MATH411 | Fall 2018 | Homework 3 (Due: Monday in class, 10/31/2018)

Paul Tomosky

10/31/18

Libraries:

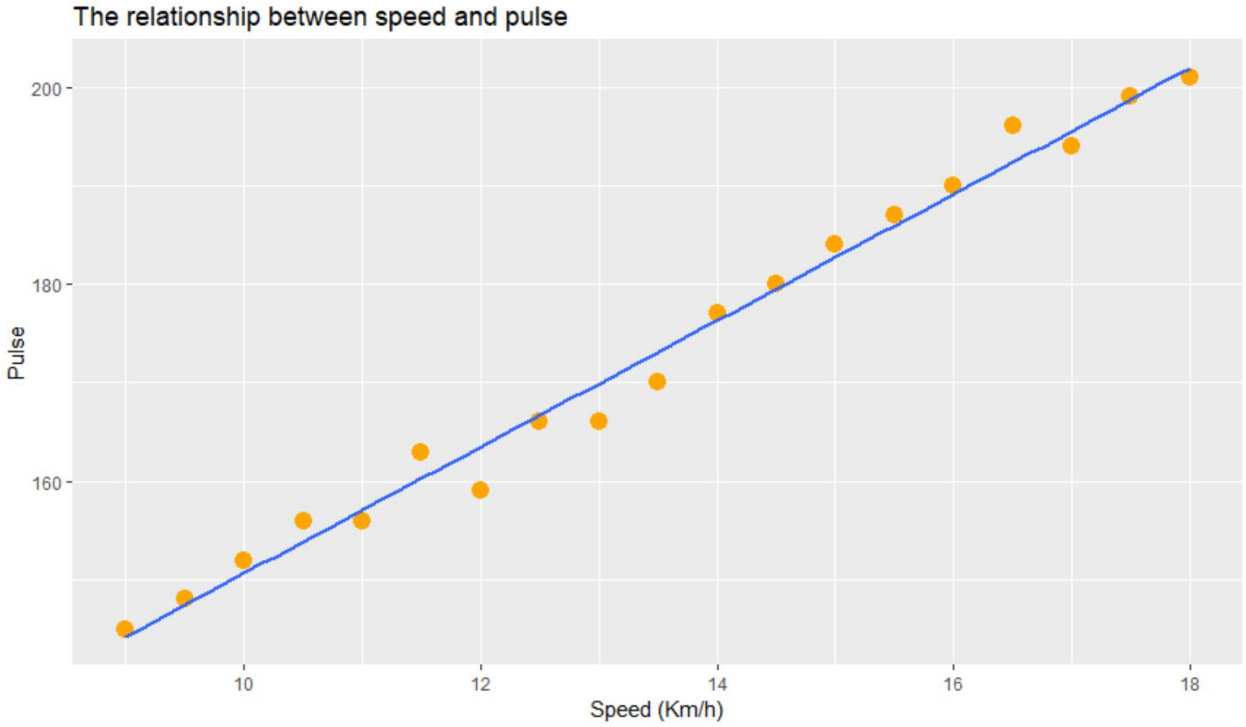
```
library(tidyverse)
library(car)
library(ggpmisc)
library(gridExtra)
library(GGally)
library(ggfortify)
```

Problem 1

The `Conconi` test measures the endurance performance of a person. It takes place on the 400m-track where one starts running slowly (9 km/h). Every 200 meters the speed is increased by 0.5 km/h. At the end of every 200m section the pulse is measured. The test continues until the speed can no longer be increased. A test was performed in summer 2012 on 19 participants. The data is contained in the file `conconi.rda`.

```
conconi = load(file.choose())
conconi %>% View()
```

(a) Visualize the data in a scatter plot with a regression line on it.




```
conconi %>%
 ggplot(aes(x = speed, y = puls))+
 geom_point(color = "orange", size = 4)+
 labs(title    = "The relationship between speed and pulse",
    y = "Pulse",
    x = "Speed (Km/h)")+
  geom_smooth(method = "lm", se = FALSE, size = 1)
```

(b) Fit a SLR for investigating the relationship between `pulse` (y) and `speed` (x) in R and answer the following questions:

slr_conconi = lm(puls ~ speed, data = conconi)

summary(slr_conconi)

```
> summary(slr_conconi)

Call:
lm(formula = puls ~ speed, data = conconi)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4947 -1.0123  0.5228  1.1825  3.6737

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.6105     2.5372   34.14   <2e-16 ***
speed         6.4070     0.1842   34.78   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.199 on 17 degrees of freedom
Multiple R-squared:  0.9861,   Adjusted R-squared:  0.9853
F-statistic:  1210 on 1 and 17 DF,  p-value: < 2.2e-16
```

confint(slr_conconi, level = 0.95)

```
> confint(slr_conconi, level = 0.95)
                2.5 %    97.5 %
(Intercept) 81.257578 91.963475
speed        6.018418  6.795617
```

(b.1) Write down the statistical model, in terms of the two variables, that you fitted.

Speed is the predictor variable and pulse in the target variable

(b.2) To what extent can we explain the scatter in the pulse by the increase in speed?

I think this can be explained by other variables like height, weight fitness, age, and others not measured in the data. The data is highly correlated with an R-squared value of 0.98.

(b.3) By what amount does the pulse increase on average when the speed is increased by 1 km/h? What other values are also plausible?

