# Asignment2

## Question1

library(tidyverse)

set.seed(17053777)

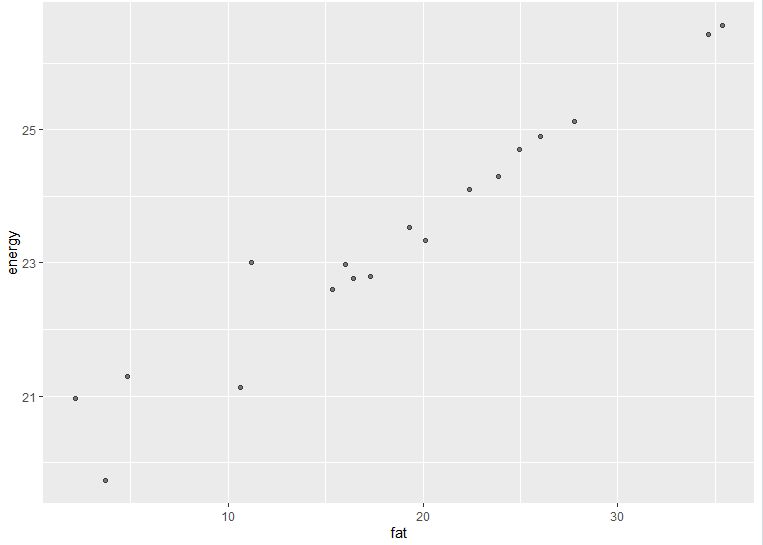
fish <- read\_csv('http://stats.apiolaza.net/data/fish.csv')

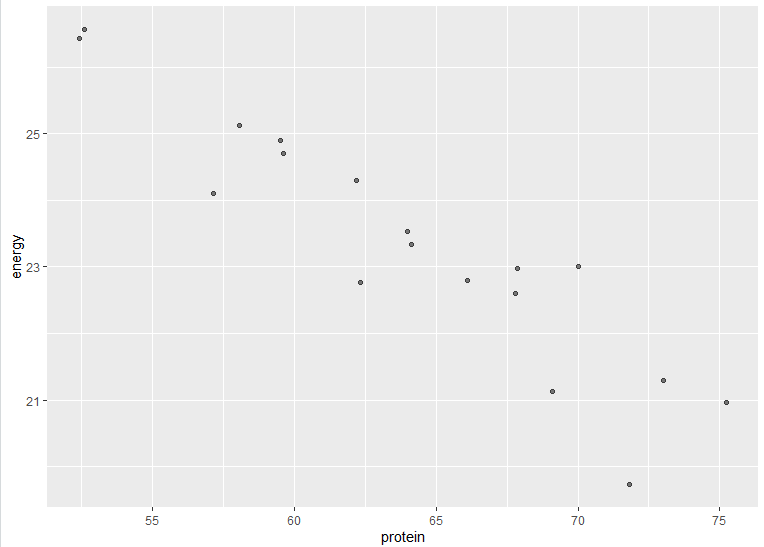
my\_fish <- fish %>% sample\_n(18)

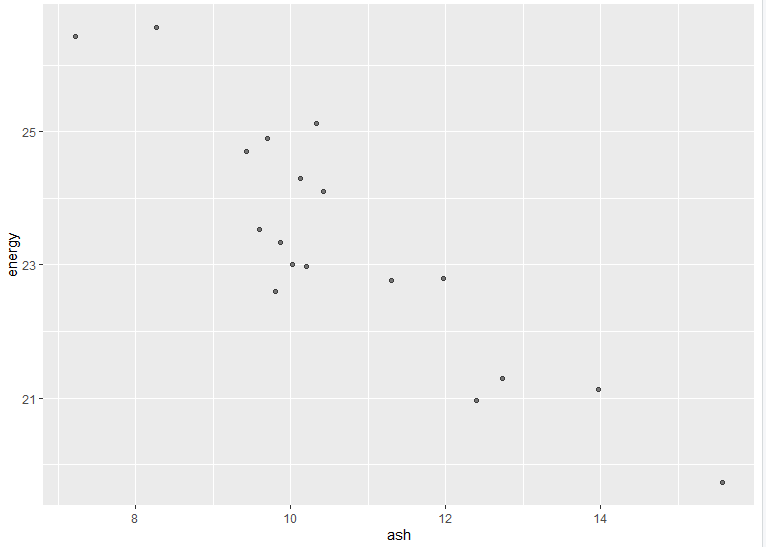
ggplot(my\_fish, aes(fat, energy)) + geom\_point(alpha = 0.5)

ggplot(my\_fish, aes(protein, energy)) + geom\_point(alpha = 0.5)

ggplot(my\_fish, aes(ash, energy)) + geom\_point(alpha = 0.5)







# As fat increases, we observe that the energy increases. They are in a strong positive linear relationship.

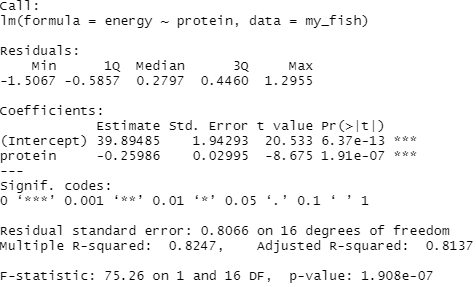
# When protein increases, we found the energy decreases. They are in a strong negative linear relationship.

