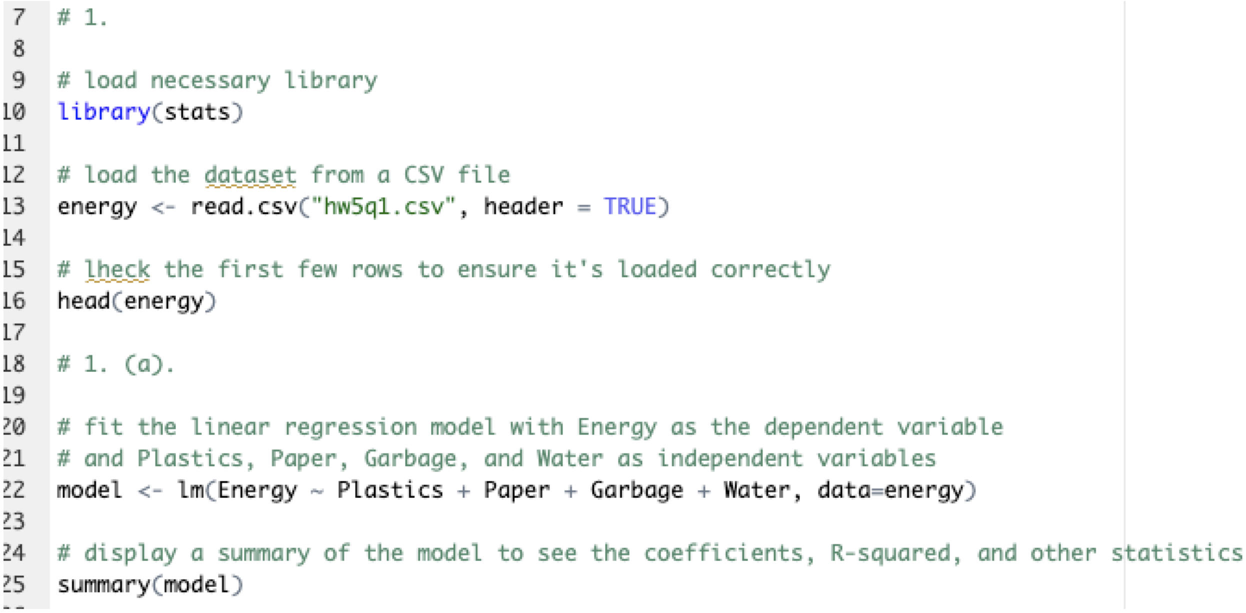
Aidan Kiser
31 March 2024
STAT 3010
Xia

Assignment 5

All screenshots that are included are from my own R Script file. You can access it with this link: https://github.com/kiseraidan/STAT-3010/blob/main/Assignment%205/Assignment5.R

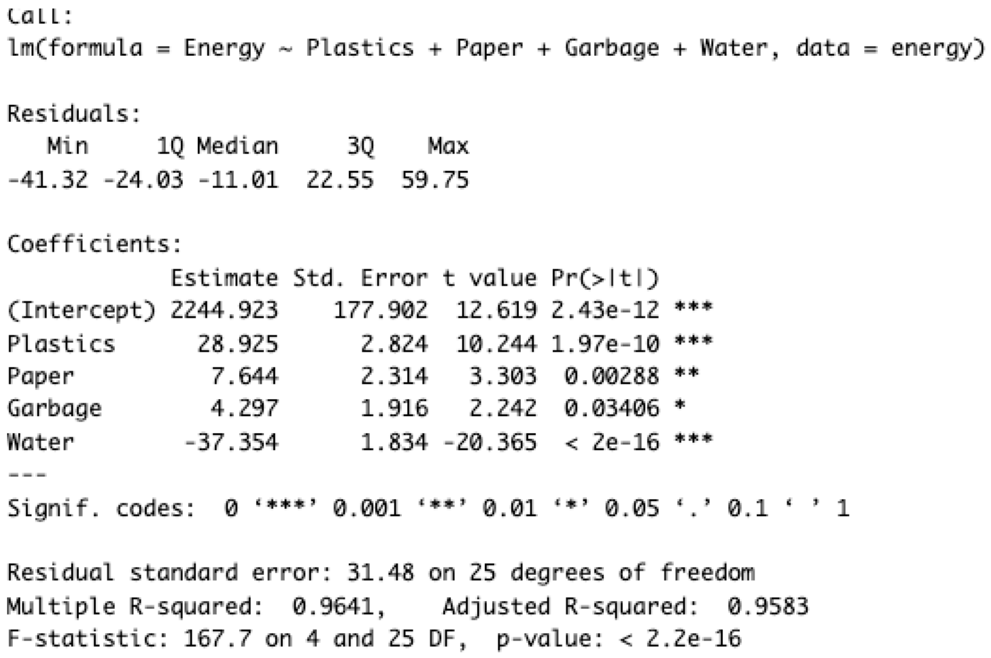
1. (a).

