Middle East Technical University

Department of Statistics

STAT 363
LINEAR MODELS I

HOMEWORK I

The Analyzing of Simple Linear Regression for Ice Cream Sales

By

Burcu NALBANT 2502235
Fatma Didem MEKE 2361392
Tuğana AKIN 2560993
Sude KONYA 2502193
Berke BAYTAK 2502011

July 2024

STAT363 HOMEWORK 1

Objective

The objective of this homework is to apply simple linear regression analysis on a dataset with two variables, including:

1. Finding the regression equation by checking the assumptions.
2. Calculating R-Squared and correlation coefficients.
3. Preparing ANOVA tables and conducting hypothesis tests.
4. Interpreting the results.

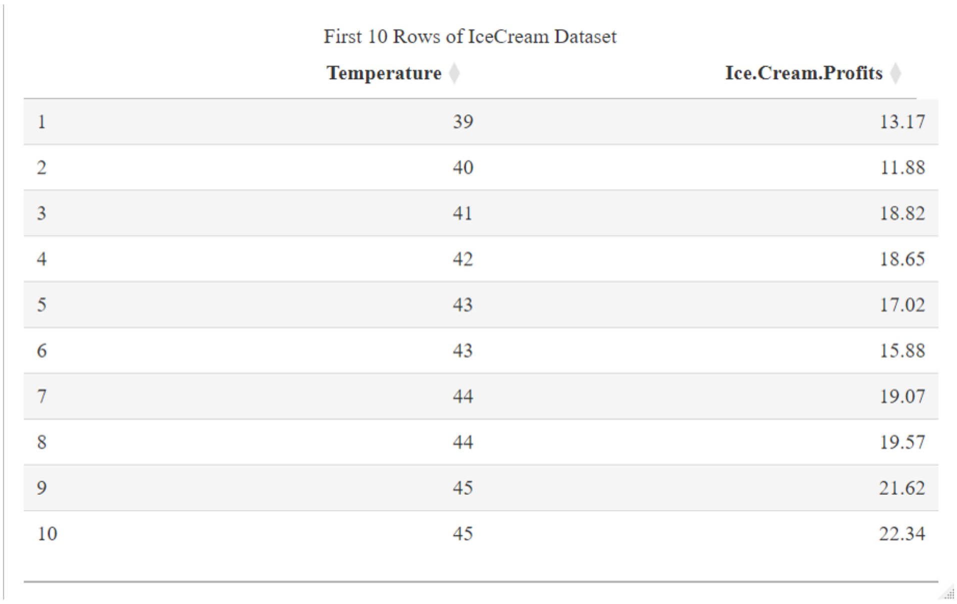
Dataset

For this analysis, we will use the **'Ice Cream Sales and Temperature'** dataset, which includes information about temperature and ice cream profits. We found this dataset here: https://www.kaggle.com/datasets/raphaelmanayon/temperature-and-ice-cream-sales

Moreover, our data set has **2 variables** and **366 entries**. The independent variable is Temperature, and the dependent variable is Ice Cream Profits.

Temperature variable is in Fahrenheit and the ice cream sales are in USD.

Our **first 10 observations** from our dataset are as follows:

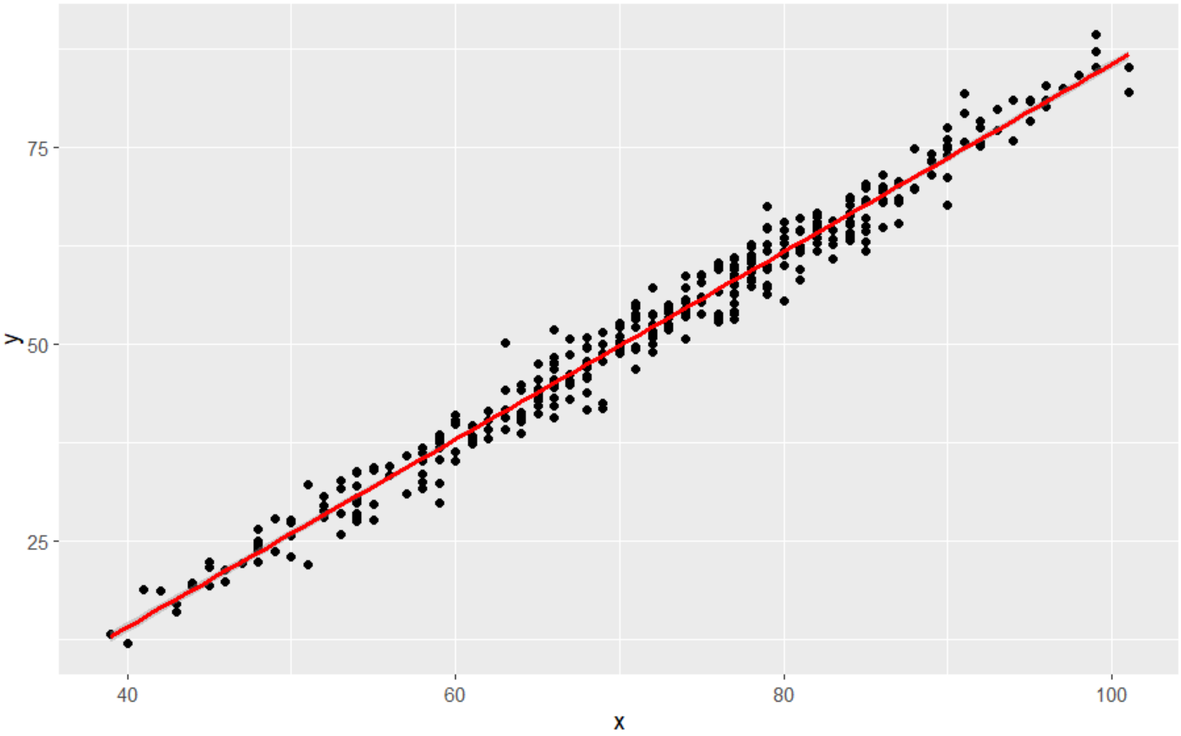


First 10 Rows of IceCream Dataset

	Temperature	Ice.Cream.Profits
1	39	13.17
2	40	11.88
3	41	18.82
4	42	18.65
5	43	17.02
6	43	15.88
7	44	19.07
8	44	19.57
9	45	21.62
10	45	22.34

1. The Regression Equation and Checking Assumptions:

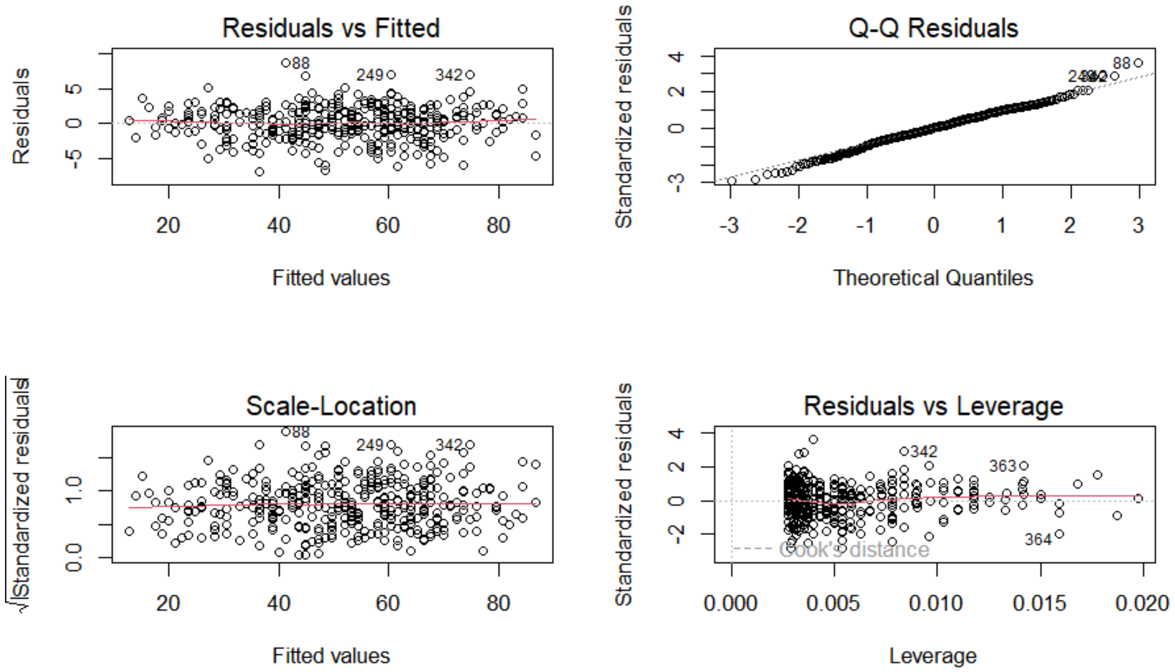
Linearity:



