Assignment 1: Review of Simple Linear Regression

Matthew D. Sherman

Department of Mathematics and Statistics, University of Canterbury

STAT202: Regression Modelling

August 3, 2023

Preface Notes:

- All provided numbers are to four decimal places.
- check_model() was used instead of plot() as per lab instructions
- All plot images were exported at 1280x720 for continuity and accuracy (it can be hard to accurately interpret plots if they are small).
- Some lines of code appear as two lines on Word due to their length. All lines of code are intended to be one line.
- The full code is included at the end of the PDF

Question One:

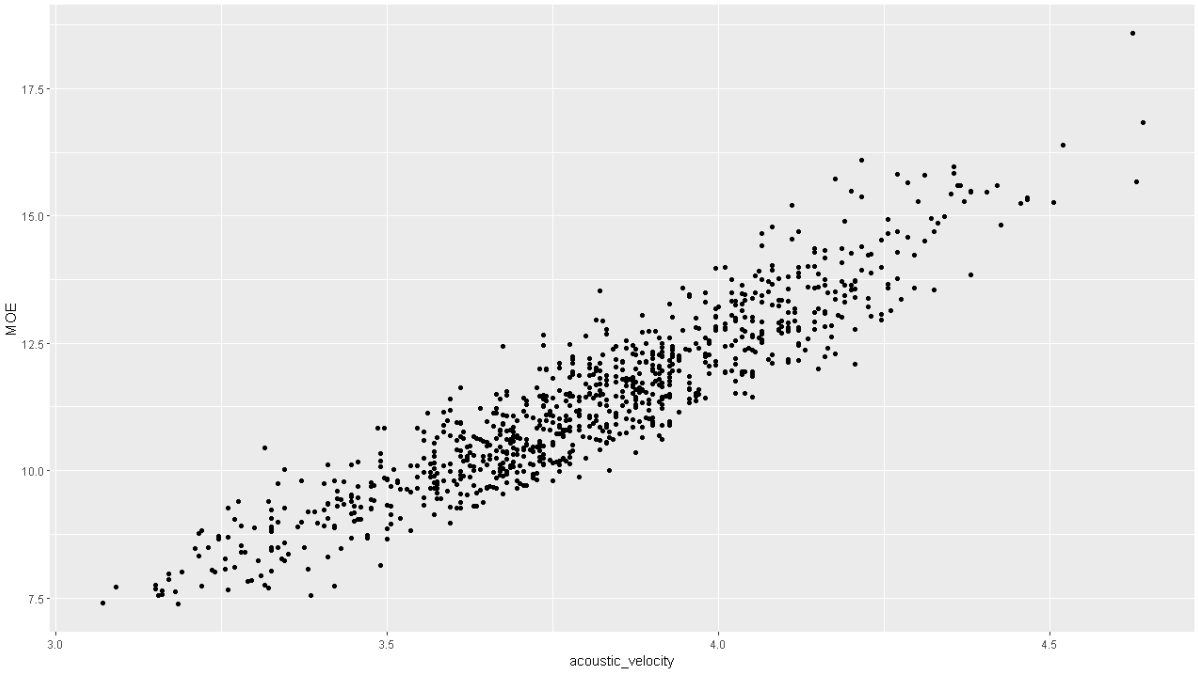
To answer this question, I wrote the following code based off the provided example.

```
eucalyptus <- read_csv('euc_tricarpa.csv')
set.seed(16645573)
my_eucalyptus <- eucalyptus %>% sample_n(900)
```

As per the question description, I used my student code (16645573) for set.seed().

Question Two:

```
ggplot(my_eucalyptus, aes(x= acoustic_velocity, y = MOE)) + geom_point()
```



Description of the plot (20 words or less):

The scatter plot shows a strong and positive relationship between the two variables MOE and acoustic velocity.

Question Three:

```
model_1 <- lm(MOE ~ acoustic_velocity, data = my_eucalyptus)
summary(model_1)
```

- Regression coefficients:
  - Intercept: -11.2915
  - Slope: 5.9525
- Standard error of the residuals: 0.6543 on 898 degrees of freedom
- Multiple R^2: 0.8595
- Adjusted-R^2: 0.8594

Linear regression equation:

$$y = -11.2915 + 5.9525x + e$$

Where y is MOE (in GPa), x is acoustic velocity (in km/s), and e is the error/residual term.

