**STAT448: Big Data**

**Assignment 1**

**Due Date: 26th July 2023**

**STAT448 - Assignment 1**

**Yuanyuan Qin (43865058) Guolei Li (28496187)**

1. 1. (15 marks) Three observations for a response random variable Y are {6, 12, 18}; the corresponding values observed for the explanatory variable X are {4, 5, 6}. Assume a linear model: Y = β0 + β1X + ε

(a) Compute ordinary least squares estimates of the coefficients β0 and β1 using linear algebra calculations by hand and with explanatory comments. (5 marks).

(b) Calculate by hand the estimates of the residuals ˆ. (4 marks).

(c) Perform the same matrix algebra calculations of parts (a) and (b) using R. (4 marks).

(d) Estimate the coefficients using the function lm in R. (2 marks).

***Answer:***

1. Computer OLS by hand

**Step 1: Set up the response variable matrix Y:**

Y = [[6],

[12],

[18]]

**Step 2: Set up the design matrix X with an additional column of ones for the intercept:**

X = [[1, 4],

[1, 5],

[1, 6]]

**Step 3: Calculate Xᵀ \* X:**

Xᵀ \* X = [[1, 1, 1], [[1, 4], [[3, 15],

[4, 5, 6]] \* [1, 5], = [15, 77]]

**Step 4: Calculate the inverse of (Xᵀ \* X):**

(Xᵀ \* X)⁻¹ = (1/6) \* [[77, -15],

[-15, 3]]

**Step 5: Calculate Xᵀ \* Y:**

Xᵀ \* Y = [[1, 1, 1], [[6],

[4, 5, 6]] \* [12],

[18]]

Xᵀ \* Y = [[(1 \* 6) + (1 \* 12) + (1 \* 18)],

[(4 \* 6) + (5 \* 12) + (6 \* 18)]]

Xᵀ \* Y = [[36],

[192]]

Step 6: Calculate β by multiplying (Xᵀ \* X)⁻¹ and Xᵀ \* Y:

β = (1/6) \* [[77, -15],

[-15, 3]] \* [[36],

[192]]

Performing the matrix multiplication:

β = (1/6) \* [[(77 \* 36) + (-15 \* 192)],

[(-15 \* 36) + (3 \* 192)]]

β = (1/6) \* [[2772 - 2880],

[-540 + 576]]

β = (1/6) \* [[-108],

[36]]

β = [[-18],

[6]]

So, the correct result for β is [[-18], [6]].

The OLS estimates for the coefficients are β₀: -18 and β₁:6.

These values represent the intercept and slope of the linear model.

