Probability and Statistics Assignment 4

Multiple Linear Regression Model

By:
Abuzar Mehdi (23i-0125)
Adullah Mansoor (23i-0131)

Introduction:

The main goal of this assignment was to analyze how the dependent variable weight is influenced by the independent variables **frequency of nut intake per week, travel time per day, frequency of fruit intake per week, CGPA, average weight of siblings**, **average hours slept daily, number of cups of tea consumed per week** and **waist size.** The data set was collected from the students of FAST NUCES Islamabad and statistical techniques were implemented on the data such as summary statistics, visualization and Multiple Linear Regression Modeling.

Literature review:

- **Nuts intake:** Nuts are nutrient-rich foods and they are linked to weight management such that the high consumption of nuts intake leads to higher weight due to the higher caloric intake. *A review in BMJ Nutrition, Prevention & Health analyzed large cohorts to understand how nut consumption affects weight change. The study found that increased nut intake is associated with better weight control and even prevention of long-term weight gain. This could be due to nuts' nutrient-dense profile, their impact on satiety, and their role in regulating energy balance*
- **Travel Time:** The traveling time affects the energy levels causing decreased physical activity, which leads to lowering weight. *Studies have consistently shown that increased physical activity correlates with better weight management. This*

aligns with general findings that maintaining a balance between energy intake (from diet) and expenditure (through activity) is crucial for healthy weight.

- **Waist Size:** Waist circumference has a strong relationship with body weight so it can be used to predict body weight more accurately such that larger waist sizes are often associated with higher body mass. *A study published in PLOS ONE compared waist circumference, BMI, and waist-to-height ratios, finding that waist circumference was highly correlated with body weight and cardiovascular risks. This measure offers additional insight into weight distribution and health risks compared to BMI alone*
- **Frequency of Fruit Intake per Week**

  Fruit intake is a critical component of a balanced diet and can significantly influence weight. Fruits are rich in fiber and water content, which contribute to satiety and reduced overall caloric intake. *A study published in Nutrition Research found that higher fruit consumption was associated with lower body weight and a reduced risk of obesity. The fiber in fruits slows digestion, helping regulate energy intake and promoting weight control.*

- **CGPA**

  Academic performance, often reflected in CGPA, can indirectly relate to weight through lifestyle factors. *High-achieving students may experience stress and irregular eating habits, which can affect weight. A study in Nutritional Neuroscience revealed a link between academic stress and dietary choices, with stressed students tending to consume energy-dense foods, leading to potential weight changes. Conversely, disciplined lifestyles associated with higher CGPAs might promote healthier eating and regular exercise.*

- **Average Weight of Siblings**

  Family weight trends can reflect shared genetics and environmental factors such as diet and activity patterns*. A study in the International Journal of Obesity highlighted that sibling weight was a strong predictor of individual weight,*

emphasizing the role of shared familial factors. Genetic predisposition and shared behaviors, such as dietary habits, significantly influence body weight within families.

- **Average Hours Slept Daily**

  Sleep plays a pivotal role in regulating metabolism and weight. Insufficient sleep has been li*nked to increased hunger hormones and reduced satiety, leading to overeating. Research in the Journal of Clinical Sleep Medicine demonstrated that individuals sleeping fewer than six hours a night had a higher risk of obesity compared to those with sufficient sleep. This highlights the importance of sleep duration in maintaining a healthy weight.*

- **Number of Cups of Tea Consumed per Week**

  Tea, particularly green tea, has been associated with weight management due to its thermogenic and fat oxidation properties*. A meta-analysis published in The American Journal of Clinical Nutrition found that regular tea consumption, particularly catechin-rich green tea, was linked to modest reductions in body weight and waist circumference. This effect is attributed to compounds that boost metabolism and reduce fat storage.*

These variables were chosen for their strong relationships with weight based on the studies.