# As ash increases, we notice that the energy decreases. They are in a strong negative linear relationship.

## Question2

m1 <- lm(energy ~ protein, data = my\_fish)

summary(m1)

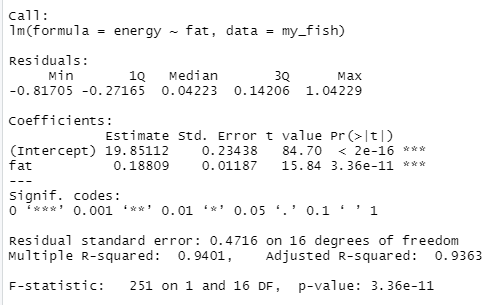


# R-squared: 0.8247, Adjusted R-squared: 0.8137

# residual standard errors: 0.8066 on 16 degree of freedom

m2 <- lm(energy ~ fat, data = my\_fish)

summary(m2)

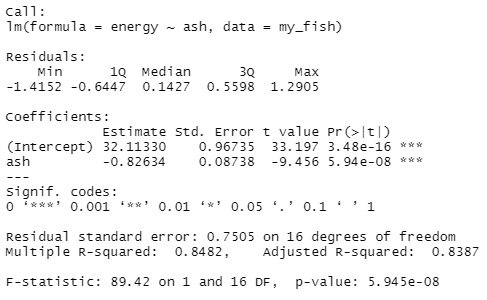


# R-squared: 0.9401, Adjusted R-squared: 0.9363

# residual standard errors: 0.4716 on 16 degree of freedom

m3 <- lm(energy ~ ash, data = my\_fish)

summary(m3)

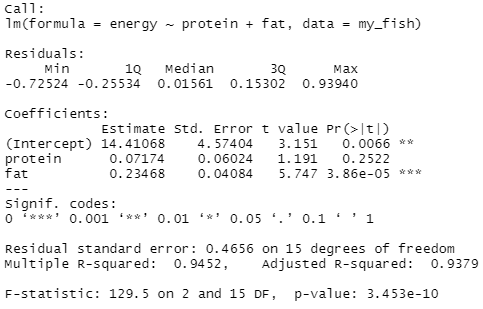


# R-squared: 0.8482, Adjusted R-squared: 0.8387

# residual standard errors: 0.7505 on 16 degree of freedom

m4 <- lm(energy ~ protein + fat, data = my\_fish)

summary(m4)

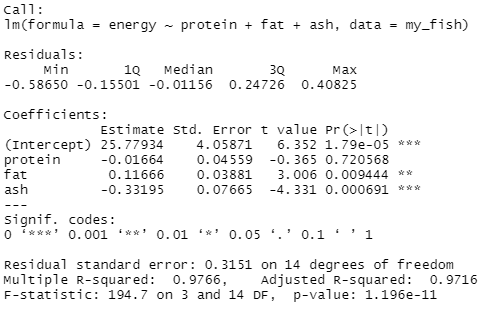


# R-squared: 0.9452, Adjusted R-squared: 0.9379

# residual standard errors: 0.4656 on 15 degree of freedom

m5 <- lm(energy ~ protein + fat + ash, data = my\_fish)

summary(m5)



# R-squared: 0.9766, Adjusted R-squared: 0.9716

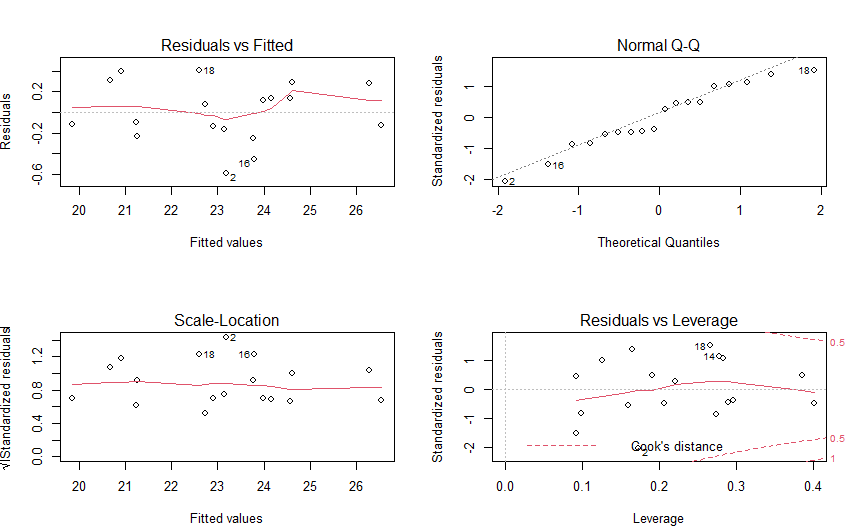
# residual standard errors: 0.3151 on 14 degree of freedom