1. **Step 1: Calculate the predicted values Ŷ for each observation:**

Ŷ₁ = -18 + 6 \* 4 = 6

Ŷ₂ = -18 + 6 \* 5 = 12

Ŷ₃ = -18 + 6 \* 6 = 18

**Step 2: Calculate the residuals (e) for each observation:**

Residual₁ = Y₁ - Ŷ₁ = 6 - 6 = 0

Residual₂ = Y₂ - Ŷ₂ = 12 - 12 = 0

Residual₃ = Y₃ - Ŷ₃ = 18 - 18 = 0

The residuals for all observations are

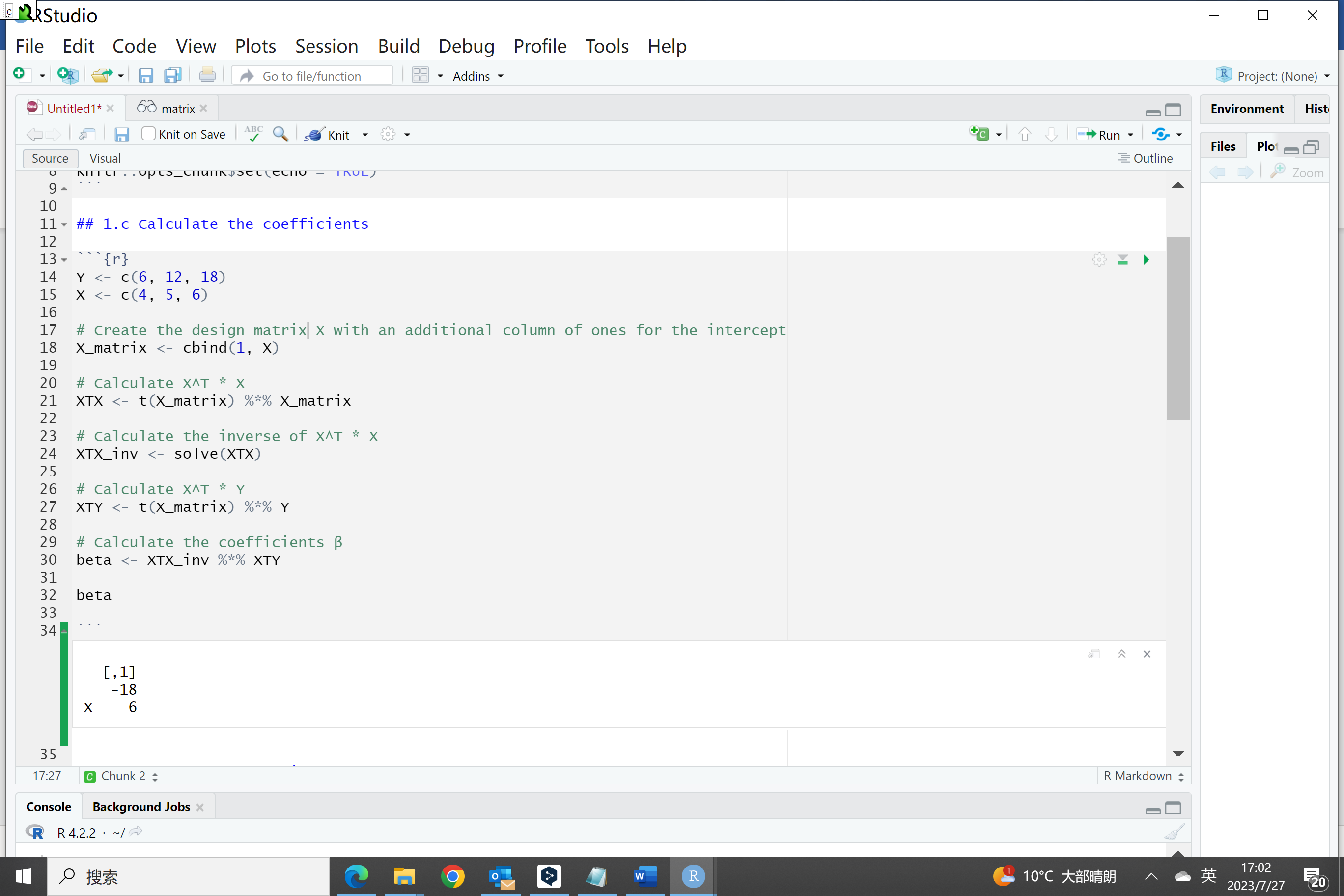
Residual = [[0],

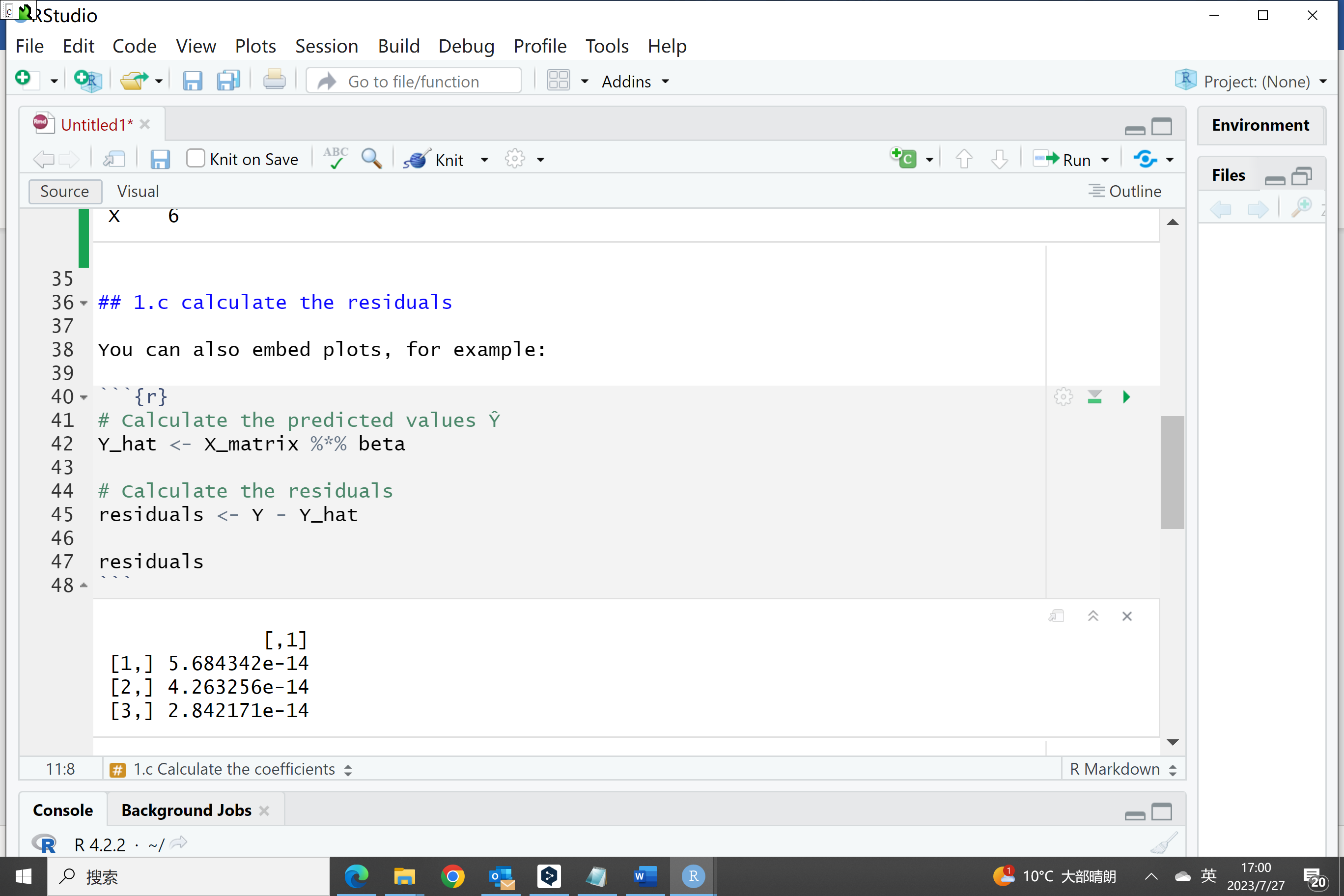
[0],

[0]]

