STAT 456 Homework 4

Hailey Lee

03/10/2024

Instructions:

1. Please use R to finish all the questions below. Although in some simple cases, you may obtain the solution directly without using R, you still need to provide the corresponding R code.
2. You are liable for missing points if you don't include output;
3. Whenever possible, please run saved variables, so our TA knows if your code goes the right way and assigns partial credits even if your final answer is wrong.
4. Please submit your solutions in a .pdf file compiled using R markdown.

1. **Time at the table.** Does how long young children remain at the lunch table help to predict how much they eat. Here are the data on 20 toddlers observed over several months at a nursery school. `Time` is the average number of minutes a child spent at the table when lunch was served. `Calories` is the average number of calories the child consumed during lunch, calculated from careful observation of what the child ate each day.

Child	Time	Calories
1	21.4	472
2	30.8	498
3	37.7	465
4	33.5	456
5	32.8	423
6	39.5	437
7	22.8	508
8	34.1	431
9	33.9	479
10	43.8	454
11	42.4	450
12	43.1	410
13	29.2	504
14	31.3	437
15	28.6	489
16	32.9	436
17	30.6	480
18	35.1	439
19	33.0	444
20	43.7	408

(a) Check the conditions for regression inference.

- **Linear relationship.** Draw a plot of residuals against the variable `Time`. Use vertical limits -100 to 100 in your plot of the residuals agains time to help you see the pattern. Does the plot show any systematic deviation from a roughly linear pattern?
- **Normal variation about the line.** Make a histogram of the residuals. Give some comments on the distribution of the residuals.

```
library(devtools)
```

```
## Loading required package: usethis
```

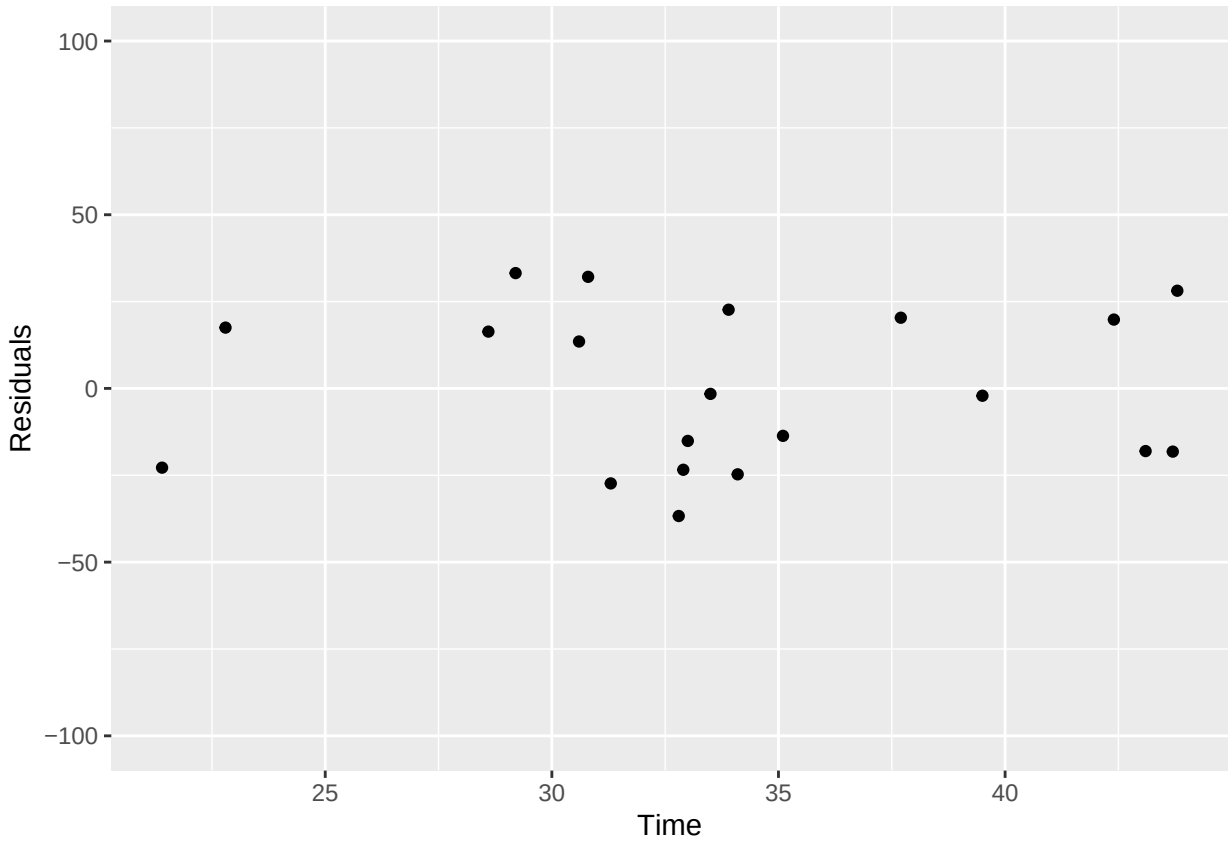
```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
```