The intercept represents the estimated value of the response variable when the predictor variable/s are zero. In the context of the problem, an intercept of -11.2915 means that when acoustic velocity is 0 km/s, the modulus of elasticity (MOE) value is estimated to be -11.2915 gigapascals (GPa). However, an acoustic velocity of 0 km/s and an MOE of -11.2915 gigapascals is unrealistic and thus the intercept should be interpreted carefully within an appropriate data range. It would be wise to centre the predictor so that the intercept can be more practically interpreted.

The slope represents the change in the response variable for a change of one unit in the predictor variable/s. In the context of the problem, a slope of 5.9525 means that for every increase of one kilometre per second (km/s) in acoustic velocity, the estimated MOE value is expected to increase by 5.9525 GPa. This indicates a positive slope between acoustic velocity and MOE.

Question Four:

```
my_eucalyptus <- my_eucalyptus %>% mutate(cent_velocity = acoustic_velocity -
mean(acoustic_velocity))
model_2 <- lm(MOE ~ cent_velocity, data = my_eucalyptus)
summary(model_2)
```

New coefficients:

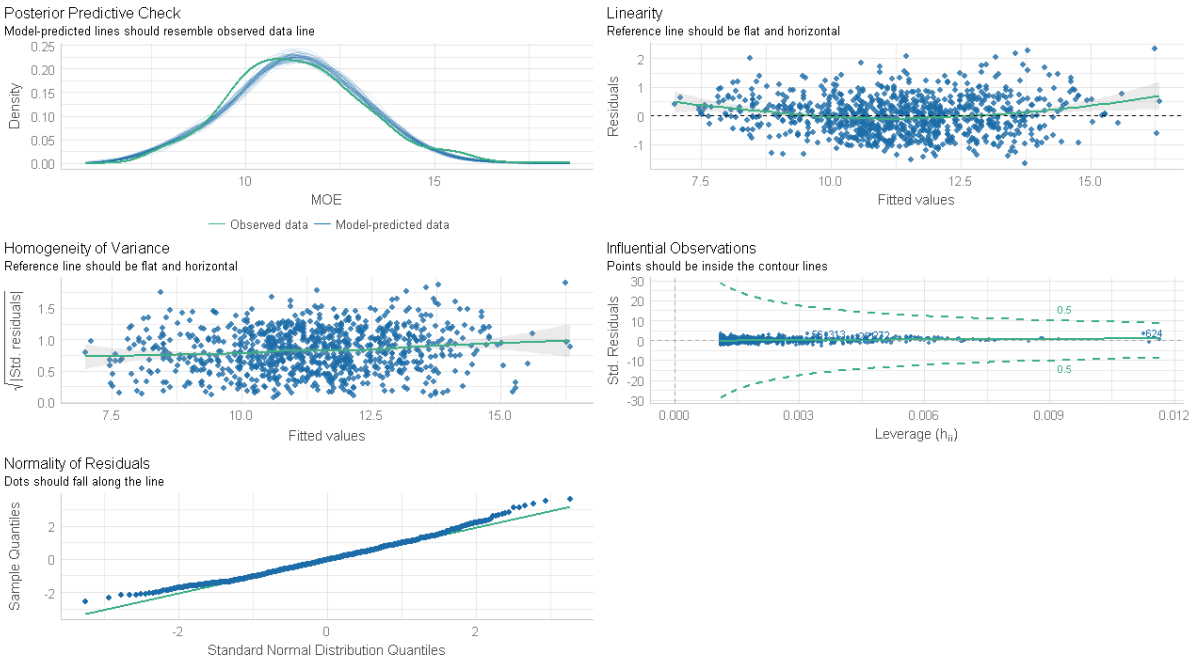
- Intercept: 11.3513
- Slope: 5.9525

The slope is the same (5.9525) for both sets of coefficients. This is because centring the predictor only affects the intercept and not the slope.