Data Collection (Task 1):

The data was collected from the students of FAST university Islamabad, then the process of data cleaning was done which included the removal of unnecessary columns such as roll numbers and phone numbers so that only necessary variables could be studied. The variables studied were:

- **Weight:** measured in kg (dependent variable)
- **Nut intake :** Number of nuts in a day
- **Travel Time:** traveling hours in a day
- **Waist size:** measured in inches

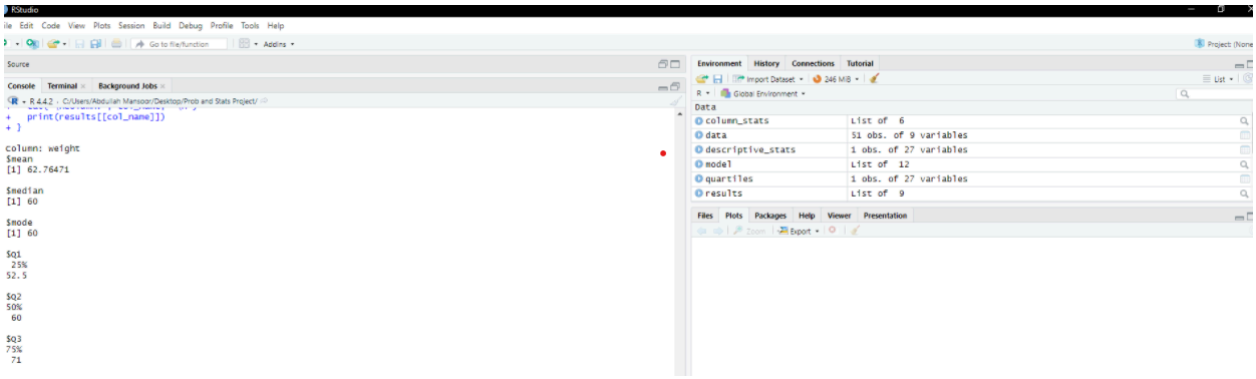
- **Average weight of siblings:** Sum of weight of each sibling / number of siblings
- **Travel time:** Average time spent traveling daily (in hours)
- **Sleep:** Average time spent sleeping per day
- **Fruit Intake:** Number of fruits eaten per week
- **CGPA:** CGPA
- **Tea intake:** Number of cups of tea consumed per week

The missing data values were replaced with the average of that variable.

Summary Statistics (Task 2):

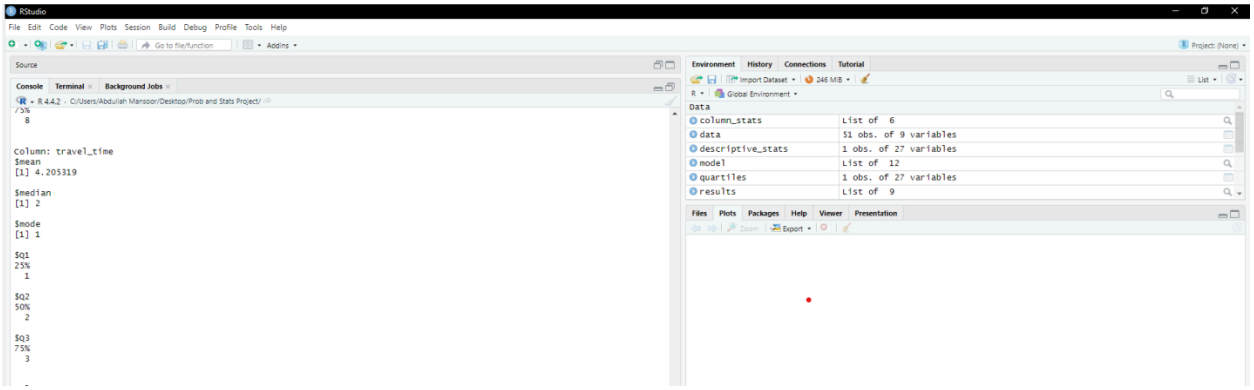
Weight:

- **Mean :** 62.76471
- **Median :** 60
- **Mode :** 60
- **Q1:** 25% (52.5)
- **Q2:** 50% (60)
- **Q3:** 75% (71)



Travel Time:

- **Mean :** 4.205110
- **Median :** 2
- **Mode :** 1
- **Q1:** 25% (1)
- **Q2:** 50% (2)
- **Q3:** 75% (3)