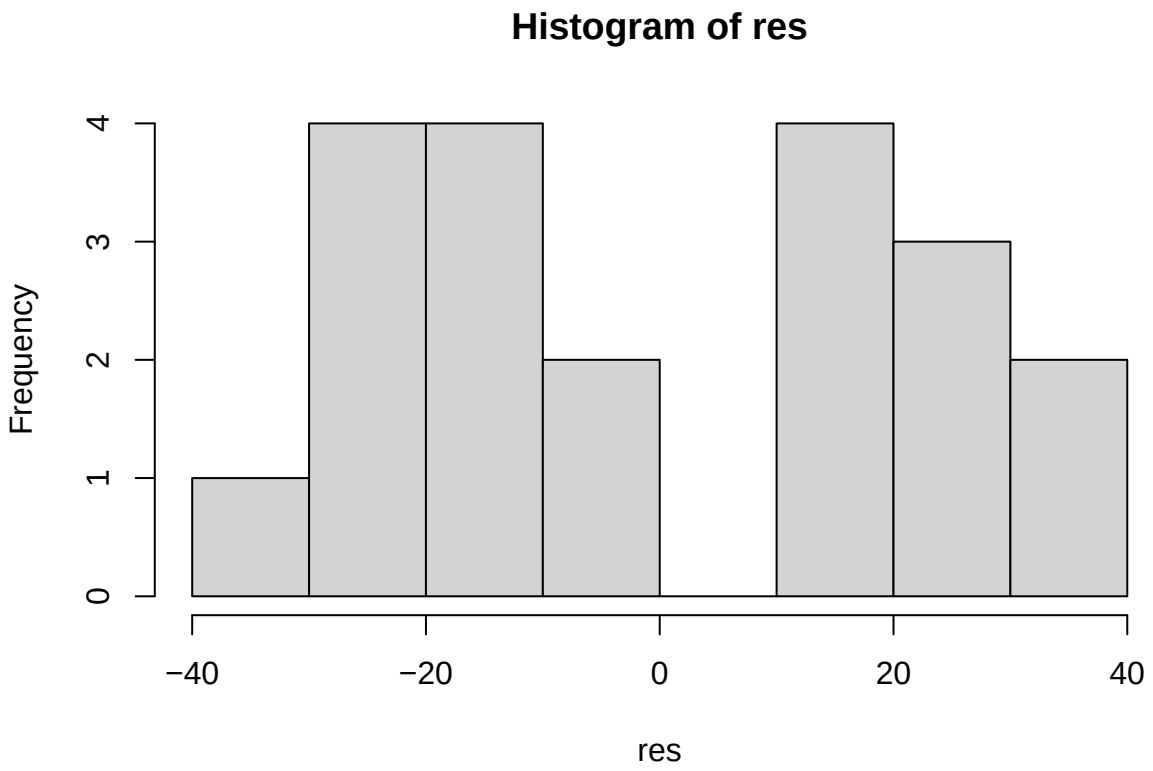
```
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
```

```r
library(dplyr)

child <- data.frame(
  Child <- 1:20,
  Time <- c(21.4, 30.8, 37.7, 33.5, 32.8, 39.5, 22.8, 34.1, 33.9, 43.8,
            42.4, 43.1, 29.2, 31.3, 28.6, 32.9, 30.6, 35.1, 33.0, 43.7),
  Calories <- c(472, 498, 465, 456, 423, 437, 508, 431, 479, 454,
                450, 410, 504, 437, 489, 436, 480, 439, 444, 408)
)

res <- resid(lm(Calories ~ Time, data=child))

child %>%
  ggplot(aes(Time, res)) +
  geom_point() +
  labs(x = "Time",
       y = "Residuals",
       main = "Time vs. Residuals") +
  ylim(-100, 100)
```




```
hist(res)
```