Centring acoustic velocity involves subtracting the mean of acoustic velocity from each acoustic velocity value (and in this example, we call the centred acoustic velocity 'cent_velocity'). Centring the predictor is done because it makes interpreting the intercept easier. This is because by doing this, the intercept now represents the average value of the response variable (MOE) when the predictor (acoustic velocity) is at its mean value. As a result of this, the intercept changes from -11.2915 (non-centred) to 11.3513 (centred). The new intercept of 11.3513 means that when the predictor (cent_velocity) is 0, then it is estimated that the response variable (MOE) is 11.3513 GPa. The new intercept can now be more easily and practically interpreted as 11.3513 GPa is realistic and not an extrapolation of the data.

Question Five:

```
check_model(model_2)
```



- Normality:
  To assess normality, we can refer to the 'Normality of Residuals' plot. To assume normality, the data points should closely fall along the diagonal line. The diagonal line represents the pattern of perfect normality, hence why it is desired that the data points fall along it as accurately as possible. As shown in the normality of residuals plot above, the data points mostly fall along the diagonal line accurately. However, the data points deviate slightly from the diagonal line at the lower and higher ends of the x-axis. Nevertheless, we can still assume that the residuals meet the normality assumption as the deviations are minimal.
- Equal Variance:
  To assess equal variance, we can refer to the 'Homogeneity of Variance' plot. To assume equal variance, the reference line should be flat and horizontal. The reference line is a representation of the spread and pattern of the points on the plot. As shown on the homogeneity of variance plot, the points are reasonably spread out and, in turn, the reference line is very flat. It is not perfectly horizontal as it has a slight positive trend, but it is still satisfactory. Hence, we can assume that the assumption of equal variance is met.

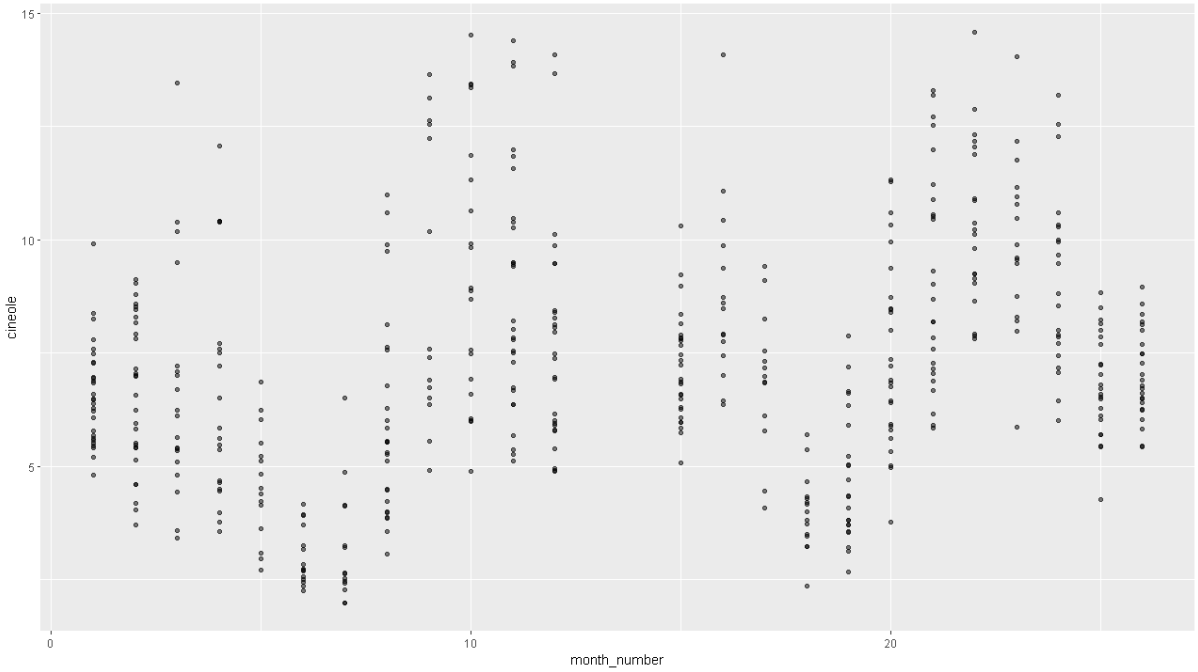
Question Six:

```
cineole <- read_csv('cineole.csv')
set.seed(16645573)
my_cineole <- cineole %>% sample_n(500)
```

Used my student code (16645573) for set.seed().