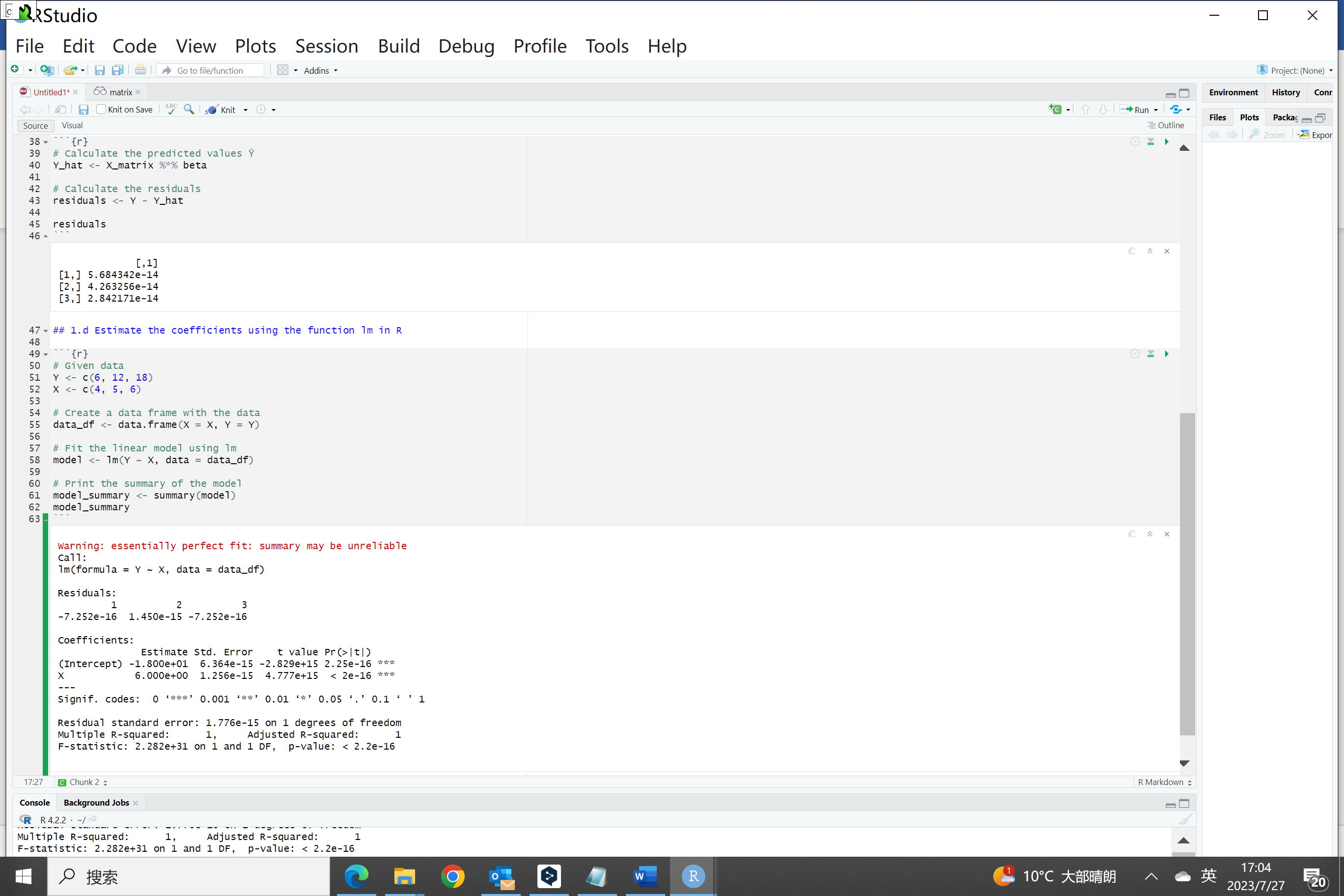
The result means that the observed values exactly match the predicted values in this linear model. There is no error, and the model perfectly fits the data.

1. Calculate the coefficients and residuals in R





1. Use lm function in R to calculate the model



The lm function provides the same results of coefficients (β₀: -18 and β₁:6) and residuals (near to zero). Since the simple linear model fit the data perfectly so a warning shows up as a reminder.

1. (15 marks) In the context of question 1, consider the case where the values observed for the explanatory variable X are {2, 2, 2}. (a) What happens to the coefficient estimates? (5 marks). (b) Using appropriate terminology gives both a statistical and geometric explanation of this situation. (10 marks).

***Answer:***

(a)

If all the values observed for the explanatory variable X are the same (e.g., {2, 2, 2}), it means that there is no variation in the X variable. In this situation, the coefficient estimates in the linear regression model will become problematic because the determinant of (Xᵀ \* X) could be zero. The inverse of Xᵀ \* X does not exist, and the coefficient estimates cannot be calculated using the method.

Xᵀ \* X = [[3, 6], [6, 12]]

det (Xᵀ \* X) = (3 \* 12) - (6 \* 6) = 36 - 36 = 0

(b)

Statistical Explanation:

