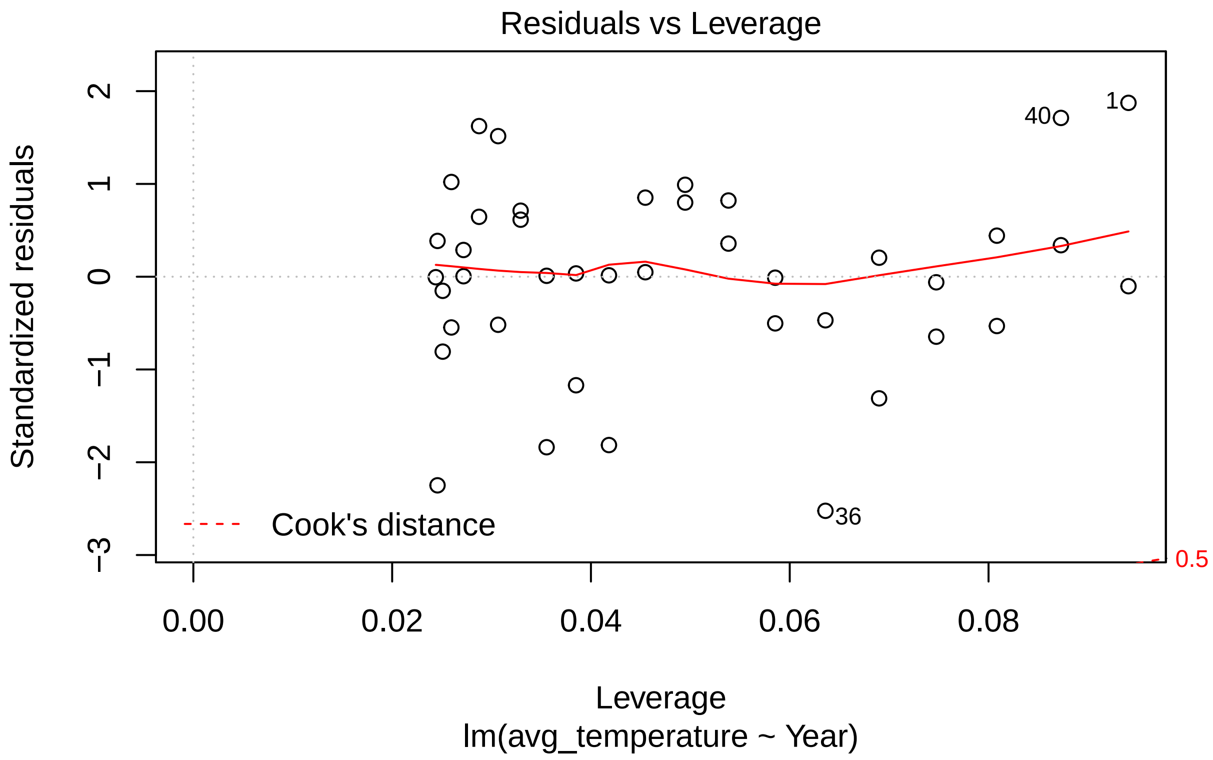




Normal Q−Q



Scale−Location



- The linear trend is directly observed when plotting the data.
- The variability is seen to be constant by looking at the residuals over time.
- The residuals are nearly normal distributed, as seen on the qq plot.
- There might be an issue with the requirement of independent observations, as the average temperature from one year potentially can influence the average temperature of the next year.

Predict the average temperature for the year 2030.

```
summary(fit)$coefficients[1] + summary(fit)$coefficients[2]* 2030
```

```
## [1] 9.796564
```

The average temperature for Denmark in 2030 year (according to the model) should reach 9.8 degrees.