Question Seven:

```
ggplot(my_cineole, aes(x= month_number, y = cineole)) + geom_point(alpha = 0.5)
```

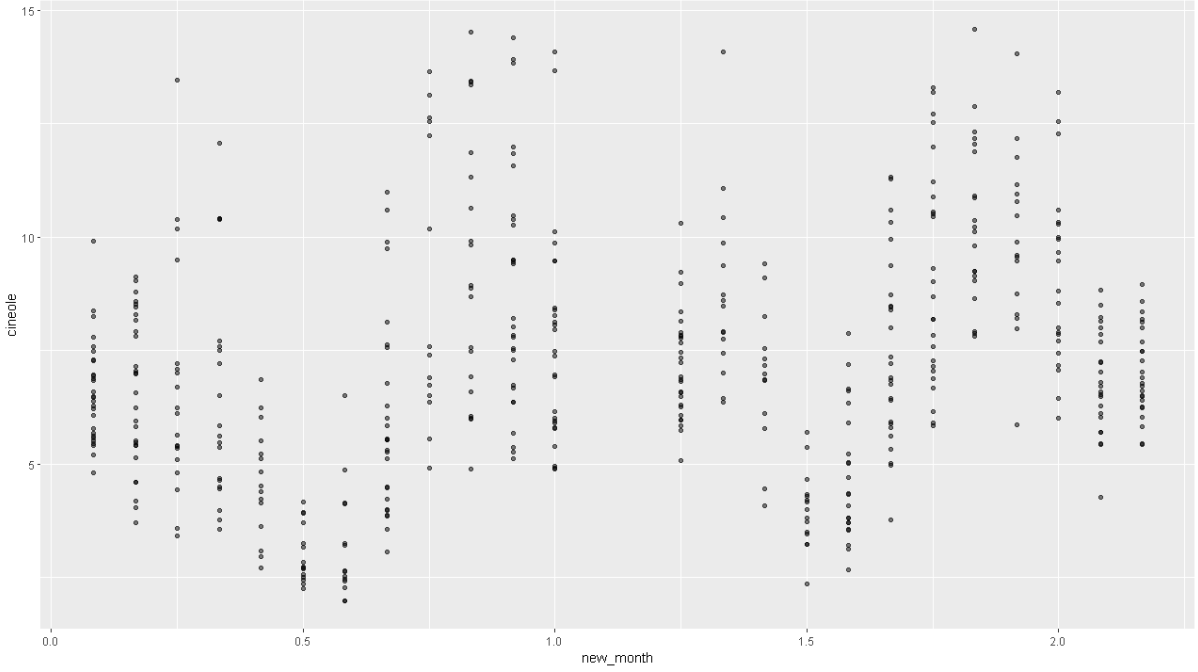


Explanation of the relationship in 50 words or less:

The scatterplot appears to show a seasonal pattern between cineole production and month number. The production levels go through recurring peaks and troughs over the months. This indicates that cineole production may be affected by the weather and climate of the different seasons that correlate with the month number.