   Code:



```
# 1.

# load necessary library
library(stats)

# load the dataset from a CSV file
energy <- read.csv("hw5q1.csv", header = TRUE)

# lheck the first few rows to ensure it's loaded correctly
head(energy)

# 1. (a).

# fit the linear regression model with Energy as the dependent variable
# and Plastics, Paper, Garbage, and Water as independent variables
model <- lm(Energy ~ Plastics + Paper + Garbage + Water, data=energy)

# display a summary of the model to see the coefficients, R-squared, and other statistics
summary(model)
```

   Output:



```
Call:
lm(formula = Energy ~ Plastics + Paper + Garbage + Water, data = energy)

Residuals:
   Min     1Q Median     3Q    Max 
-41.32 -24.03 -11.01  22.55  59.75 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2244.923    177.902  12.619 2.43e-12 ***
Plastics      28.925      2.824  10.244 1.97e-10 ***
Paper          7.644      2.314   3.303  0.00288 ** 
Garbage        4.297      1.916   2.242  0.03406 *  
Water        -37.354      1.834 -20.365  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.48 on 25 degrees of freedom
Multiple R-squared:  0.9641,	Adjusted R-squared:  0.9583 
F-statistic: 167.7 on 4 and 25 DF,  p-value: < 2.2e-16
```

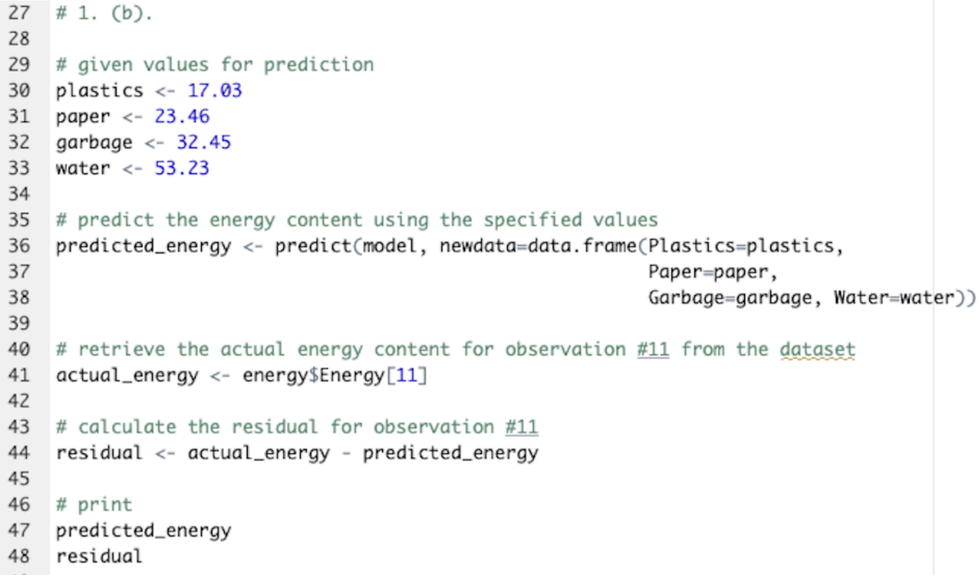
Report:

- Plastics: Show a strong positive impact on energy content, with each percentage increase contributing approximately 28.93 kcal/kg.
- Paper: Also positively affects energy content, though with a lesser magnitude (7.64 kcal/kg per percentage increase) compared to plastics.
- Garbage: Positively related but with a smaller effect (4.30 kcal/kg per percentage increase) on energy content.
- Water (Moisture): Contrasts with other materials by negatively impacting energy content, decreasing it by about 37.35 kcal/kg for each percentage increase.

  The model's R-squared value at 0.964 indicates a high ability to explain the energy content variability through these predictors. However, a cautionary note on potential multicollinearity suggests the need for further scrutiny of predictor relationships to confirm the model's stability and accuracy.

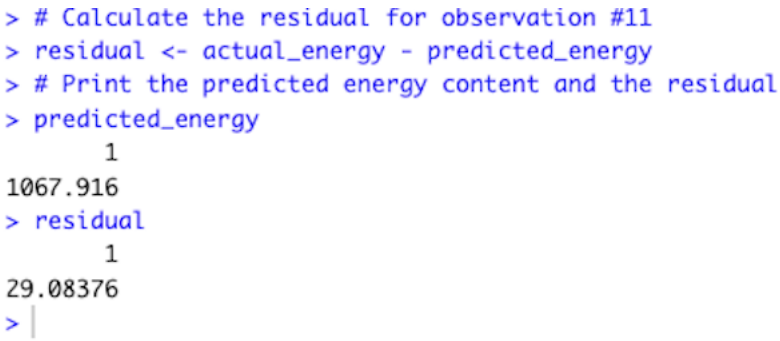
(b).