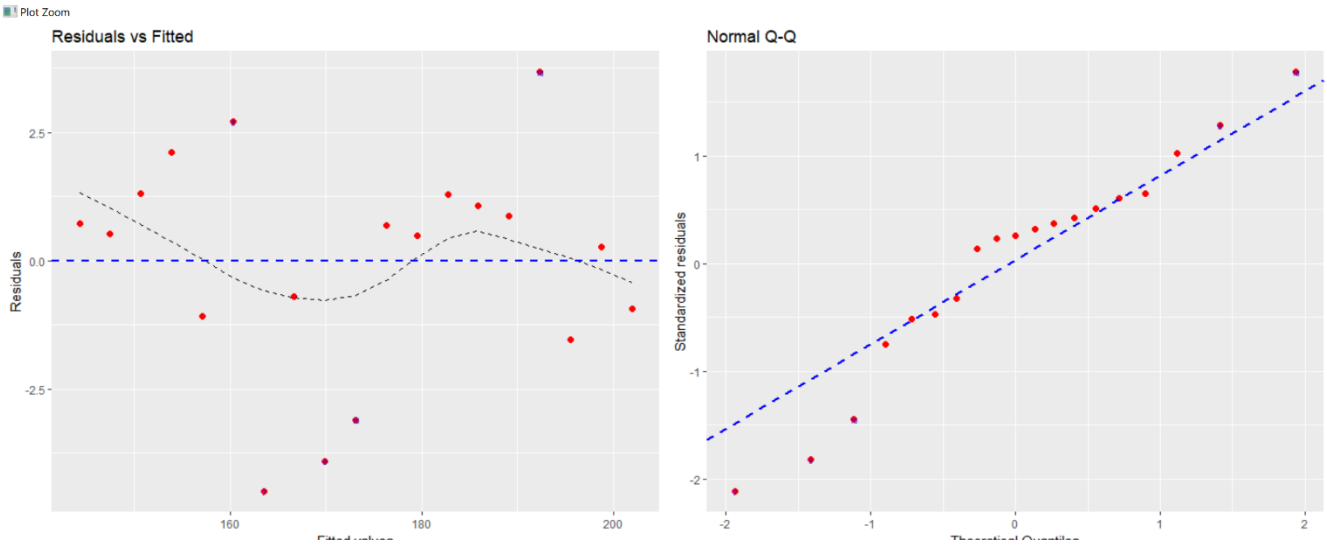
For every +1 km/h increase in speed, the pulse increases by 6.41. Based on this model, this seems like a plausible answer. Given the data, we can be 95% confident that the speed increase is between 6.02 and 6.79

(b.4) How large is the resting heart rate (i.e. when there is no movement)? In what interval do you expect this value to be? Does it seem plausible?

The resting heart rate is 86.6105. This is plausible because this is a normal rate for someone who is at rest. Given the data, we can be 95% confident that the resting heart rate (the intercept) is between 81.2 and 91.6

(c) Plot the residuals against the predictor as well as the normal plot of the residuals. Decide which of the following four assumptions are fulfilled.

```
autoplot(slr_conconi, which = 1:6, colour = 'red', size = 2,
     smooth.colour = 'black', smooth.linetype = 'dashed',
     ad.colour = 'blue', ad.size = 1,
     label.size = 2, label.n = 5, label.colour = 'blue',
     ncol = 3)
```



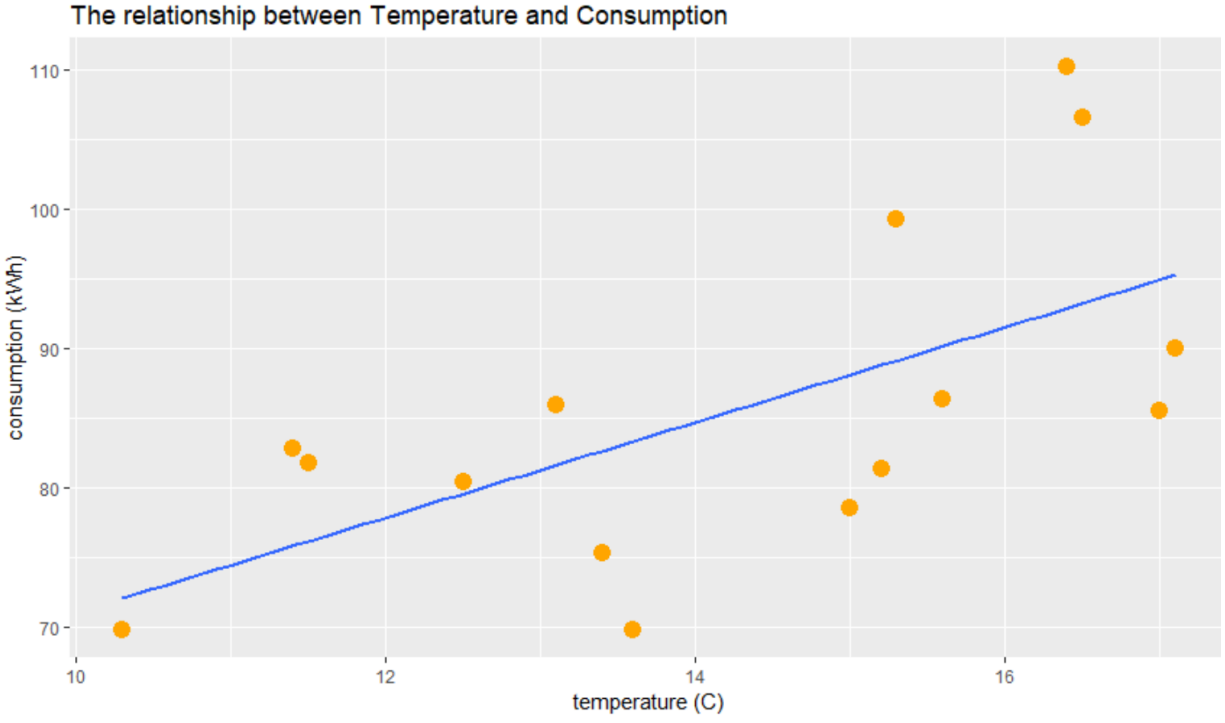
- The regression line captures the relation correctly, i.e. $E(\epsilon_i) = 0$.
  - **Yes**
- The variance of the error is constant, i.e. $Var(\epsilon_i) = \sigma_\epsilon^2$.
  - **Yes**
- The errors follow a Normal distribution, i.e. $\mathcal{N}(0, \sigma_\epsilon^2)$.
  - **Yes**
- The errors are uncorrelated, i.e. $Cov(\epsilon_i, \epsilon_i) = 0 \quad \forall i \neq j$
  - **No**