The Time vs. Residual plot shows that every value of residuals are within 50 units from the

expected value of Calories; however, it doens't really show linear pattern. Looking at histogram, it also does not look unimodal or roughly symmetric.

(b) Is there significant evidence that more time at the table is associated with more calories consumed? (Hint: This is to test if $H_0 : \beta_1 \leq 0$ *v.s.* $H_a : \beta_1 > 0$, and if the p-value < 0.05, then there is significant evidence that more time at the table is associated with more calories consumed; otherwise, there is not enough evidence.)

```
cor.test(child$Time, child$Calories, method = "pearson", alternative = "greater")
```

```
##
##  Pearson's product-moment correlation
##
## data:  child$Time and child$Calories
## t = -3.6208, df = 18, p-value = 0.999
## alternative hypothesis: true correlation is greater than 0
## 95 percent confidence interval:
##  -0.8251654  1.0000000
## sample estimates:
##        cor
## -0.6491667
```

Since p-value is greater than 0.05, there is not enough evidence that more time at the table is associated with more calories consumed.

2. **Ball Thrower.** [See Homework 3, Q3] A child throws a tennis ball straight up in the air. The table below shows the height of the ball (measured in feet from the ground) at $n = 7$ times, measured in seconds, where Time $= 0$ corresponds to the time at which the ball was released.

Time	Height
0.0	8
0.5	48
1.0	72
1.5	80
2.0	72
2.5	48
3.0	8

(a) Now define a new variable $Z = \text{Time}^2$, and consider the linear regression model:

$$Y_i = \gamma_0 + \gamma_1 Z_i + \epsilon_i, \tag{1}$$

where Y is Height and Z is Time^2, and obtain the numerical values for the least-squares estimates of $\widehat{\gamma}_0$ and $\widehat{\gamma}_1$.

```r
tennis <- data.frame(
  Time = c(0.0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0),
  Height = c(8, 48, 72, 80, 72, 48, 8)
)
tennis["Z"] = tennis$Time^2

coefficients(lm(Height ~ Z, data = tennis))
```

```
## (Intercept)           Z
##   56.000000   -2.461538
```

(b) Report the Pearson correlation between X (Time) and Z.

```r
cor.test(tennis$Time, tennis$Z, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  tennis$Time and tennis$Z
## t = 7.746, df = 5, p-value = 0.0005732
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7512548 0.9943790
## sample estimates:
##       cor
## 0.9607689
```

(c) What is the R^2 for the model in part (a)? Discuss your results.

```r
summary(lm(Height ~ Z, data = tennis))
```

```
##
## Call:
## lm(formula = Height ~ Z, data = tennis)
##
## Residuals:
##       1       2       3       4       5       6       7
## -48.000  -7.385  18.462  29.538  25.846   7.385 -25.846
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.000     17.187   3.258   0.0225 *
## Z             -2.462      3.813  -0.645   0.5471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.5 on 5 degrees of freedom
```

```
## Multiple R-squared:  0.07692,    Adjusted R-squared:  -0.1077
## F-statistic: 0.4167 on 1 and 5 DF,  p-value: 0.5471
```

The value of R^2 is 0.07692, which is relatively small.

(d) Compare the model in 2(a) Homework 4 and the model in 3(b) Homework 3. Which model is more reasonable? Is Time relevant for predicting Height? How are they related?

I would say the model from 3(b) of Homework 3 is more reasonable because R-squared is much smaller which can indicates more predictive model. Since the p-value is greater than 0.05 in 3(b), there is significant evidence that the more time throwing tennis balls is associated with higher height.