Code:



```
# 1. (b).

# given values for prediction
plastics <- 17.03
paper <- 23.46
garbage <- 32.45
water <- 53.23

# predict the energy content using the specified values
predicted_energy <- predict(model, newdata=data.frame(Plastics=plastics,
                                                      Paper=paper,
                                                      Garbage=garbage, Water=water))

# retrieve the actual energy content for observation #11 from the dataset
actual_energy <- energy$Energy[11]

# calculate the residual for observation #11
residual <- actual_energy - predicted_energy

# print
predicted_energy
residual
```

Output:



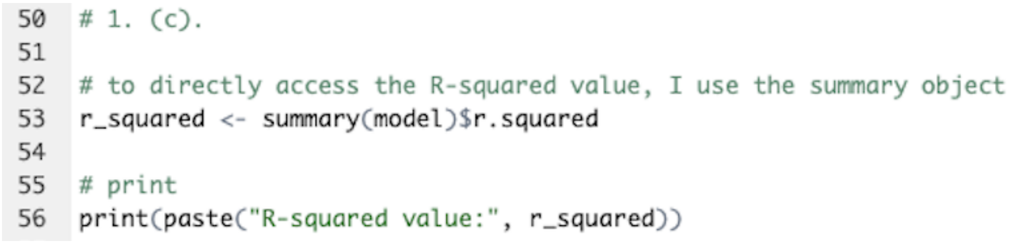
```
> # Calculate the residual for observation #11
> residual <- actual_energy - predicted_energy
> # Print the predicted energy content and the residual
> predicted_energy
       1
1067.916
> residual
       1
29.08376
> 
```

Report:

My R script uses the predict() function with the fitted model ‘model’ to estimate the energy content based on values for plastics, paper, garbage, and water. It then calculates the residual by subtracting this predicted value from the actual energy content of observation #11 in dataset. The predicted energy content and residual are 1067.916 and 29.08376, respectively.

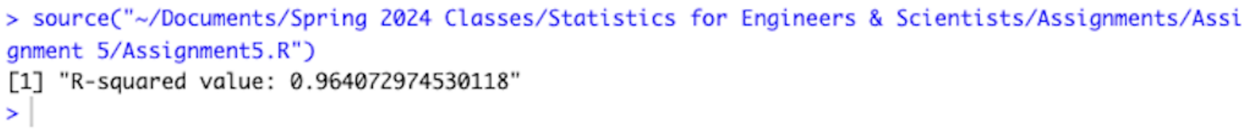
(c).

Code:



```r
# 1. (c).

# to directly access the R-squared value, I use the summary object
r_squared <- summary(model)$r.squared

# print
print(paste("R-squared value:", r_squared))
```

Output:



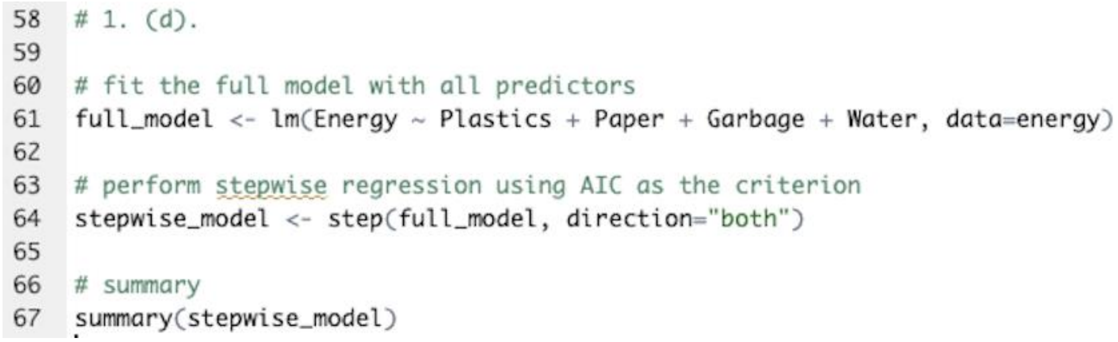
```
> source("~/Documents/Spring 2024 Classes/Statistics for Engineers & Scientists/Assignments/Assi
gnment 5/Assignment5.R")
[1] "R-squared value: 0.964072974530118"
> 
```

Report:

> The proportion of observed variation in energy content that can be explained by the approximate relationship between energy content and the four predictors (plastics, paper, garbage, and water) is quantified by the R-squared value of the regression model. From the summary provided earlier, the R-squared value for the model was 0.964. This means that approximately 96.4% of the variance in energy content can be explained by the model that includes the four predictors.