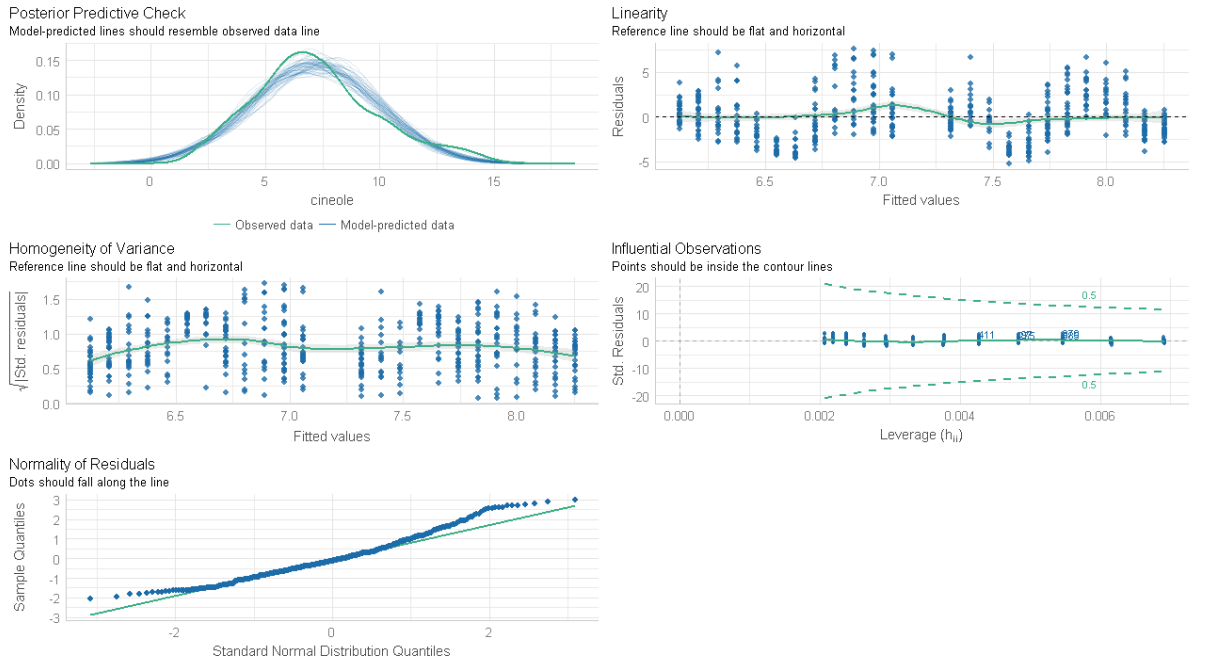
Question Eight:

```
my_cineole <- my_cineole %>% mutate(new_month = month_number / 12)
ggplot(my_cineole, aes(x= new_month, y = cineole)) + geom_point(alpha = 0.5)
```



Question Nine:

```
oil_1 <- lm(cineole ~ new_month, data = my_cineole)
check_model(oil_1)
```



Comment on linearity using the linearity (residuals vs fitted) plot (50 words or less):

The reference line is satisfactory for the lower/higher ends of the x-axis, but it oscillates in the central region. This results in a reference line which while generally horizontal, is not entirely flat (due to the seasonal pattern). Hence, we can say that the assumption of linearity is not met.

Question Ten:

```
oil_2 <- lm(cineole ~ (sin(2*pi*new_month) + cos(2*pi*new_month)), data =
my_cineole)
summary(oil_2)
```

Regression Coefficients (oil_2):

- Intercept: 7.0283
- sin(2 * pi * new_month): -0.8595
- cos(2 * pi * new_month): 1.7635

Residual Standard Error (RSE):

- oil_1: 2.575
- oil_2: 2.315

Explanation (100 words or less):

The RSE for oil_2 (2.315) being lower than the RSE for oil_1 (2.575) suggests that oil_2 has a better fit than oil_1 and is therefore more accurate. This is because a smaller RSE captures more variance and in turn results in fewer prediction errors. The RSE for oil_2 is smaller because using (sin(2*pi*new_month) + cos(2*pi*new_month)) instead of new_month as the response, oil_2 can account for seasonal changes (as mentioned in Q7) better. This is because using