Problem 2

While fitting and visualizing simple linear regression models as well as conducting the corresponding tests becomes a routine after a while, assessing whether a model fits remains a challenging task. We will practice this with two additional data sets:

(a) The file `gas.rda` contains the gas `consumption` (in kWh) and the differences of `temperature` (in $^{\circ}C$) inside and outside of 15 houses which are heated with gas. The measurements were collected over a long time span and then averaged. The goal is to explain the gas consumption with the temperature difference. Plot the regression line and perform a residual analysis.

```
gas = load(file.choose())

gas %>% View()

gas %>%

 ggplot(aes(x = temp, y = verbrauch))+

 geom_point(color = "orange", size = 4)+

 labs(title    = "The relationship between Temperature and Consumption",

    y = "consumption (kWh)",

    x = "temperature (C)")+

 geom_smooth(method = "lm", se = FALSE, size = 1)
```



gas_slr = lm(verbrauch ~ temp, data = gas)

coef(gas_slr)

```
> coef(gas_slr)
(Intercept)        temp
   36.89366     3.41274
```

summary(gas_slr)

```
> summary(gas_slr)

Call:
lm(formula = verbrauch ~ temp, data = gas)

Residuals:
    Min      1Q  Median      3Q     Max
-13.497  -7.391  -2.235   6.280  17.367

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   36.894     16.961   2.175   0.0487 *
temp           3.413      1.177   2.900   0.0124 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.601 on 13 degrees of freedom
Multiple R-squared:  0.3929,   Adjusted R-squared:  0.3462
F-statistic: 8.413 on 1 and 13 DF,  p-value: 0.0124
```

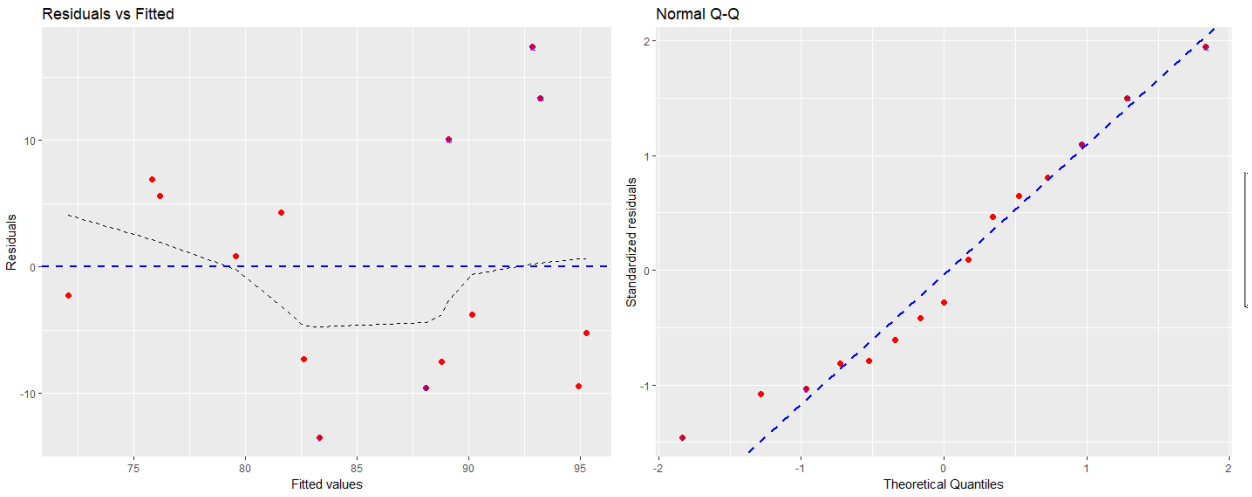
autoplot(gas_slr, which = 1:6, colour = 'red', size = 2,

smooth.colour = 'black', smooth.linetype = 'dashed',

ad.colour = 'blue', ad.size = 1,

label.size = 2, label.n = 5, label.colour = 'blue',

ncol = 3)



confint(gas_slr, level = 0.95)