# As we know, adjusted R-squared increases when the new term improves the model more than would be expected by chance. From m1 to m5, we can observe that the R-squared and Adjusted R-squared increase and residual standard errors decrease. We can say the model of fit improves.

## Question3

par(mfrow = c(2, 2))

plot(m5)



# We need to check whether the residual is normally distributed with 0 mean and constant variance.

# From the Residuals vs Fitted graph we obverse that the plot is randomly scatter around the 0 mean which indicates that the residuals and the fitted values are uncorrelated.

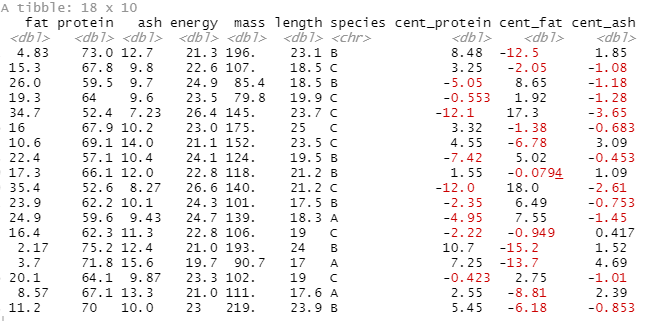
## Question4

my\_fish = my\_fish %>% mutate(cent\_protein = protein - mean(protein))

my\_fish = my\_fish %>% mutate(cent\_fat = fat - mean(fat))

my\_fish = my\_fish %>% mutate(cent\_ash = ash - mean(ash))

m6 <- lm(energy ~ cent\_protein + cent\_fat + cent\_ash, data = my\_fish)



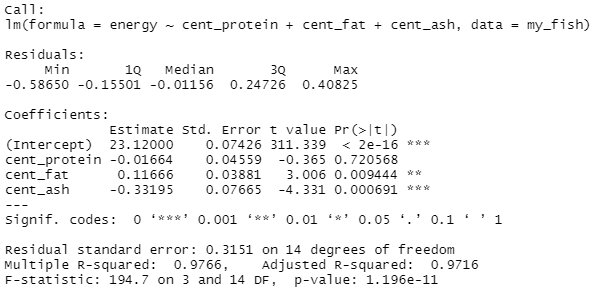
## Question5

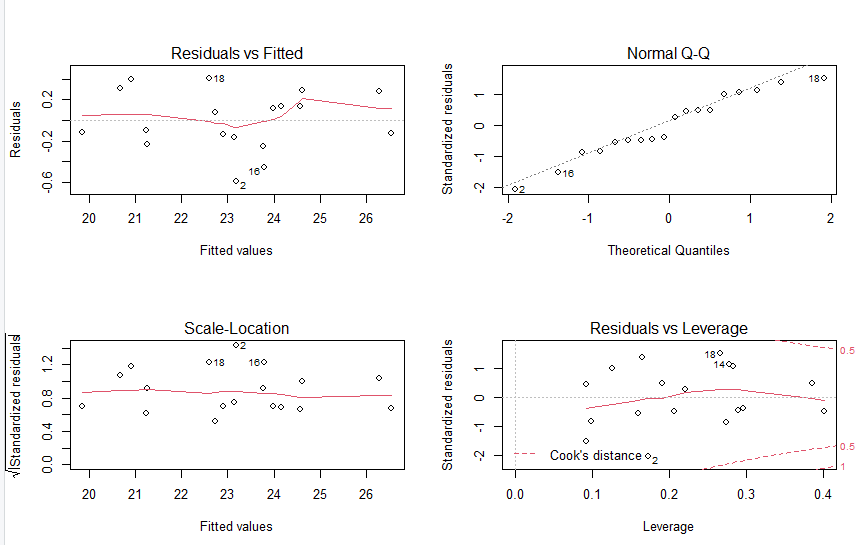
m6 <- lm(energy ~ cent\_protein + cent\_fat + cent\_ash, data = my\_fish)

summary(m6)

par(mfrow = c(2, 2))

plot(m6)





# The only difference between m5 and m6 is we centred the variables. The intercept is different but the slope is exactly the same. m6 still the same model but coordinator changed. The slope indicates the steepness of a line and the intercept indicates the locations where it intersects an axis.