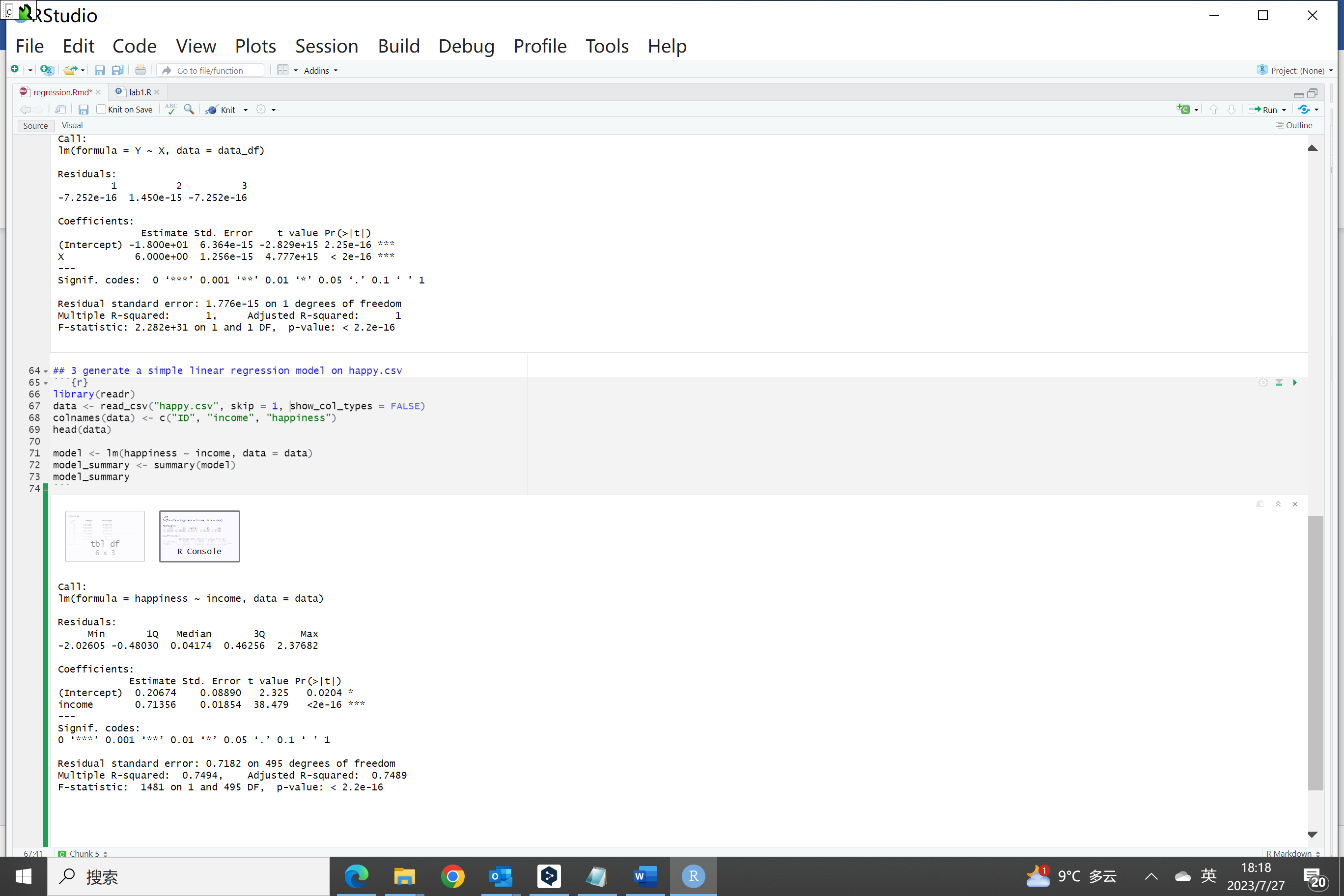
This situation is known as multicollinearity, where one or more explanatory variables in the model are highly correlated or perfectly correlated. In this case, all the X values are the same, resulting in a situation where there is no variation in the X variable. As a result, the model cannot distinguish the individual effects of the explanatory variable X on the response variable Y. The coefficient estimates become unreliable, as the model fails to determine the unique contribution of X to the predictions.

Geometric Explanation:

In simple linear regression, the model tries to fit a line that minimizes the sum of squared residuals between the predicted and observed Y values. When all the X values are the same, it means that all data points lie on a single vertical line and there is no spread along the X-axis. Consequently, the model cannot find a unique line to fit through these points, as there are infinitely many lines that pass through the same X value. This situation leads to the non-existence of the inverse of Xᵀ \* X and, therefore, the inability to calculate reliable coefficient estimates.

1. (30 marks) Using the data in the provided CSV file (happy.csv), generate a simple linear regression model to describe the relationship between Income and Happiness. Then answer the questions below. Note: 1 unit of Income (feature) is $10,000 and Happiness (target) is a scale, of 1-10.
2. Using the model summary, state the regression equation for happiness (3 decimal places for coefficients is sufficient). (4 marks).

***Answer:***



Based on the model summary provided, the regression equation for happiness could be stated: happiness ≈ 0.207 + 0.714 \* income.

β₀: 0.207 and β₁: 0.714

This equation represents the estimated relationship between income and happiness based on the linear regression model.

(b) Using the β1 coefficient from the model equation, provide an interpretation of this coefficient in relation to the response variable. (4 marks).

***Answer:***

The coefficient β₁ in the model equation represents the estimated effect of the explanatory variable (income) on the response variable (happiness). Specifically, in this model, the coefficient β₁ is approximately 0.714.

For every one-unit increase in income, the predicted happiness score (happiness) is expected to increase by approximately 0.714 units. This means that as an individual’s income increases, their predicted happiness level is estimated to increase by about 0.714 on average.

1. Using the model summary, does income have a significant effect on happiness? Justify your answer. (5 marks).

***Answer:***

Yes, income has a significant effect on happiness based on the model summary. In the summary output, the p-value associated with the coefficient for "income" is reported as "<2e-16," which means the p-value is effectively zero. In statistical hypothesis testing, a p-value below a significance level (commonly set at 0.05) indicates that the effect of the variable is statistically significant.

The coefficient for "income" is also positive (0.714), suggesting that higher income levels are associated with higher predicted happiness scores on average.

(d)Again using the model summary, does your model provide a good fit for the observed data? Justify your answer. (5 marks).

***Answer:***

Yes. From the Adjusted R square, F score, and residuals, I think this model provides a good fit for the observed data.

Adjusted R-squared: The Adjusted R-squared takes into account the number of predictors in the model and adjusts the R-squared accordingly. In this case, the Adjusted R-squared is reported as 0.7489, which means approximately 74.89% of the variance in happiness can be explained by the variation in income in the model.