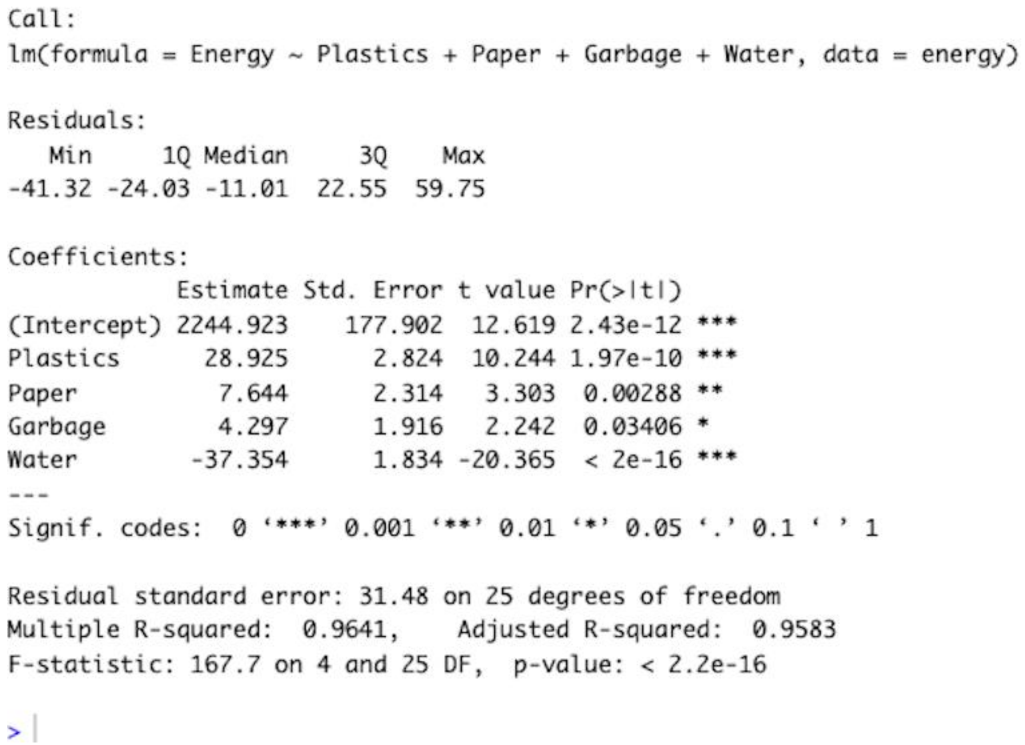
(d).

Code:



```r
# 1. (d).

# fit the full model with all predictors
full_model <- lm(Energy ~ Plastics + Paper + Garbage + Water, data=energy)

# perform stepwise regression using AIC as the criterion
stepwise_model <- step(full_model, direction="both")

# summary
summary(stepwise_model)
```

Output:



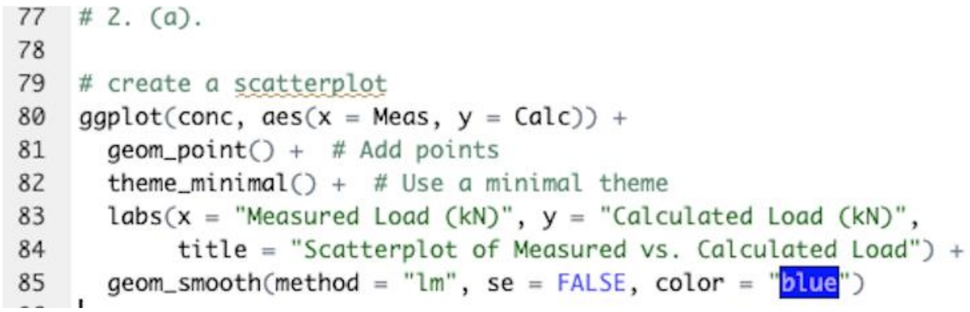
```
Call:
lm(formula = Energy ~ Plastics + Paper + Garbage + Water, data = energy)

Residuals:
   Min     1Q Median     3Q    Max 
-41.32 -24.03 -11.01  22.55  59.75 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2244.923    177.902  12.619 2.43e-12 ***
Plastics      28.925      2.824  10.244 1.97e-10 ***
Paper          7.644      2.314   3.303  0.00288 ** 
Garbage        4.297      1.916   2.242  0.03406 *  
Water        -37.354      1.834 -20.365  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.48 on 25 degrees of freedom
Multiple R-squared:  0.9641,	Adjusted R-squared:  0.9583 
F-statistic: 167.7 on 4 and 25 DF,  p-value: < 2.2e-16

> 
```

Report:

The selected model should theoretically provide a robust framework for predicting the energy content of waste based on its composition, balancing the complexity of the model with its explanatory power.