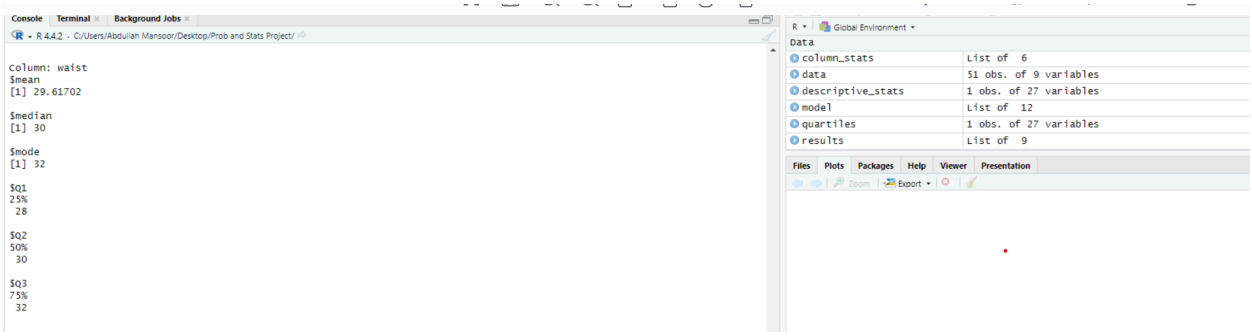


Nut Intake:

- **Mean :** 7.767647
- **Median :** 7.767647
- **Mode :** 7.767647
- **Q1:** 25% (3.5)
- **Q2:** 50% (7.767647)
- **Q3:** 75% (7.767647)

Waist Size:

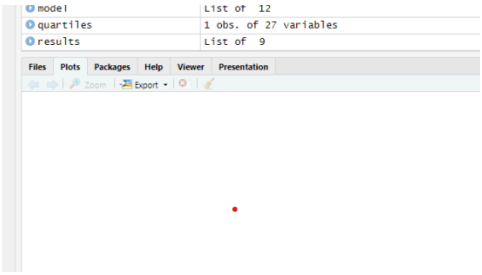
- **Mean :** 29.61702
- **Median :** 30
- **Mode :** 32
- **Q1:** 25% (28)
- **Q2:** 50% (30)
- **Q3:** 75% (32)


Console
Terminal
Background Jobs
R 4.4.2 · C:/Users/Abdullah Mansoor/Desktop/Prob and Stats Project/
Column: waist
$mean
[1] 29.61702
$median
[1] 30
$mode
[1] 32
$Q1
25%
28
$Q2
50%
30
$Q3
75%
32
Global Environment
Data
column_stats List of 6
data 51 obs. of 9 variables
descriptive_stats 1 obs. of 27 variables
model List of 12
quartiles 1 obs. of 27 variables
results List of 9
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export

Some of other variables were also observed:

```
Column: cgpa
$mean
[1] 3.075278

$median
[1] 3.075278

$mode
[1] 3.075278

$Q1
  25%
2.775

$Q2
     50%
3.075278

$Q3
     75%
3.077639
```

```
Column: fruit_intake
$mean
[1] 6.569231

$median
[1] 6.569231

$mode
[1] 6.569231

$Q1
25%
  3

$Q2
     50%
6.569231

$Q3
75%
  7
```




```
Column: avg_weight_of_sibings
$mean
[1] 58.38587

$median
[1] 58.38587

$mode
[1] 58.38587

$Q1
25%
 50

$Q2
     50%
58.38587

$Q3
    75%
68.25
```

Interpretation:

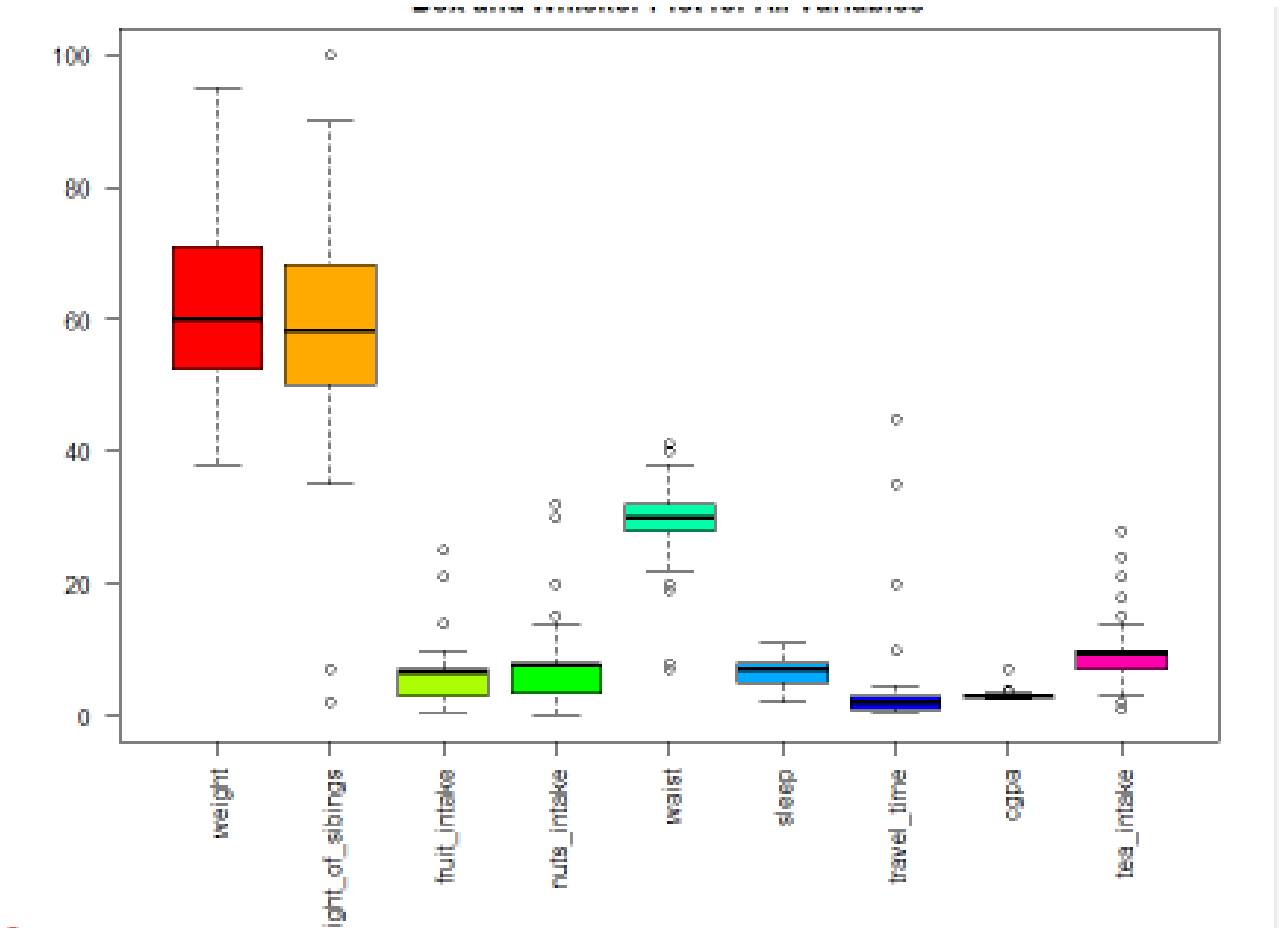
- **Weight** : The weight distribution shows that the majority of students weigh close to the median (60 kg), with a moderate spread of values as indicated by the interquartile range (71 - 52.5 = 18.5). The alignment of mean and median suggests a fairly symmetrical weight distribution.
- **Nuts intake**: The data shows clustering around a single value (7.77), suggesting a high concentration of nut intake within the student group. The lack of variability (equal Q2 and Q3) indicates that many students consume a similar quantity of nuts.

- **Traveling Hours**: The travel time is skewed towards shorter durations, with a high concentration of students traveling minimal distances. The relatively low median compared to the mean indicates the presence of outliers with longer travel times.
- **Waist Size** : Waist sizes are fairly symmetrically distributed, with the mean and median close in value. The interquartile range (32 - 28 = 4) indicates relatively low variation in waist size among the student group.

General Observations

- **Weight and Waist Size Relationship:** Waist size shows a reasonable range with minimal variation, which is consistent with its role as a physical measure that correlates with weight.
- **Travel Time and Weight:** The skewed travel time distribution might influence the energy expenditure and, consequently, weight in the population. Students with longer travel times might walk more, influencing their weight.
- **Nut Intake and Weight:** The clustering in nut intake data suggests a common dietary pattern among students, potentially influencing weight positively or negatively, depending on nut type and portion size.