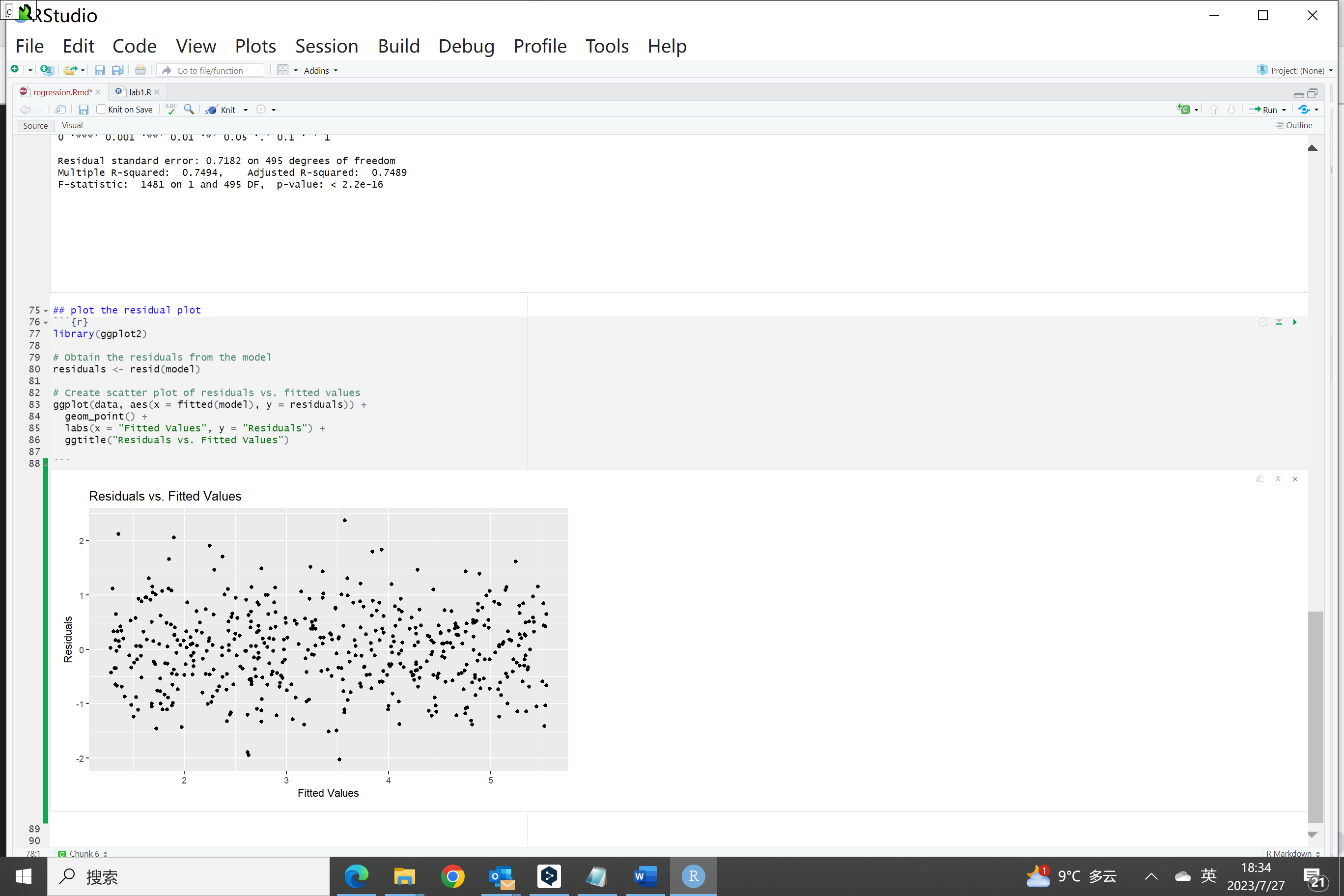
F-statistic: The F-statistic is a test statistic that assesses the overall significance of the model. It tests whether the model as a whole is a significant improvement over the intercept-only model (null model). In this case, the F-statistic is reported as 1481 with an extremely small p-value ("< 2.2e-16"), indicating that the model is highly significant.

Residual standard error: The residual standard error measures the average deviation of the observed data points from the fitted values. It provides a measure of the model's accuracy in predicting the response variable. In this case, the residual standard error is reported as 0.7182, which shows the model could fit the observed data well.

1. Validate your regression model using appropriate residual plots. What do you observe? Is the fit adequate? Do the residuals suggest a better fit is possible? (5 marks).