3. Consider the dataset `pressure` which is an R build in dataset. You may type `help(pressure)` to get more information on this dataset.

```
help(pressure)
data(pressure)
```

(a) The temperatures are provided on the Celsius scale. Convert them to the Fahrenheit scale and store them in an appropriate vector.

```
temp_F = (pressure$temperature * 9/5) + 32
temp_F
```

```
##  [1]  32  68 104 140 176 212 248 284 320 356 392 428 464 500 536 572 608 644 680
```

(b) Create a data frame consisting of the temperature in the Fahrenheit scale and Pressure.

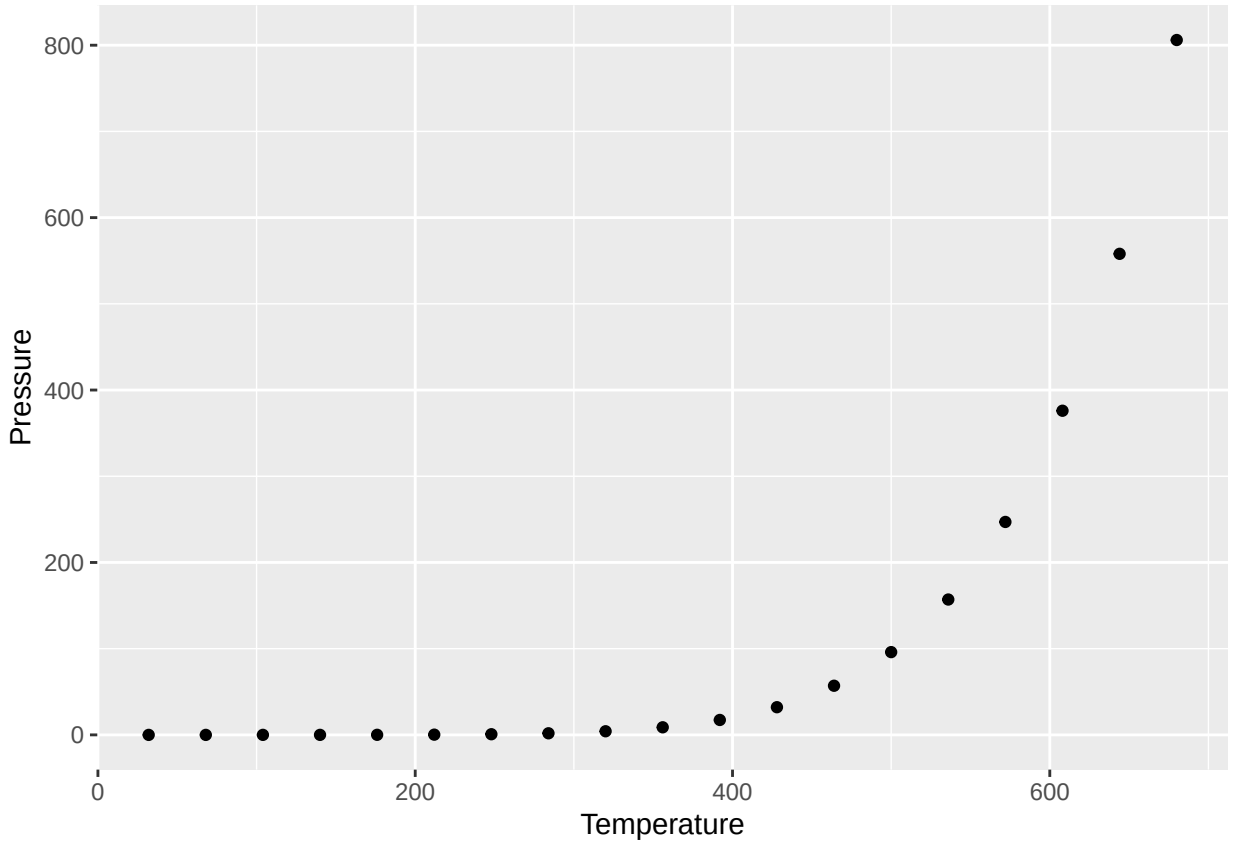
```
new_pressure <- data.frame(
  Temperature = temp_F,
  Pressure = pressure$pressure
)
new_pressure
```

```
##    Temperature Pressure
## 1           32   0.0002
## 2           68   0.0012
## 3          104   0.0060
## 4          140   0.0300
## 5          176   0.0900
## 6          212   0.2700
## 7          248   0.7500
## 8          284   1.8500
## 9          320   4.2000
## 10         356   8.8000
## 11         392  17.3000
## 12         428  32.1000
```

```
## 13          464  57.0000
## 14          500  96.0000
## 15          536 157.0000
## 16          572 247.0000
## 17          608 376.0000
## 18          644 558.0000
## 19          680 806.0000
```