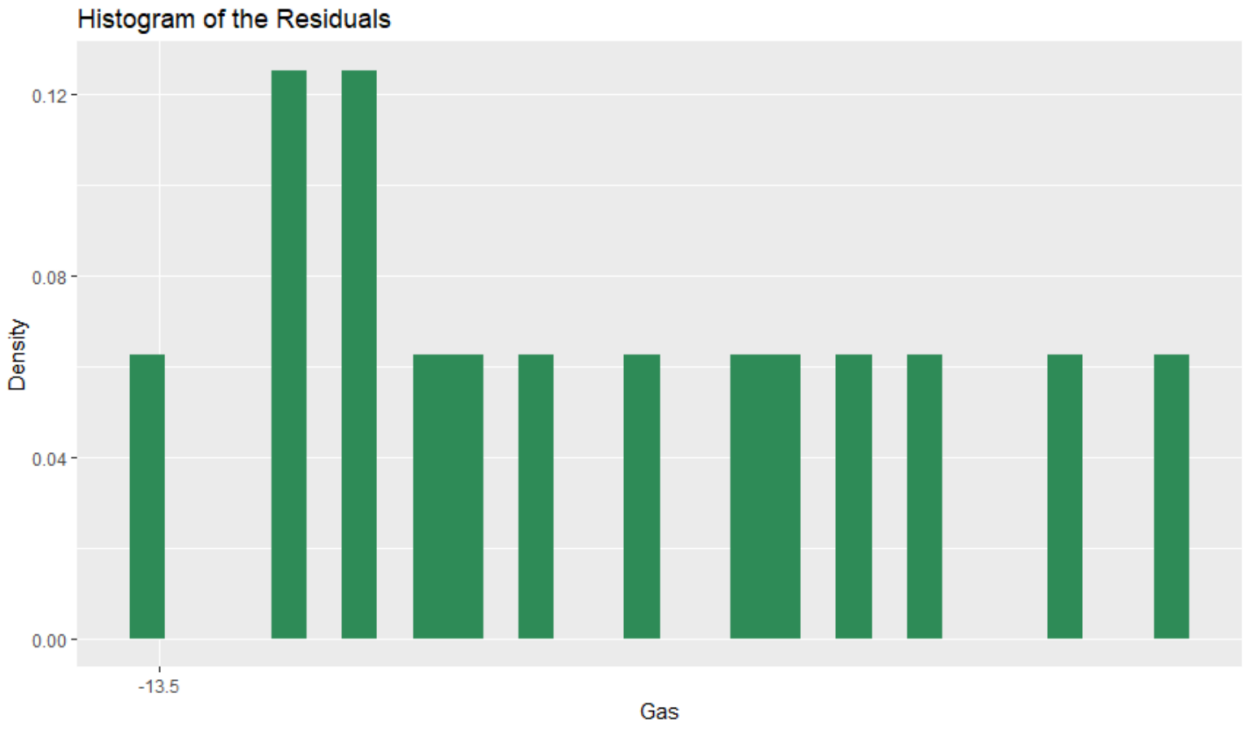
```
> confint(gas_slr, level = 0.95)
                 2.5 %    97.5 %
(Intercept) 0.2522626 73.535050
temp        0.8708088  5.954672
```

The Q-Q plot shows that the error is slightly skewed, but since the error is mostly normal, this model could be good at indicating the trend but will have a wide confidence interval

ggplot(gas_slr, aes(x = gas_slr$resid)) +

geom_histogram(aes(y = ..density..), fill = "seagreen")+

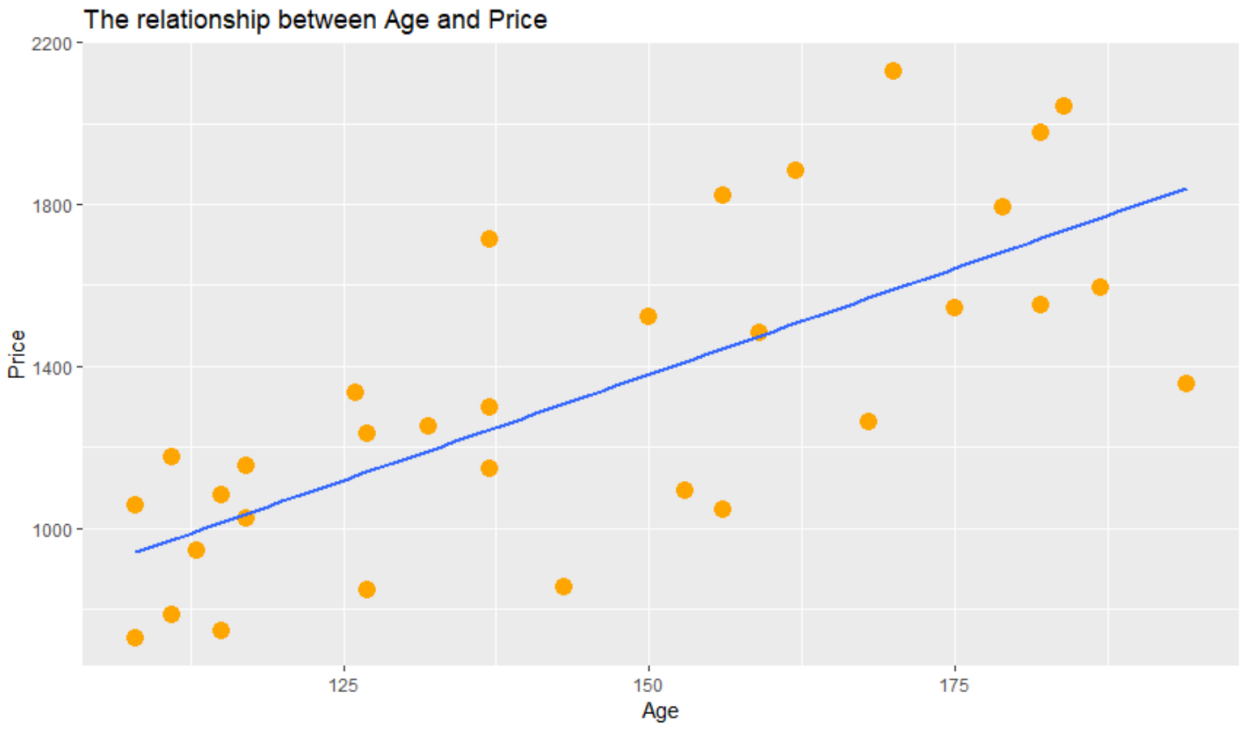
scale_x_continuous(breaks = round(seq(min(gas_slr$resid), max(gas_slr$resid), by = 100), 2))