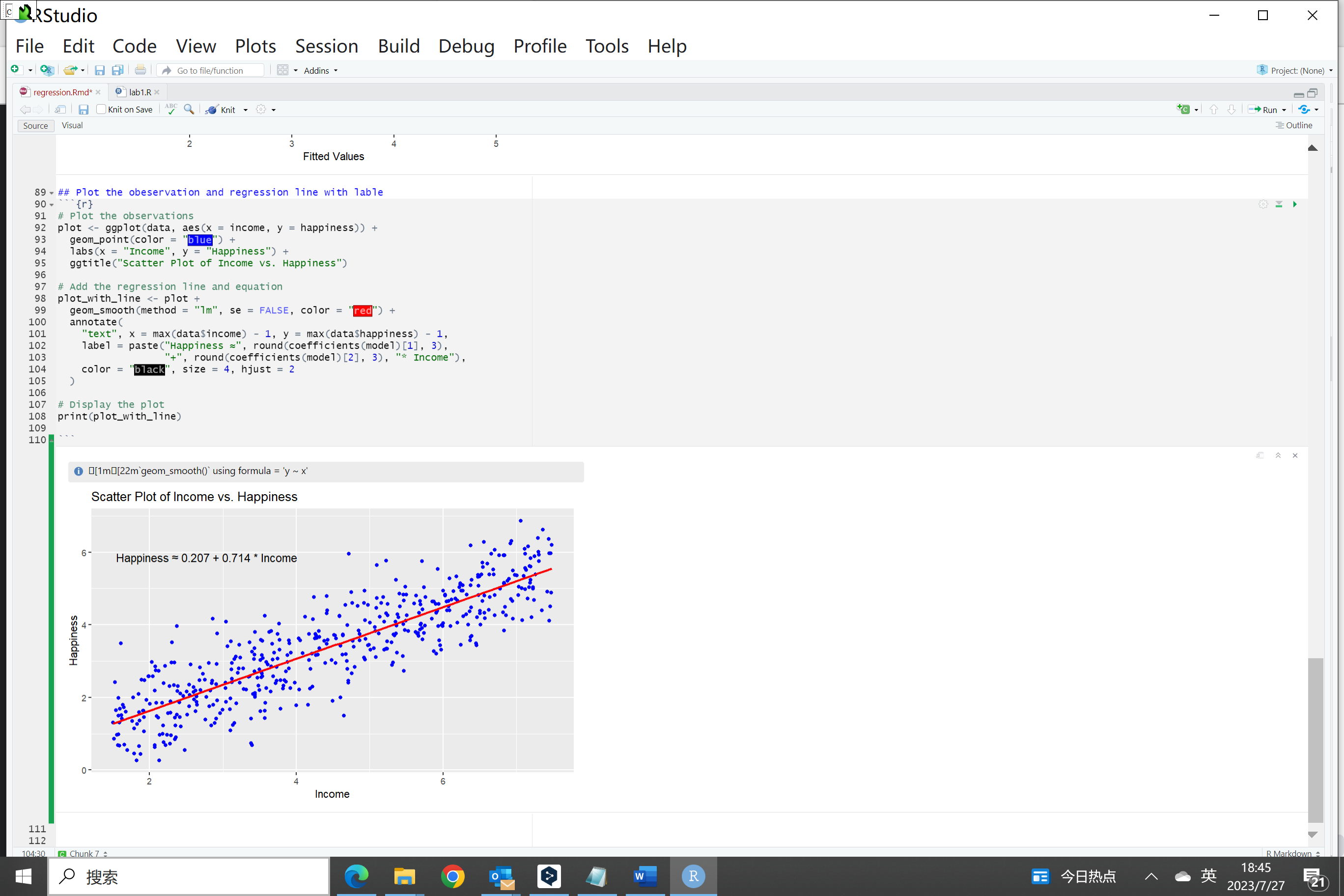
***Answer:***

This is the R code to plot the residuals against the fitted value.



In this scenario, if the residual plot shows random scatter without any systematic patterns, it suggests that the current model is a good fit for the data. The absence of patterns in the residuals indicates that the linear regression model is adequately capturing the variation in the response variable based on the income variable.

(f) Provide a well-labeled plot of the observations, and also include the regression line and regression equation. (4 marks).



1. Using your regression equation, make Happiness predictions for the following income values, 3.65, 6.87, and 8.49 (to 2 decimal places). Why might it not be valid to make predictions outside the income range in your data? (3 marks).

***Answer:***

For Income = 3.65:

Happiness ≈ 0.207 + 0.714 \* 3.65 ≈ 2.81

For Income = 6.87:

Happiness ≈ 0.207 + 0.714 \* 6.87 ≈ 5.10

For Income = 8.49:

Happiness ≈ 0.207 + 0.714 \* 8.49 ≈ 6.27

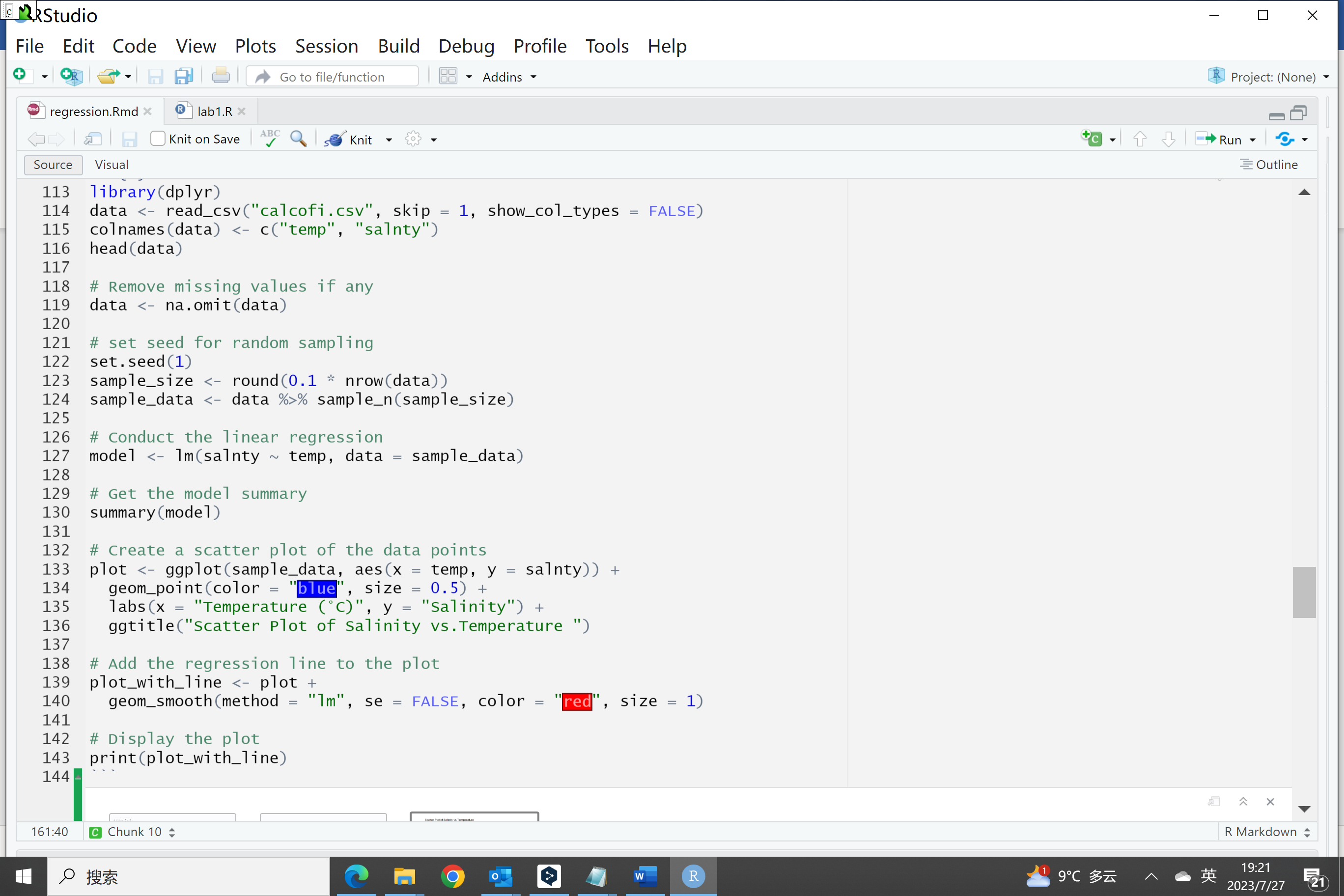
Regression models are best suited for making predictions within the range of the observed data. Predictions outside this range may be less reliable because the model has not seen any data points in those regions. When we do extrapolation (predict data of out the observed range), the assumption is that the relationship between income and happiness continues to hold outside the observed range. However, this assumption may not be valid, as the data might follow a different pattern outside the observed range.