(c) Plot temperature against pressure in the Fahrenheit scale.

```
new_pressure %>%
  ggplot(aes(x = Temperature, y = Pressure)) +
  geom_point()
```



(d) Perform a simple linear regression with temperature (in Fahrenheit) against pressure, but **with no intercept** in the model. Report a summary of the results and plot the fitted line in part (c).

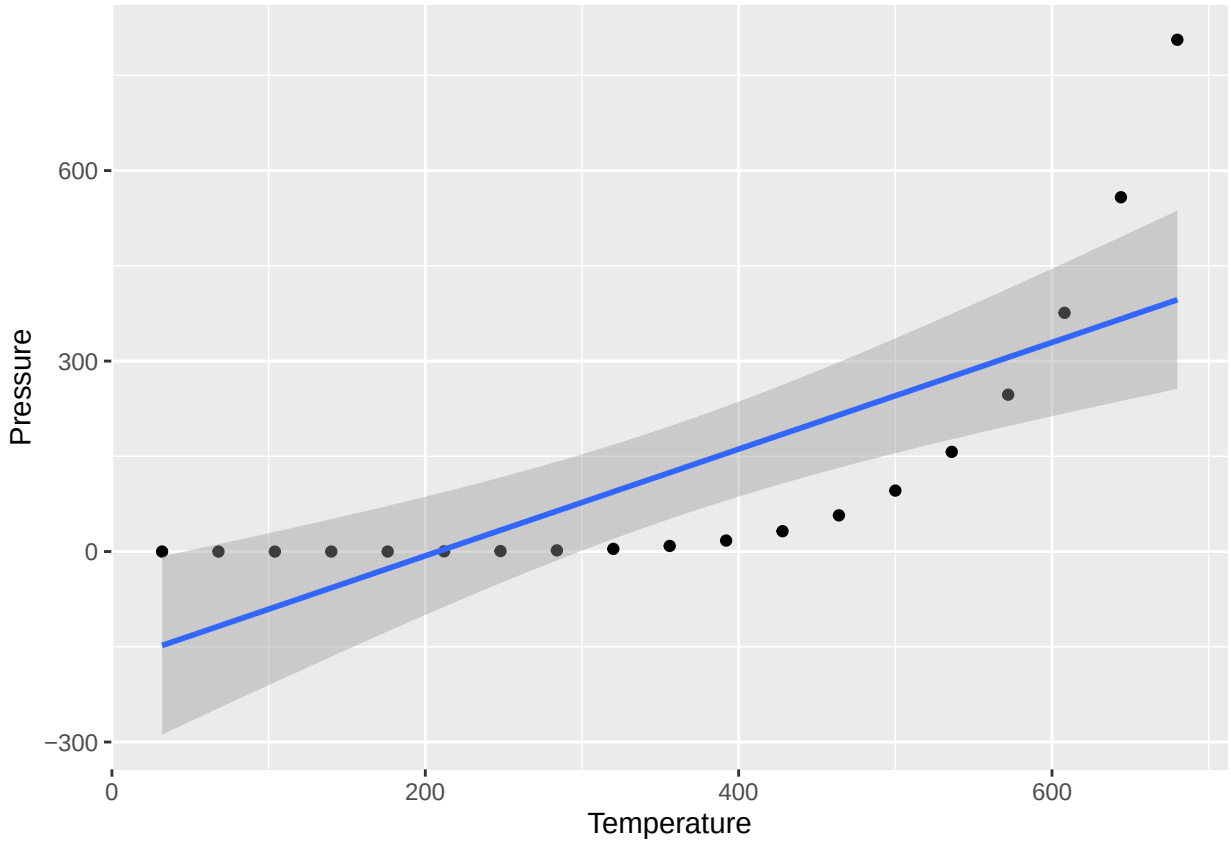
```
summary(lm(Pressure ~ 0 + Temperature, data = new_pressure))
```

```
##
## Call:
## lm(formula = Pressure ~ 0 + Temperature, data = new_pressure)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -166.72 -140.35  -91.98  -25.15  490.13
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## Temperature  0.46452    0.09609   4.834 0.000133 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 170.4 on 18 degrees of freedom
## Multiple R-squared:  0.5649, Adjusted R-squared:  0.5408
## F-statistic: 23.37 on 1 and 18 DF,  p-value: 0.000133
```