The histogram of the residuals shows a clear skew to the left.

(b) The file `antikeUhren.rda` contains the `age` and the `price` of antique clocks that are auctioned.The goal is to predict the price with the age of the clock. Plot the regression line and perform a residual analysis.

```
antikeUhren = load(file.choose())

antikeUhren %>% View()

antikeUhren %>%
 ggplot(aes(x = Alter, y = Preis))+
 geom_point(color = "orange", size = 4)+
 labs(title    = "The relationship between Age and Price",
    y = "Price",
    x = "Age")+
 geom_smooth(method = "lm", se = FALSE, size = 1)
```



coef(antikeUhren_slr)

```
> coef(antikeUhren_slr)
(Intercept)       Preis
77.40365990  0.05088613
```

summary(antikeUhren_slr)

```
> summary(antikeUhren_slr)

Call:
lm(formula = Alter ~ Preis, data = antikeUhren)

Residuals:
    Min      1Q  Median      3Q     Max
-27.572 -13.466  -5.601  12.551  47.595

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 77.403660  12.016054   6.442 4.09e-07 ***
Preis        0.050886   0.008692   5.854 2.10e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.03 on 30 degrees of freedom
Multiple R-squared:  0.5332,   Adjusted R-squared:  0.5177
```

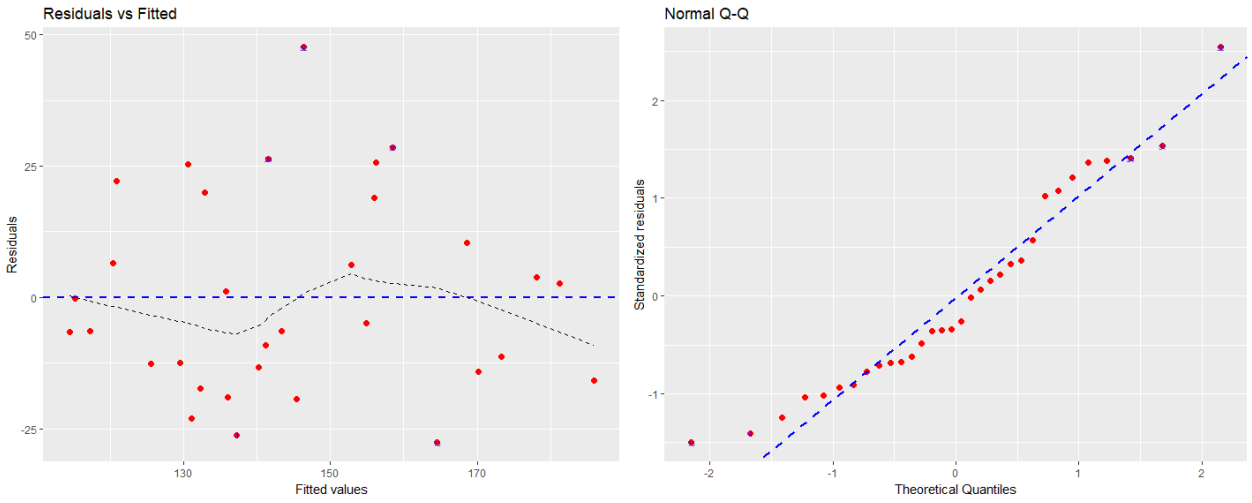
autoplot(antikeUhren_slr, which = 1:6, colour = 'red', size = 2,

smooth.colour = 'black', smooth.linetype = 'dashed',

ad.colour = 'blue', ad.size = 1,

label.size = 2, label.n = 5, label.colour = 'blue',

ncol = 3)