1. (40 marks) In this question we are going to be working with the CalCOFI Dataset (CC BY 4.0 License). This dataset contains comprehensive oceanographic measurement data from California from 1949 up to the present day. The dataset includes features such as water temperature, salinity, measurement depth, O2 level, and more. For our purposes, we will only be using the water temperature (T degC) feature to predict the water salinity (Salnty).

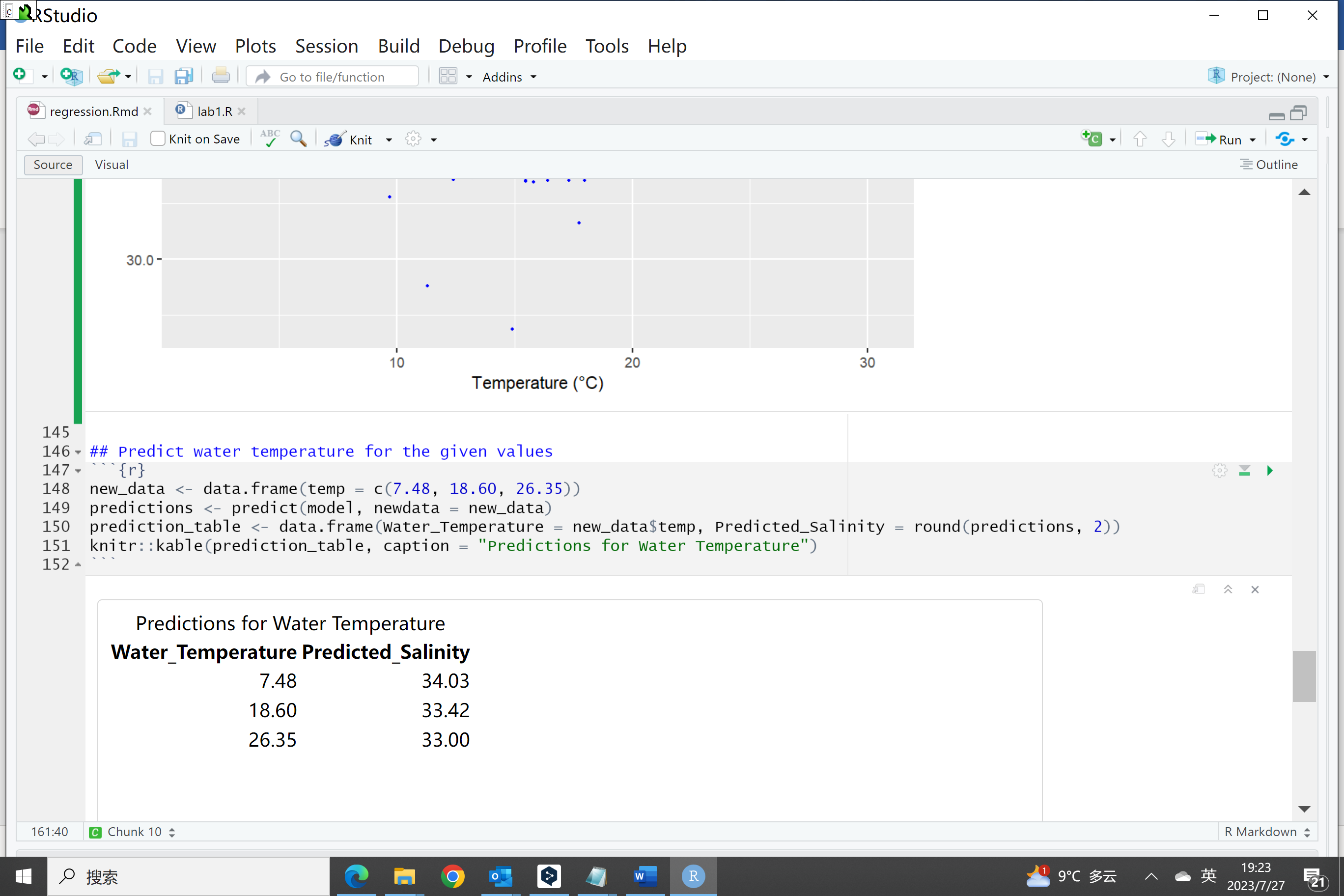
Using the data in the provided CSV file (calcofi.csv), extract a 10% random sample (as the dataset is quite large). Using your random sample, explore the relationship between two variables of interest through simple linear regression. Apply appropriate data preprocessing, model fitting, and evaluation techniques to analyze the relationship and make predictions for the following water temperature values 7.48, 18.60 and 26.35 (to 2 decimal places). Discuss the significance of the regression coefficients, interpret the model’s performance metrics, and draw meaningful conclusions from your analysis.

***Answer:***

Here is the R code to extract 10% random samples, do data preprocessing and simple linear regression.