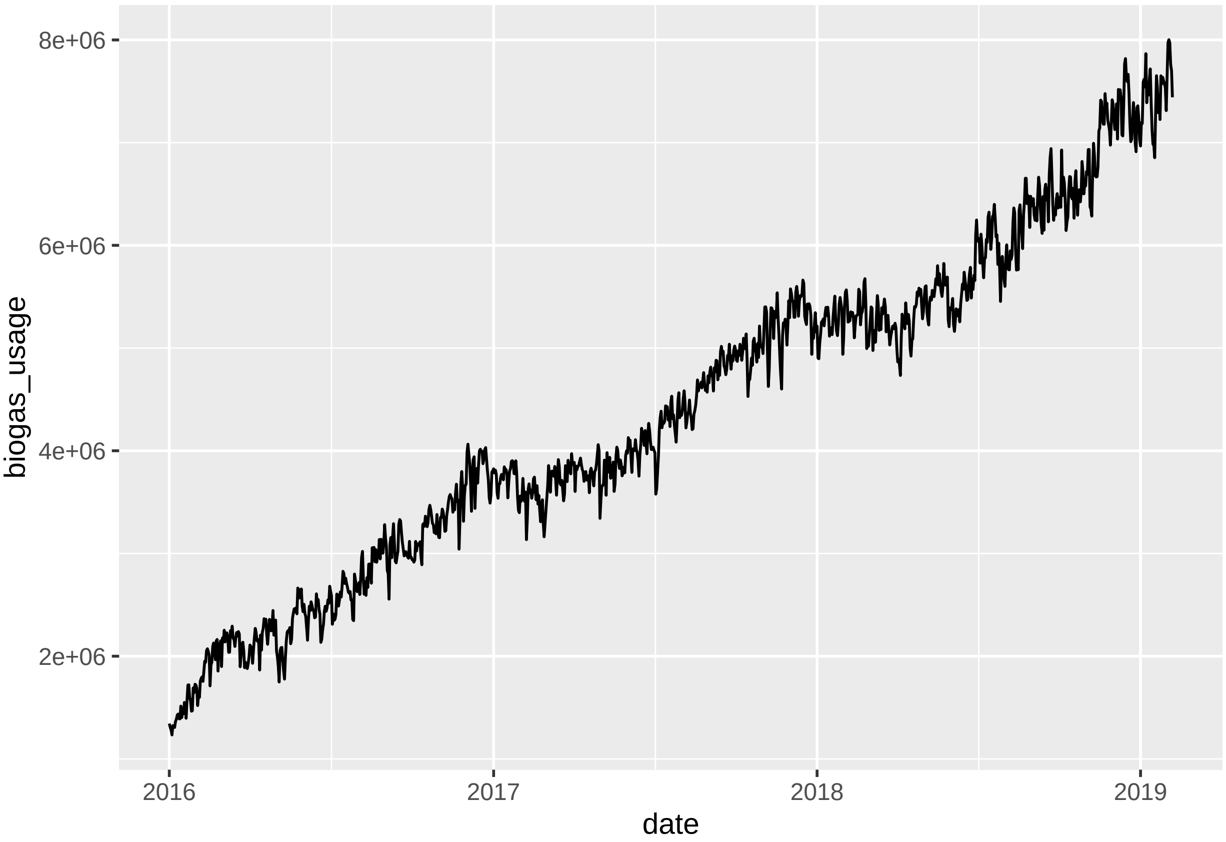
AD4. Biogas usage over time

```
biogas <- readr::read_delim('biogas_usage.csv', delim = ';')
```

```
## Parsed with column specification:
## cols(
##   date = col_datetime(format = ""),
##   year = col_double(),
##   daynumber = col_double(),
##   biogas_usage = col_double()
## )
```

To get a sense of the development of the useage of biogas over time, a scatter plot of the *date* and the *biogas_usage* is inserted below.

```
biogas %>%
  ggplot() +
  geom_line(aes(x = date, y = biogas_usage))
```



Fit a linear model to the biogas consumption. Use daynumber as the independent variable.

Details from the fitting process is shown below.

```
biogas %>%
  lm(biogas_usage ~ daynumber, data = .) -> biogas_fit
summary(biogas_fit)
```

```
##
## Call:
## lm(formula = biogas_usage ~ daynumber, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -917318 -193385   -1900  185313  941426
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1666647.8    17256.4   96.58   <2e-16 ***
## daynumber      4829.0       26.4  182.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 290600 on 1131 degrees of freedom
## Multiple R-squared:  0.9673, Adjusted R-squared:  0.9673
## F-statistic: 3.346e+04 on 1 and 1131 DF,  p-value: < 2.2e-16
```

Predict the biogas consumption for day number 3650

```
summary(biogas_fit)$coefficients[1] + summary(biogas_fit)$coefficients[2]* 3650
```

```
## [1] 19292534
```