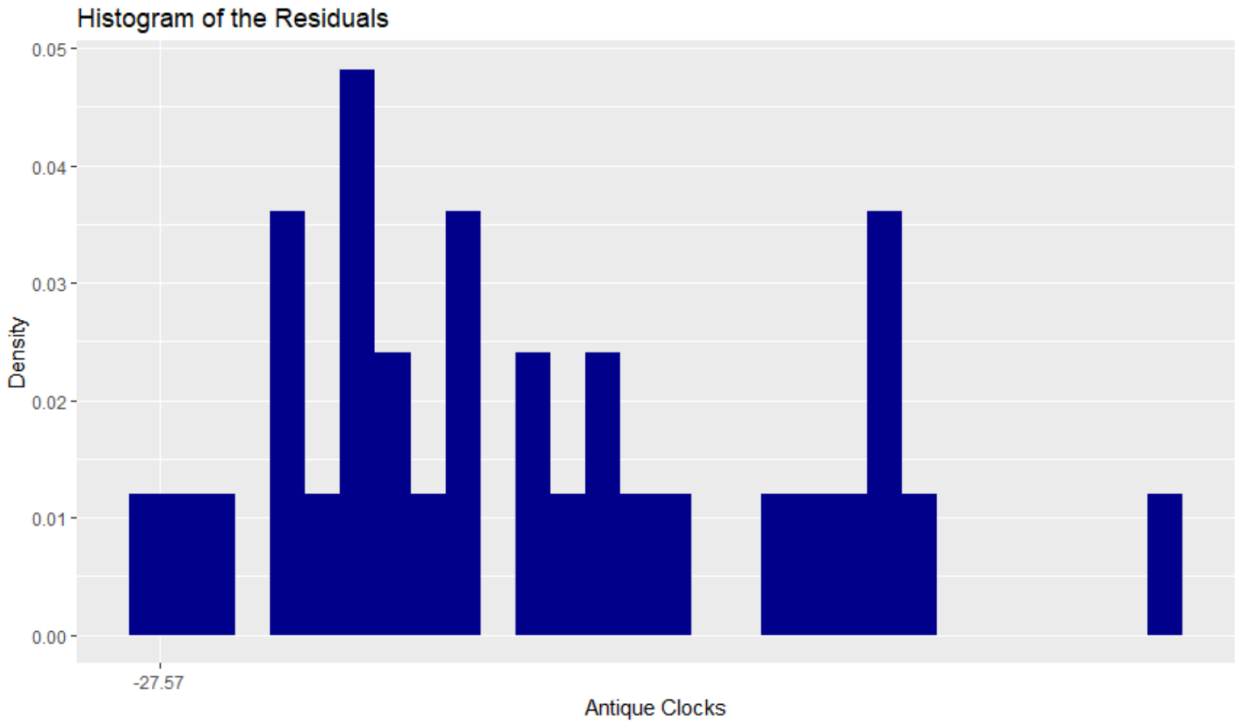


confint(gas_slr, level = 0.95)

```
> confint(gas_slr, level = 0.95)
                2.5 %    97.5 %
(Intercept) 0.2522626 73.535050
temp        0.8708088  5.954672
```

The Q-Q plot shows that the error is normal, since the error is mostly normal, this model could be good at indicating trends but have a wide confidence interval.

```
ggplot(antikeUhren_slr, aes(x = antikeUhren_slr$resid)) +
 geom_histogram(aes(y = ..density..), fill = "darkblue")+
 scale_x_continuous(breaks = round(seq(min(antikeUhren_slr$resid),
max(antikeUhren_slr$resid), by = 100), 2))+
 labs(title    = "Histogram of the Residuals",
    y = "Density",
    x = "Antique Clocks")
```



The histogram of the residuals shows a clear skew to the right.

Problem 3

In an experiment marine bacteria were exposed to x-rays during 15 intervals of six minutes. The following table contains the amount of bacteria after each interval

- `Interval`: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
- `Amount`: 255, 211, 197, 166, NA, 106, 104, 60, 56, 38, 36, 32, 21, 19, 15

interval = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
bacteria = c(255, 211, 197, 166, NA, 106, 104, 60, 56, 38, 36, 32, 21, 19, 15)
bacteria_experiment = data.frame(interval, bacteria)

(a) Show the relation between the number of surviving bacteria and the number of radiation intervals. Does it make sense to fit a OLS regression to the data?

bacteria_experiment_slr = lm(bacteria ~ interval , data = bacteria_experiment)