The prediction on this model for the new datasets are put below



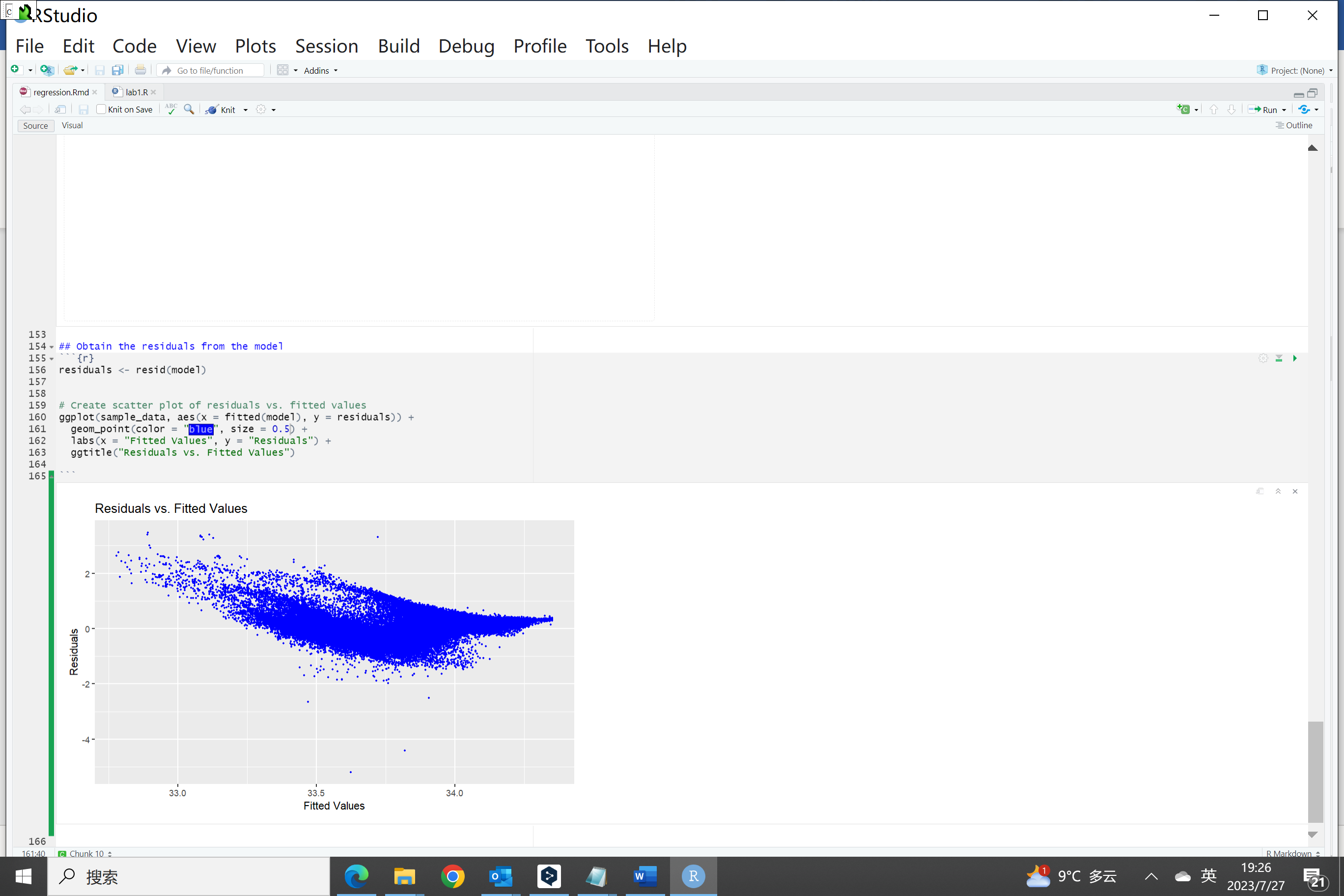
Here are the model summary and residual plot from R



The model summary:



The residual plot of this simple linear regression model:



Both coefficients are highly statistically significant (indicated by <2e-16 in the model summary output), with very low p-values. This means that the linear regression model finds a strong statistical relationship between temperature and salinity, and there is evidence to suggest that the temperature variable is significantly associated with changes in salinity.

The estimated coefficient (β₁) for temperature is -0.0542. It represents the change in the expected salinity for a one-unit increase in temperature (1 degree Celsius). Since the coefficient is negative, it suggests that as the temperature increases, the expected salinity decreases by approximately 0.0542 units. This negative coefficient indicates an inverse relationship between temperature and salinity in the linear model.

The estimated intercept (β0) is 34.4309. However, its interpretation is not very meaningful in this context because the temperature variable (temp) is not centered around zero in the dataset.

Adjusted R-squared: The model's adjusted R-squared value (0.2485) suggests that only about 24.85% of the variability in salinity can be explained by the linear relationship with temperature. This indicates that other factors or variables not included in the model may also play a significant role in determining salinity.

Residual Standard Error: The residual standard error (0.3993) measures the average amount that the observed salinity values deviate from the regression line. It represents the standard deviation of the residuals.

Residual plot: The residual plot has some pattern. The residual displays a curve according to the fitted water salinity values. It suggests that the simple linear regression model may not fully capture the underlying relationship between the dependent variable and the independent variable. In other words, the relationship between temperature and salinity might be more complex than a straight line, which the simple linear regression assumes.

Conclusions: The regression analysis indicates a significant inverse relationship between temperature and salinity. As the temperature increases, the model predicts a decrease in salinity. However, from the Adjusted R-squared, residual plot, and model fitting curve, the current simple regression model's explanatory power is limited. The linear regression model might not fully capture the complexities of the relationship between temperature and salinity. It is essential to consider other factors, explore nonlinear relationships, and conduct further analysis to better understand the factors influencing salinity in the dataset.