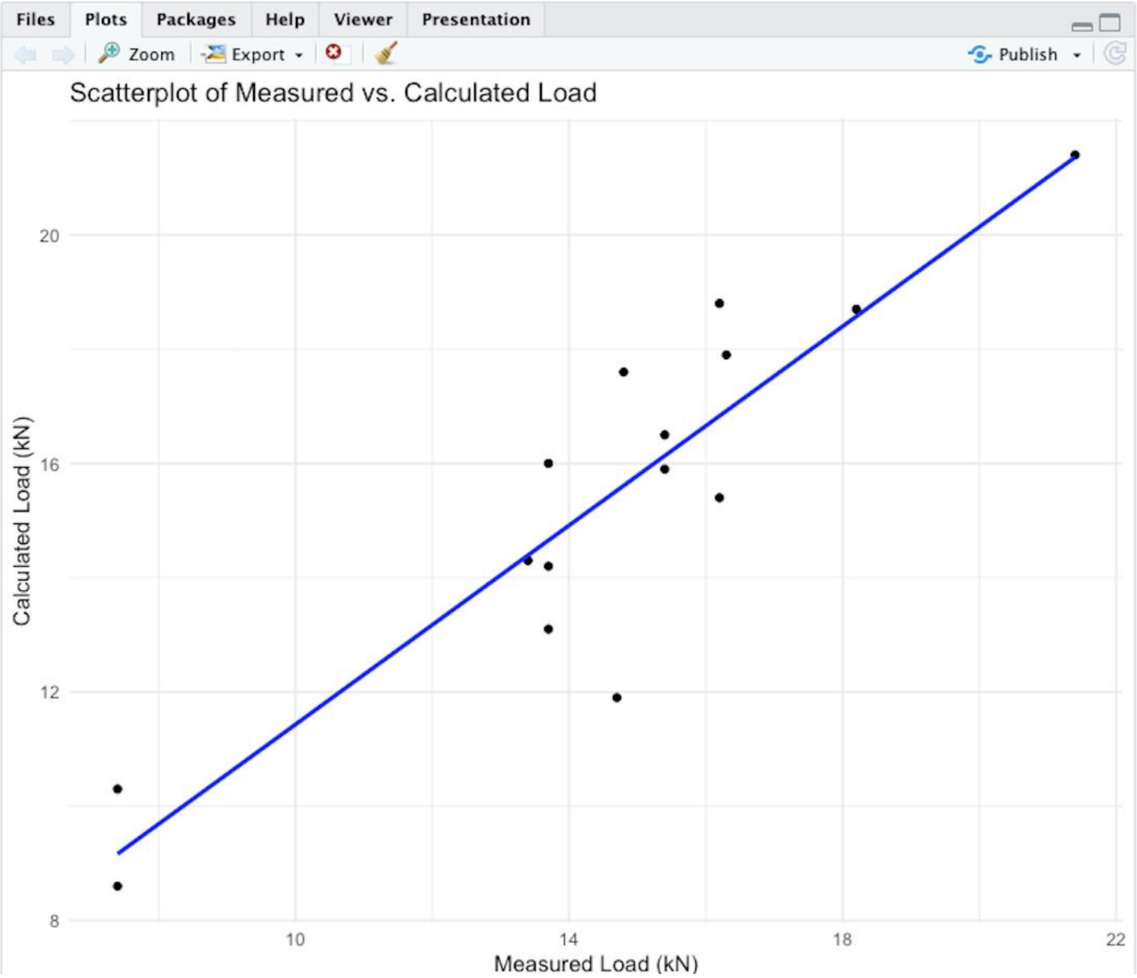
2. (a).