Box and Whisker Plots:



1. Weight

- **Interpretation:** The box for weight is fairly symmetric, suggesting a normal distribution around the median. The variability is moderate, with most values concentrated within the interquartile range (IQR). This aligns with the descriptive statistics, where the mean (62.76) and median (60) were close, indicating a fairly symmetric distribution.
- **Outliers:** As no outlier is specifically marked, the boxplot suggests no significant anomalies in weight data.

2. Travel Time

- **Interpretation:** The travel time box shows a skew toward the lower range, with most values clustering around 1-3 units. This indicates that many students have shorter

commute times. The right side is elongated, suggesting a few students experience longer travel times.

- **Outliers:** The presence of outliers in the higher travel time range suggests some students have significantly longer travel durations compared to the majority.

3. Nut Intake

- **Interpretation:** Nut intake shows a very tight interquartile range, suggesting little variability in how much students consume. This corresponds with the descriptive statistics, where the median and Q1-Q3 quartiles are very close, indicating a consistent dietary habit among most students.
- **Outliers:** While there are some outliers, these indicate a small number of students who consume much more or less than the majority.

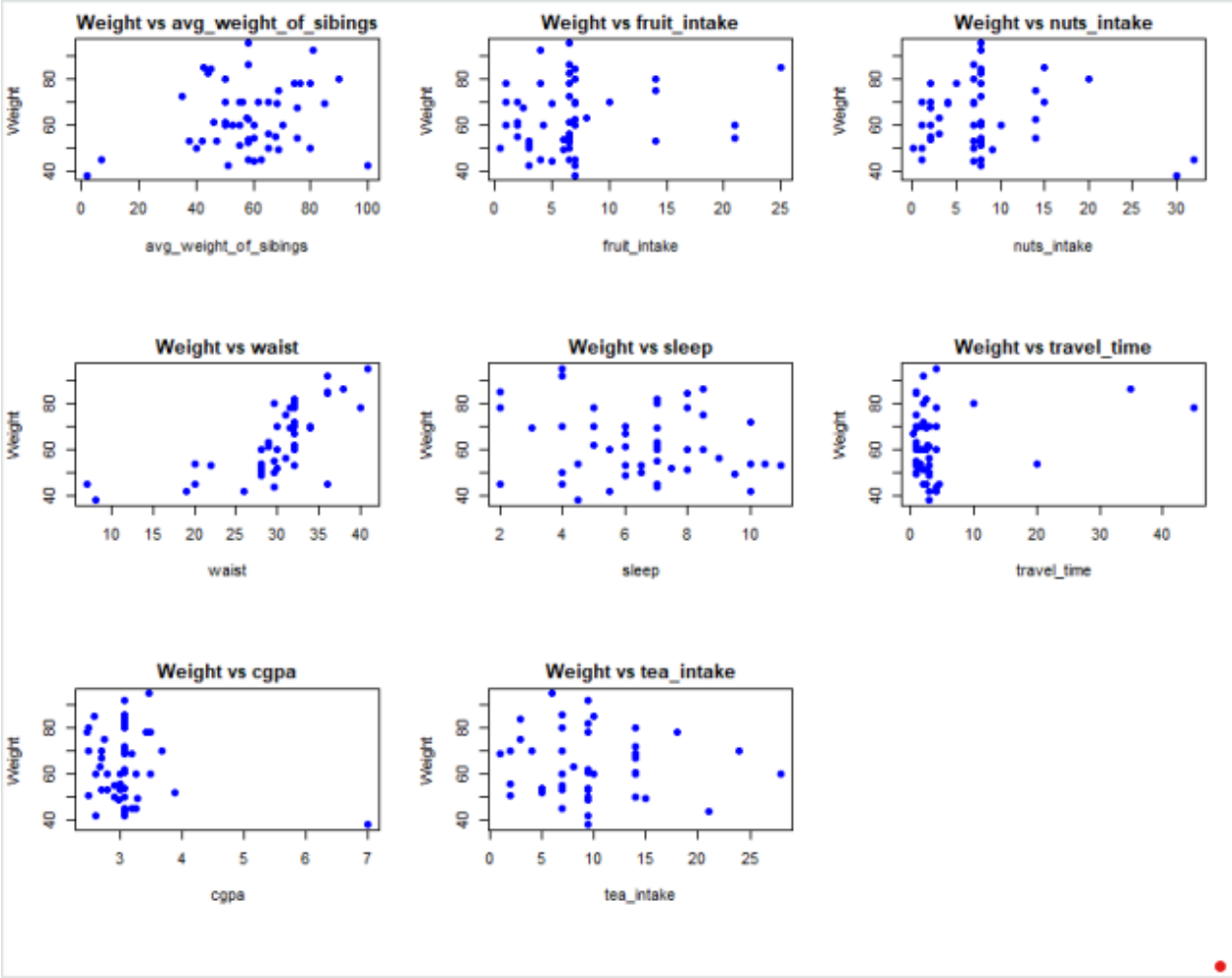
4. Waist Size

- **Interpretation:** The waist size distribution is relatively tight, with a slight tendency towards larger waist sizes, reflecting the overall physical characteristics of the student population. The boxplot supports the descriptive statistics showing a moderate spread in waist size with the median at 30.
- **Outliers:** No significant outliers are observed, further supporting the consistency in waist size across the dataset.

5. Other Variables

- **Interpretation:** Other variables (e.g., sleep, tea intake) exhibit wider distributions with multiple outliers, indicating greater variability in these aspects across students.

Scatter Plot (Task 4):



Scatter plots between weight and each independent variable reveal the following:

- **Weight vs. Travel Time:** A weak negative trend indicates that longer travel times might slightly reduce weight due to increased physical activity.
- **Weight vs. Nut Intake:** A weak positive trend suggests that higher nut intake correlates with a slightly increased weight, possibly due to increased caloric intake.
- **Weight vs. Waist Size:** A strong positive correlation confirms the direct relationship between waist circumference and weight.

The scatter plot of the other variables have either very weak or no relation with the weight.

