# Assignment1

## Question1

library(tidyverse)

eucalyptus <- read\_csv('euc\_tricarpa.csv') # Read Eucalyptus tricarpa data set

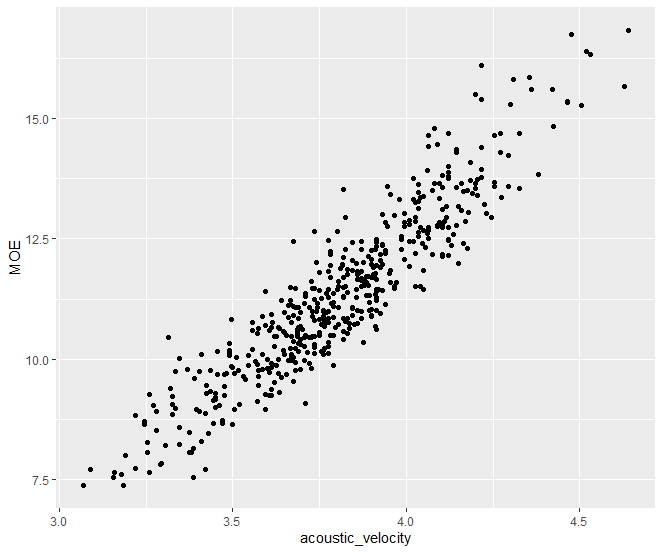
set.seed(17053777)

my\_eucs <- eucalyptus %>% sample\_n(500)

## Question2

ggplot(my\_eucs, aes(acoustic\_velocity, MOE)) + geom\_point()

# As acoustic velocity increases, we observe that MOE increases as well. They are in a strong positive linear relationship.



## Question3

model1 <- lm(MOE ~ acoustic\_velocity, data = my\_eucs)

summary(model1)

# regression coefficients: MOE = -11.0226 + 5.8747 \* acoustic\_velocity

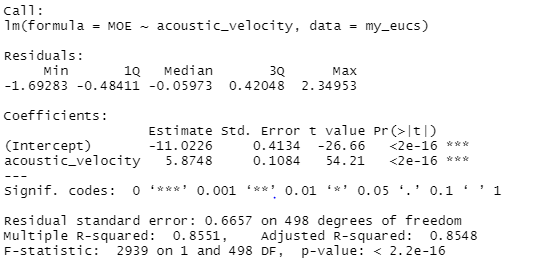
# standard error of the residuals: 0.6657 on 498 degrees of freedom

# multiple R-squared: 0.8551

# adjusted R-squared: 0.8548

# the meaning of the intercept: when the value of x is 0, the value of y is the intercept. In this case, when acoustic velocity is 0, the value of MOE is -11.4511.

# the meaning of the slope: when the value of x increased by 1 unit, how much the value in y will increase. In this case, As the acoustic velocity increases 1 unit, the value of MOE increases by 5.8747.



## Question4

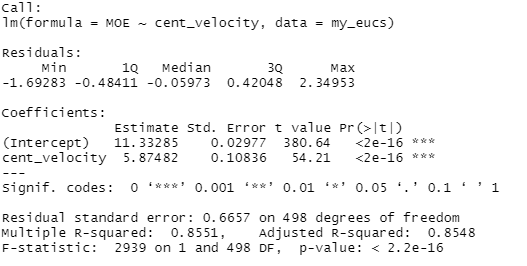
my\_eucs = my\_eucs %>% mutate(cent\_velocity = acoustic\_velocity - mean(acoustic\_velocity))

model2 <- lm(MOE ~ cent\_velocity, data = my\_eucs)

summary(model2)

# regression coefficients: MOE = 11.33285 + 5.87482 \* cen\_velocity

# Comparing to the non-centred analyses, the biggest difference is the intercept in model2 regression coefficient is positive. As acoustic\_velocity - mean(acoustic\_velocity) = 0, which means when acoustic velocity equals the average value of acoustic velocity, the intercept is 11.33285.