Code:



```
# 2. (a).

# create a scatterplot
ggplot(conc, aes(x = Meas, y = Calc)) +
  geom_point() +  # Add points
  theme_minimal() +  # Use a minimal theme
  labs(x = "Measured Load (kN)", y = "Calculated Load (kN)",
       title = "Scatterplot of Measured vs. Calculated Load") +
  geom_smooth(method = "lm", se = FALSE, color = "blue")
```

Output:

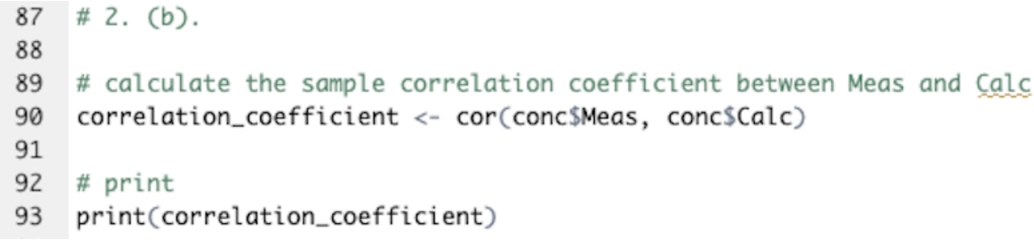


Report:

The scatterplot demonstrates a positive correlation between the measured and calculated loads of carbon FRP-concrete samples: as the measured load increases, so does the calculated load. The points are generally close to the regression line, suggesting that the calculated load is a good predictor of the measured load, though not perfectly matching. The trend is consistent across low to high loads, indicating a reliable mathematical model with some variability.

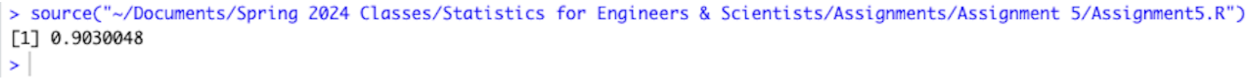
(b).

Code:



```
# 2. (b).

# calculate the sample correlation coefficient between Meas and Calc
correlation_coefficient <- cor(conc$Meas, conc$Calc)

# print
print(correlation_coefficient)
```

Output:



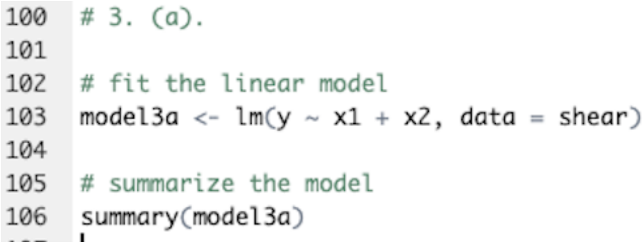
```
> source("~/Documents/Spring 2024 Classes/Statistics for Engineers & Scientists/Assignments/Assignment 5/Assignment5.R")
[1] 0.9030048
> 
```

Report:

Yes, The scatterplot indicated a positive relationship between the measured and calculated loads, suggesting that as one increases, so does the other.

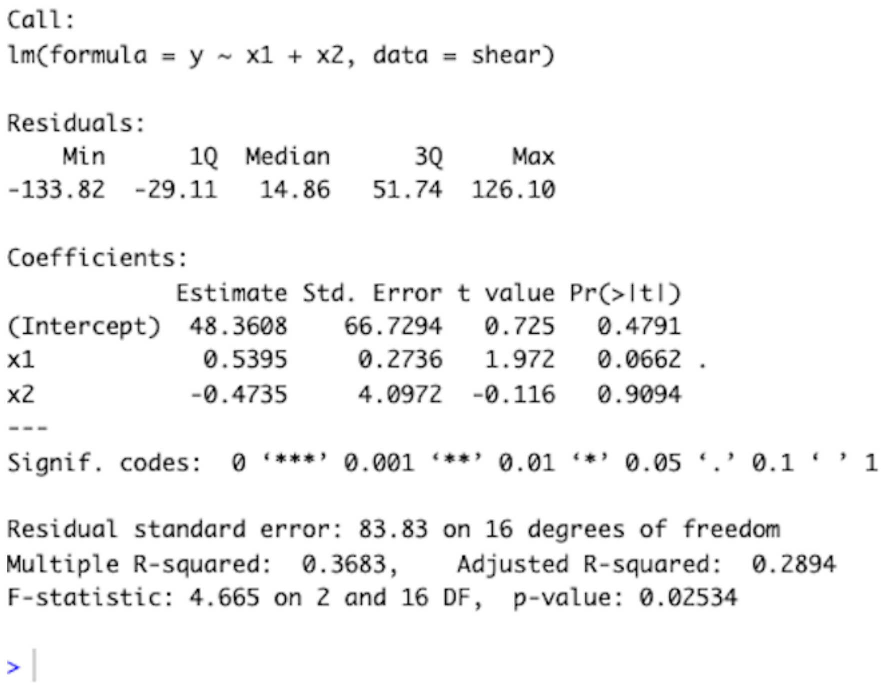
3. (a).