MLRM using R (Task 5):

The summary of the MLRM with all the dependent variables is given:

```
Call:
lm(formula = weight ~ tea_intake + cgpa + sleep + fruit_intake +
    avg_weight_of_siblings + nuts_intake + travel_time + waist,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-25.320  -5.103   1.347   4.926  16.560

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             0.79294   15.21075   0.052  0.95867
tea_intake             -0.06073    0.25869  -0.235  0.81554
cgpa                   -1.33881    2.41106  -0.555  0.58165
sleep                  -0.56943    0.62654  -0.909  0.36862
fruit_intake           -0.08967    0.30013  -0.299  0.76660
avg_weight_of_siblings -0.01868    0.08779  -0.213  0.83255
nuts_intake             1.05572    0.31899   3.310  0.00192 **
travel_time             0.33201    0.16312   2.035  0.04815 *
waist                   2.10863    0.27324   7.717  1.4e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.978 on 42 degrees of freedom
Multiple R-squared:  0.6624,    Adjusted R-squared:  0.5981
F-statistic:  10.3 on 8 and 42 DF,  p-value: 7.86e-08
```

Significant predictors: nuts_intake, travel_time, and **waist.**

Interpretation: These variables have a statistically significant relationship with weight, while others may not have a meaningful impact.

Model performance: The model explains **66.24%** of the variance in weight, with a reasonably low residual error. However, the adjusted R^2 suggests some predictors may be irrelevant.

Due to there being a lot of missing values replaced by the average in the columns and the variables frequency of fruit intake per week, CGPA, average weight of siblings, average hours slept daily and number of cups of tea consumed per week, not being significant,

we will make an MLRM with only waist, nuts_intake and travel_time as the main predictors.

```
Residuals:
    Min      1Q  Median      3Q     Max
-23.639  -4.929   1.311   5.837  16.399

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.0847     8.3606  -1.206   0.2338
nuts_intake   1.0548     0.2373   4.446 5.33e-05 ***
travel_time   0.3528     0.1539   2.292   0.0264 *
waist         2.1330     0.2386   8.940 1.05e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.608 on 47 degrees of freedom
Multiple R-squared:  0.6527,    Adjusted R-squared:  0.6306
F-statistic: 29.45 on 3 and 47 DF,  p-value: 7.281e-11
```

Model Fit: The model explains **65.27%** of the variance in weight, which is fairly good.

Significant Variables:

nuts_intake: Strong positive effect on weight.

travel_time: Moderate positive effect on weight.

waist: Strong positive effect on weight (most influential predictor).