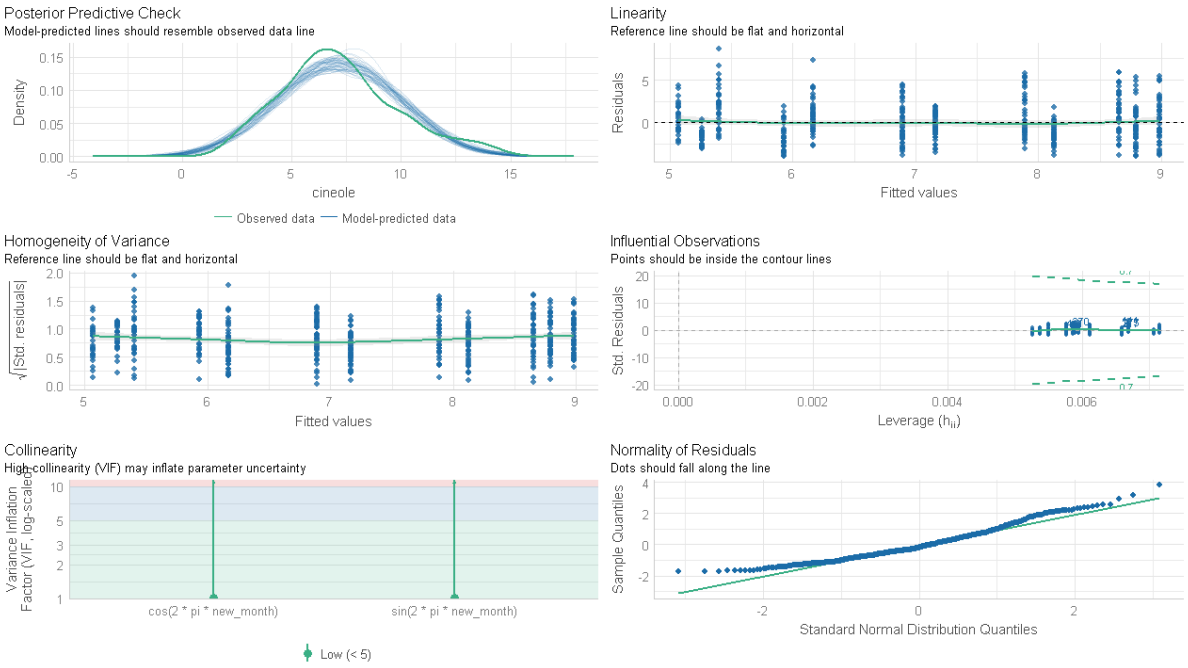
(sin(2*pi*new_month) + cos(2*pi*new_month)), periodic/cyclical patterns can be accounted for, and thus it is more accurate than new_month alone.

Question Eleven:

Using check_model(oil_2) to generate plots to comment on linearity

```
check_model(oil_2)
```

Result of running check_model(oil_2):



Comment on linearity using the linearity (residuals vs fitted) plot (50 words or less):

The points appear to be scattered randomly around the x-axis with no defining pattern, indicating linearity. Moreover, the green reference line is approximately flat and horizontal and follows the y-axis line very accurately with only minute deviations throughout. Hence, we can say that the assumption of linearity is met.

Question Twelve:

I have uploaded this PDF to LEARN and have included the entire code below.

```
library(tidyverse)
library(performance)


eucalyptus <- read_csv('euc_tricarpa.csv')
set.seed(16645573)
my_eucalyptus <- eucalyptus %>% sample_n(900)


ggplot(my_eucalyptus, aes(x= acoustic_velocity, y = MOE)) + geom_point()


model_1 <- lm(MOE ~ acoustic_velocity, data = my_eucalyptus)
summary(model_1)


my_eucalyptus <- my_eucalyptus %>% mutate(cent_velocity = acoustic_velocity -
mean(acoustic_velocity))
model_2 <- lm(MOE ~ cent_velocity, data = my_eucalyptus)
summary(model_2)


check_model(model_2)


cineole <- read_csv('cineole.csv')
set.seed(16645573)
my_cineole <- cineole %>% sample_n(500)


ggplot(my_cineole, aes(x= month_number, y = cineole)) + geom_point(alpha = 0.5)


my_cineole <- my_cineole %>% mutate(new_month = month_number / 12)
ggplot(my_cineole, aes(x= new_month, y = cineole)) + geom_point(alpha = 0.5)


oil_1 <- lm(cineole ~ new_month, data = my_cineole)
check_model(oil_1)


oil_2 <- lm(cineole ~ (sin(2*pi*new_month) + cos(2*pi*new_month)), data = my_cineole)
summary(oil_2)


check_model(oil_2)
```