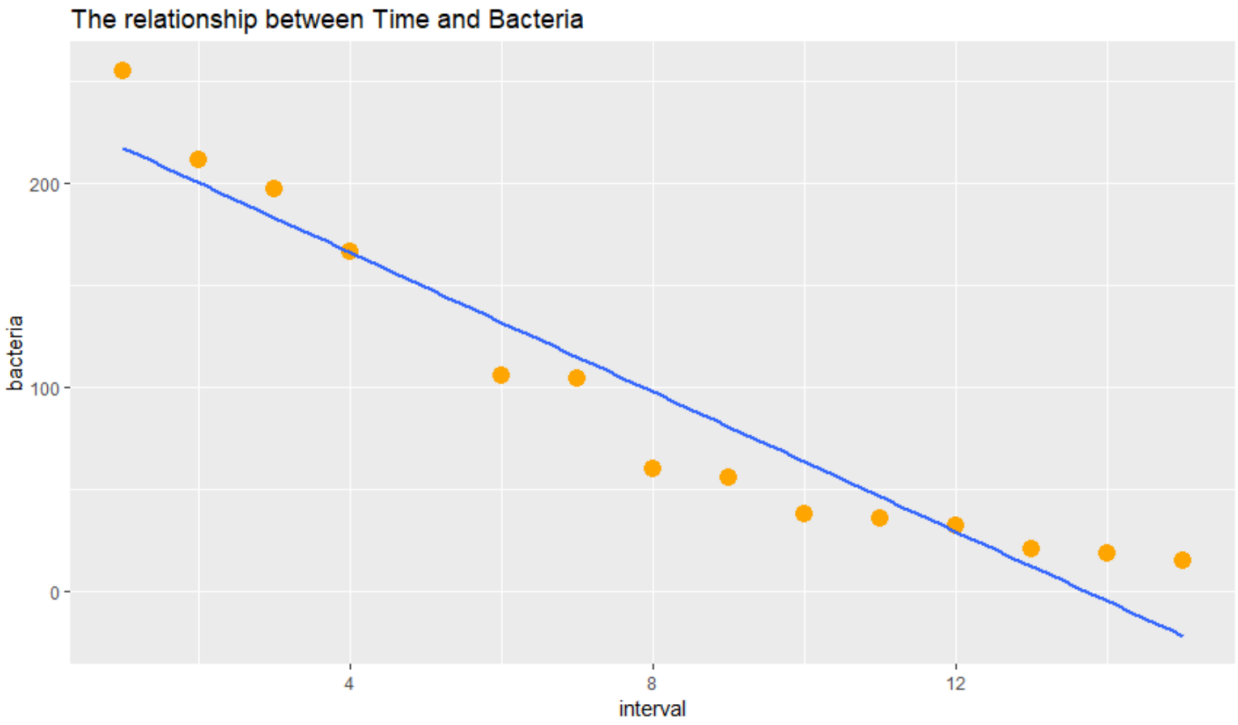
coef(bacteria_experiment_slr)

```
> coef(bacteria_experiment_slr)
(Intercept)    interval
  233.94452   -17.03672
```

The data shows a slight curve, it could potentially benefit from the OLS algorithm.

(b) Fit a simple linear regression model and check the model assumptions.

```
bacteria_experiment %>%
  ggplot(aes(x = interval, y = bacteria))+
  geom_point(color = "orange", size = 4)+
  labs(title    = "The relationship between Time and Bacteria",
       y = "bacteria",
       x = "interval")+
  geom_smooth(method = "lm", se = FALSE, size = 1)
```



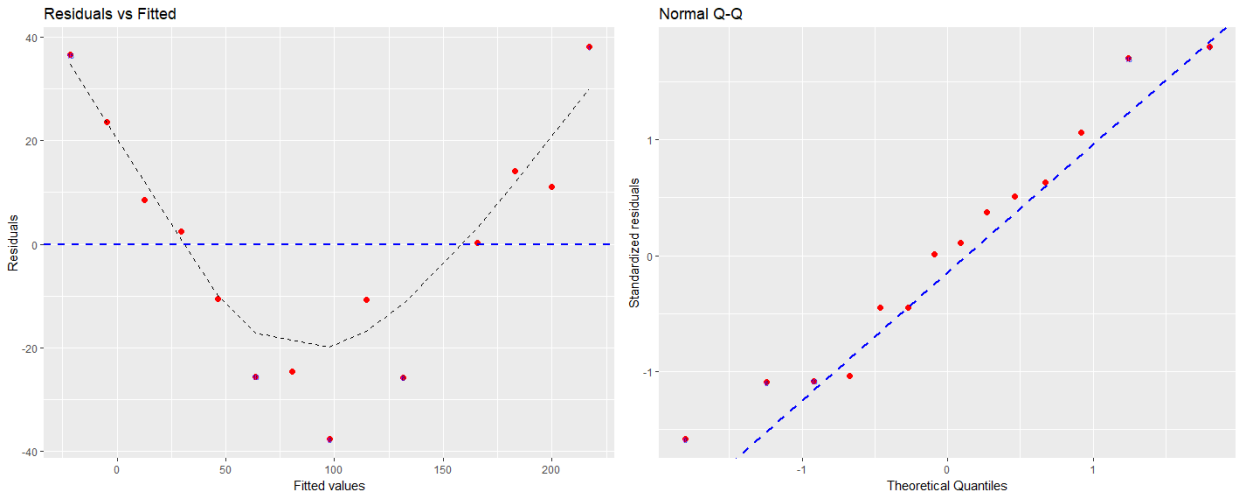
confint(bacteria_experiment_slr)

```
> confint(bacteria_experiment_slr, level = 0.95)
                2.5 %   97.5 %
(Intercept) 203.46647 264.4226
interval    -20.30835 -13.7651
```

summary(bacteria_experiment_slr, level = 0.95)