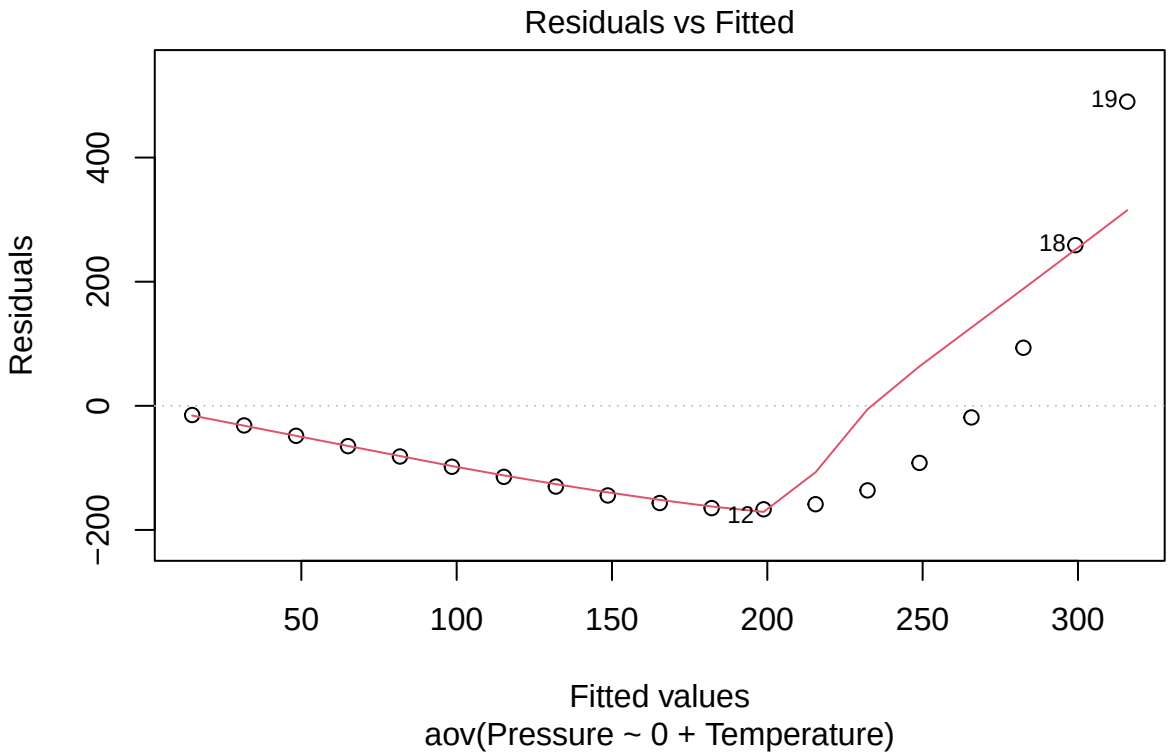
```
new_pressure %>%
  ggplot(aes(x = Temperature, y = Pressure)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



(e) Plot the residuals against the fitted values. Is pressure adequate to explain the relationship with temperature?

```r
one.way <- aov(Pressure ~ 0+Temperature, data = new_pressure)
plot(one.way, 1)
```



The points line up with the line, so the pressure is adequate to explain the relationship with temperature.