Code:



```
# 3. (a).

# fit the linear model
model3a <- lm(y ~ x1 + x2, data = shear)

# summarize the model
summary(model3a)
```

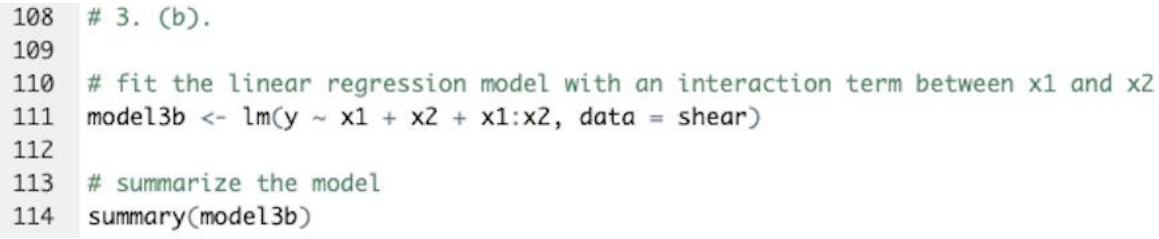
Output:



```
Call:
lm(formula = y ~ x1 + x2, data = shear)

Residuals:
    Min      1Q  Median      3Q     Max 
-133.82  -29.11   14.86   51.74  126.10 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  48.3608    66.7294   0.725   0.4791  
x1            0.5395     0.2736   1.972   0.0662 .
x2           -0.4735     4.0972  -0.116   0.9094  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 83.83 on 16 degrees of freedom
Multiple R-squared:  0.3683,	Adjusted R-squared:  0.2894 
F-statistic: 4.665 on 2 and 16 DF,  p-value: 0.02534

> 
```

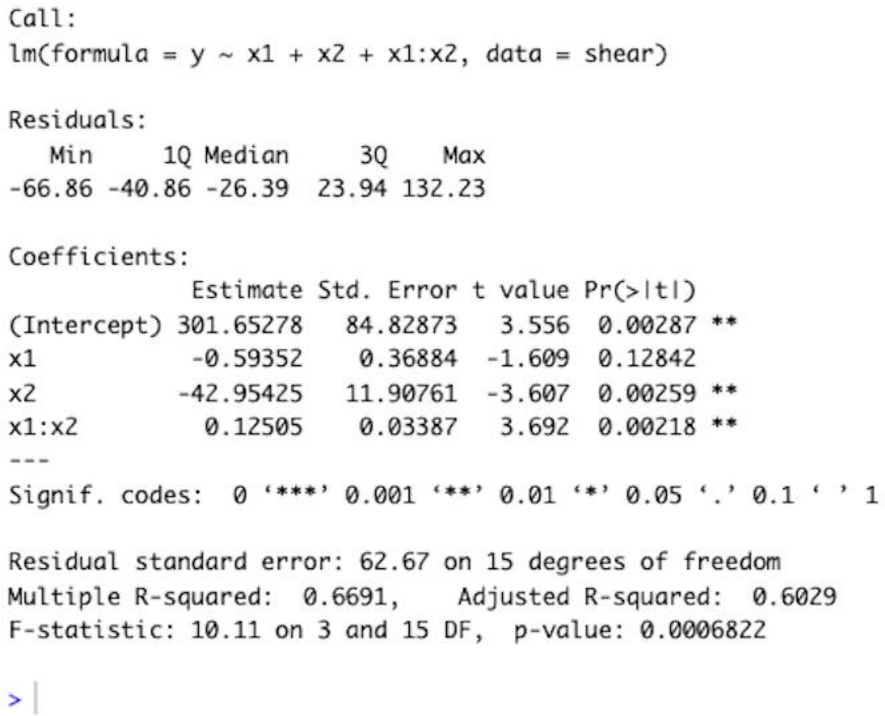
(b).

Code:



```r
# 3. (b).

# fit the linear regression model with an interaction term between x1 and x2
model3b <- lm(y ~ x1 + x2 + x1:x2, data = shear)

# summarize the model
summary(model3b)
```

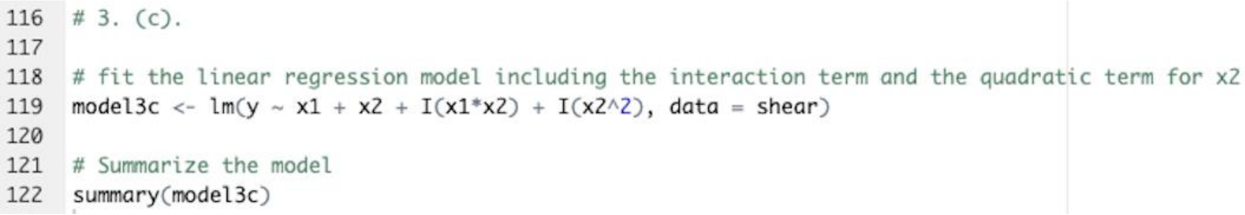
Output:



```
Call:
lm(formula = y ~ x1 + x2 + x1:x2, data = shear)

Residuals:
   Min     1Q Median     3Q    Max 
-66.86 -40.86 -26.39  23.94 132.23 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 301.65278   84.82873   3.556  0.00287 **
x1           -0.59352    0.36884  -1.609  0.12842   
x2          -42.95425   11.90761  -3.607  0.00259 **
x1:x2         0.12505    0.03387   3.692  0.00218 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 62.67 on 15 degrees of freedom
Multiple R-squared:  0.6691,	Adjusted R-squared:  0.6029 
F-statistic: 10.11 on 3 and 15 DF,  p-value: 0.0006822

> 
```