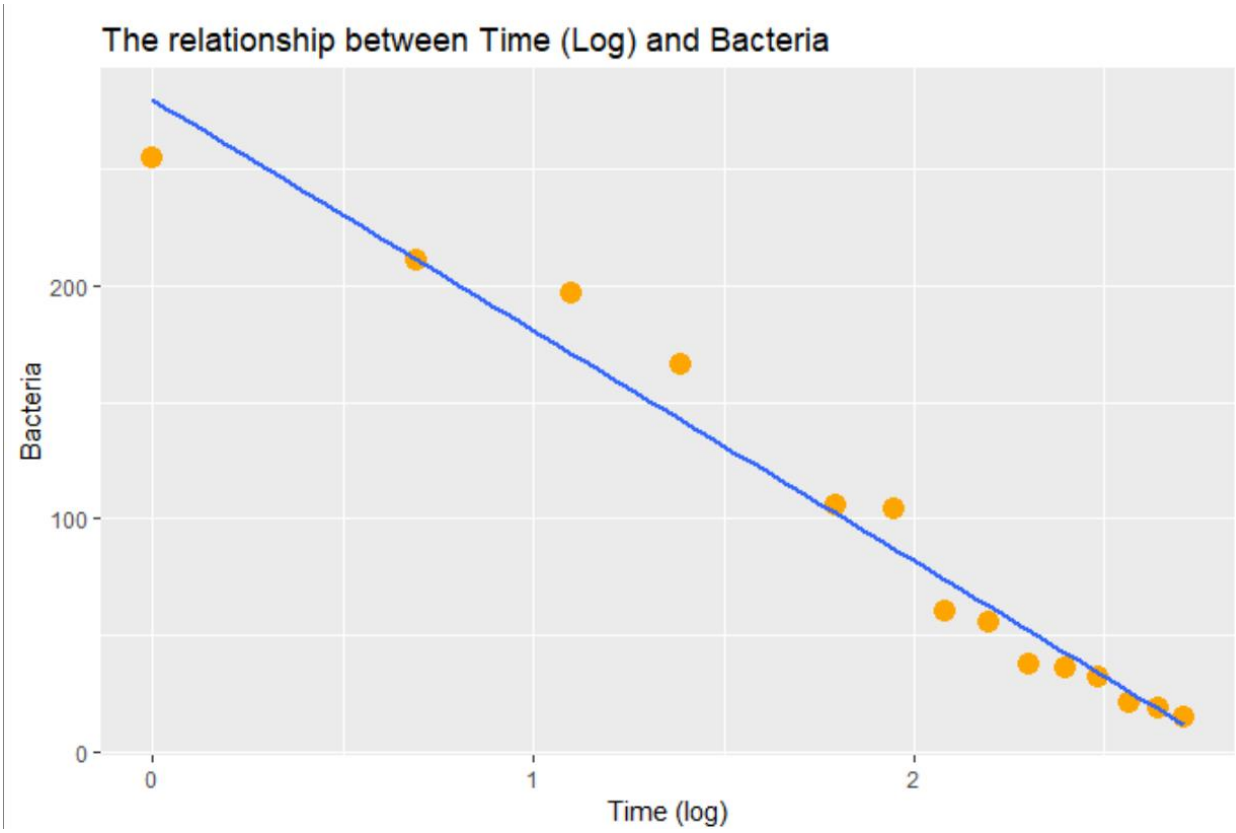
```
> summary(bacteria_experiment_slr)

Call:
lm(formula = bacteria ~ interval, data = bacteria_experiment)

Residuals:
    Min      1Q  Median      3Q     Max
-37.651 -21.132   1.349  13.406  38.092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  233.945     13.988   16.72 1.11e-09 ***
interval     -17.037      1.502  -11.35 9.01e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.69 on 12 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9147,   Adjusted R-squared:  0.9076
F-statistic: 128.7 on 1 and 12 DF,  p-value: 9.006e-08
```

```
autoplot(bacteria_experiment_slr, which = 1:6, colour = 'red', size = 2,
         smooth.colour = 'black', smooth.linetype = 'dashed',
         ad.colour = 'blue', ad.size = 1,
         label.size = 2, label.n = 5, label.colour = 'blue',
         ncol = 3)
```



(c) Improve the model by transforming the target variable or/and the predictor. **Hint**: The theory suggests that per radiation interval the proportion of bacteria that is killed remains constant.

```
bacteria_experiment_mutated = bacteria_experiment %>%
  mutate(log_interval = log(interval),
         log_bacteria = log(bacteria))
bacteria_experiment_mutated %>%
  ggplot(aes(x = log_interval, y = bacteria))+
  geom_point(color = "orange", size = 4)+
  labs(title    = "The relationship between Time (Log) and Bacteria",
     y = "Bacteria",
     x = "Time (log)")+
  geom_smooth(method = "lm", se = FALSE, size = 1)
```



(d) Predict the missing value for the fifth interval and compute a 95% prediction interval. In addition, provide the expected number of bacteria at the beginning, i.e. before the first radiation interval. Also compute a 95% confidence interval for this value.

predict(bacteria_experiment_mutated_slr, data.frame(log_interval = 1.6094379))

```
> predict(bacteria_experiment_mutated_slr, data.frame(log_interval =
1.6094379))
       1
120.5566
```

The predicted missing value for the 5th index is 120.56

coef(bacteria_experiment_mutated_slr)

```
> coef(bacteria_experiment_mutated_slr)
 (Intercept) log_interval
   279.79693    -98.94156
```

```
> confint(bacteria_experiment_mutated_slr, level = 0.95)
                 2.5 %    97.5 %
(Intercept)   257.2559 302.33795
log_interval -110.0293 -87.85385
```

The predicted interval before the first interval is 279.80. the 95% CI for this value is 257.25-302.34.