## Question5

model3 <- lm(MOE ~ (acoustic\_velocity + dry\_density), data = my\_eucs)

summary(model3)

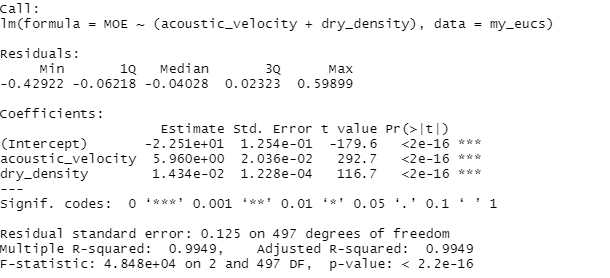
# regression coefficients: MOE = -22.51 + (5.960 \* acoustic\_velocity + 0.01434 \* dry\_density)

# the standard error of the residuals: 0.125 on 497 degrees of freedom

# multiple R-squared: 0.9949

# adjusted R-squared: 0.9949

# For the same data set, the adjusted R-squared increases when the new term improves the model more than would be expected by chance and smaller standard error of residuals means predictions are better. We can observe that the both standard error of the residuals an adjusted R-squared values of model3 are higher than the values of model1 and model2, so the model fir improved.



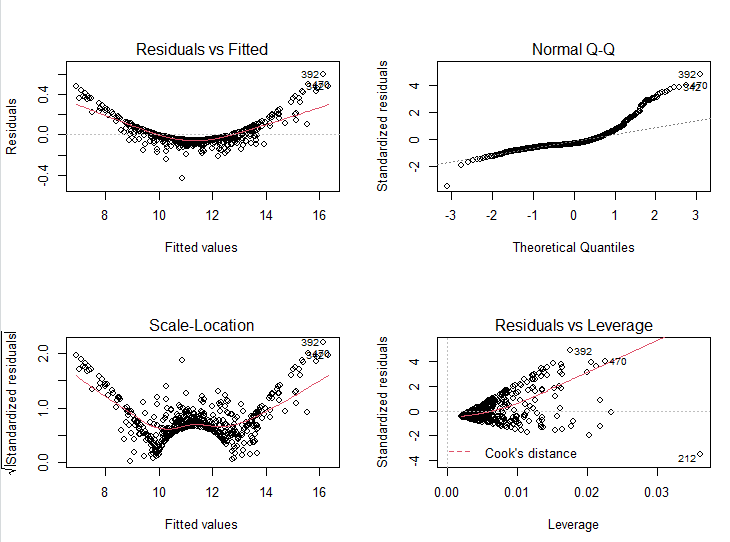
## Question6

par(mfrow = c(2, 2))

plot(model3)

# From the graph Normal Q-Q, we can observe that the line of theoretical quantiles less than 1 is approximately straight but the data afterwards cannot fit the assumptions, so we cannot say that the residuals meet the normality.

# From the graph Residuals VS Fitted, we can observe that the residuals seem not to be scattered randomly around 0, then the equal variance assumptions are not likely to be met



## Quetion7

library(readr)

write\_csv(x = my\_eucs, path='eucs.csv')

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.924723 |  |  |  |  |  |  |  |
| R Square | 0.855113 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.854822 |  |  |  |  |  |  |  |
| Standard Error | 0.665743 |  |  |  |  |  |  |  |
| Observations | 500 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 1302.673 | 1302.673 | 2939.155 | 4.8E-211 |  |  |  |
| Residual | 498 | 220.7203 | 0.443214 |  |  |  |  |  |
| Total | 499 | 1523.393 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| Intercept | -11.0226 | 0.413429 | -26.6614 | 5.88E-98 | -11.8349 | -10.2103 | -11.8349 | -10.2103 |
| X Variable 1 | 5.87482 | 0.108364 | 54.21397 | 4.8E-211 | 5.661914 | 6.087726 | 5.661914 | 6.087726 |
|  |  |  |  |  |  |  |  |  |