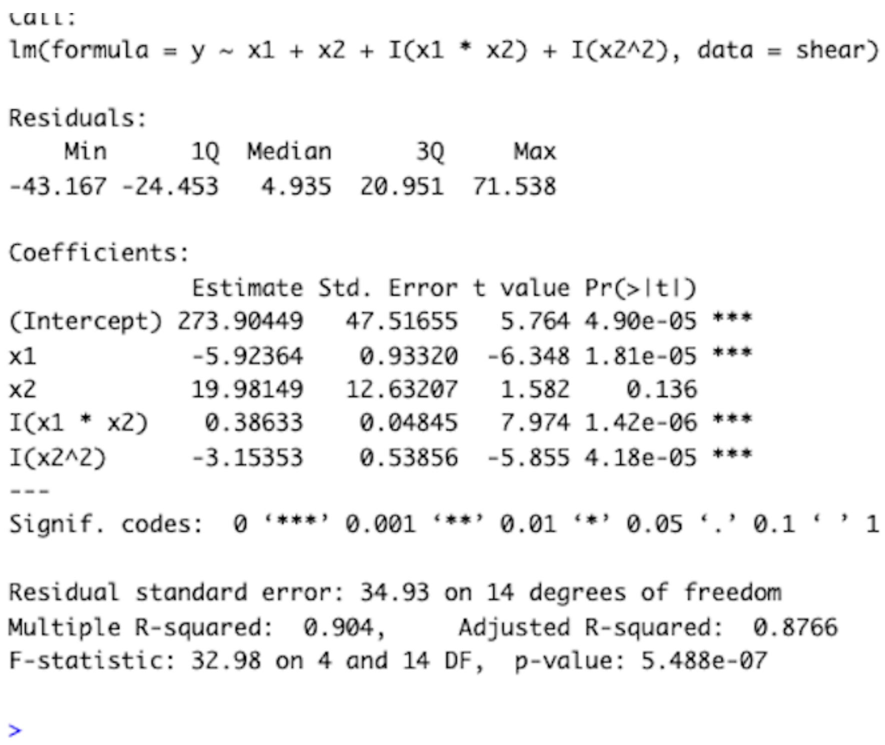
(c).

Code:



```
# 3. (c).

# fit the linear regression model including the interaction term and the quadratic term for x2
model3c <- lm(y ~ x1 + x2 + I(x1*x2) + I(x2^2), data = shear)

# Summarize the model
summary(model3c)
```

Output:



```
Call:
lm(formula = y ~ x1 + x2 + I(x1 * x2) + I(x2^2), data = shear)

Residuals:
    Min      1Q  Median      3Q     Max 
-43.167 -24.453   4.935  20.951  71.538 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 273.90449   47.51655   5.764 4.90e-05 ***
x1           -5.92364    0.93320  -6.348 1.81e-05 ***
x2           19.98149   12.63207   1.582    0.136    
I(x1 * x2)    0.38633    0.04845   7.974 1.42e-06 ***
I(x2^2)      -3.15353    0.53856  -5.855 4.18e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.93 on 14 degrees of freedom
Multiple R-squared:  0.904,	Adjusted R-squared:  0.8766 
F-statistic: 32.98 on 4 and 14 DF,  p-value: 5.488e-07

> 
```

(d).

- Model 1: included only the main effects of x1 and x2, provided a relatively low explanation of the variance in Vmax (*R*^2=0.368) and had marginally significant predictors.
- Model 2: introduced an interaction term between x1 and x2, resulting in a significantly improved fit (*R*^2=0.669), indicating a better but still partial capture of the complexity in the data.
- Model 3: enhanced the model by including both the interaction term and a quadratic term for x2, yielding a substantial increase in explanatory power (*R*^2=0.904) and demonstrating statistically significant predictors, including the interaction and quadratic terms.

The justification for selecting Model 3 as the best model is its superior explanatory power and the statistical significance of its predictors. It captures the nonlinear relationship and the interaction effects between the predictors and the response variable more effectively than the simpler models.