Ice Cream Profits = −33.7+1.19 × Temperature

This graph illustrates the relationship between temperature and ice cream profits. As you can see, as the temperature increases, the ice cream profits also increase. The red line represents the linear regression line, which shows the trend of the relationship. For this reason people are more likely to purchase ice cream when the weather is hot. The regression lines are lines that are drawn on graphs to show which line best fits the data points. This line shows that there is **a strong positive correlation** between temperature and ice cream profits. The higher the temperature, the higher the profits.

This indicates that ice cream profits rise by about $1.19 for every degree that the temperature rises. The intercept, -$33.7, is the expected ice cream profits when the temperature is 0 degrees.



Independence:

The observations should be independent of each other and there should not be any correlation between the residuals In Residuals vs Fitted plot, the points seem randomly scattered and it does not appear that there is a relationship. Thus, it indicates that errors are independent.

In addition, the residuals have a mean close to zero, as you can see in the residuals vs fitted plot, indicating that the model's predictions are unbiased.

Homoscedasticity:

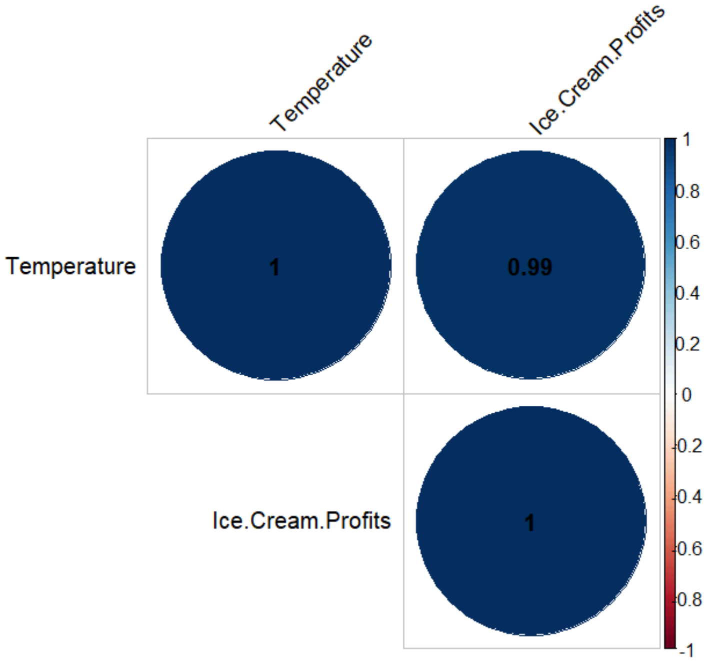
The variance of the response variable should be constant across all levels of the predictor variable. This means that the spread of the response variable should be roughly equal at all temperatures. In the Scale-location plot, you can see a horizontal line with equally (randomly) spread points which indicates equal variances.

Normality:

The residuals should be normally distributed. As you can see in the Q-Q Residuals plot, most of the data points fall close to the line which indicates that errors are normally distributed. This means that the model is a good fit for the data.

To sum up, the regression model gives **a good fit** for the data. The residuals are **independent, homoscedastic, and normally distributed** which suggests that the assumptions of the regression model are met.

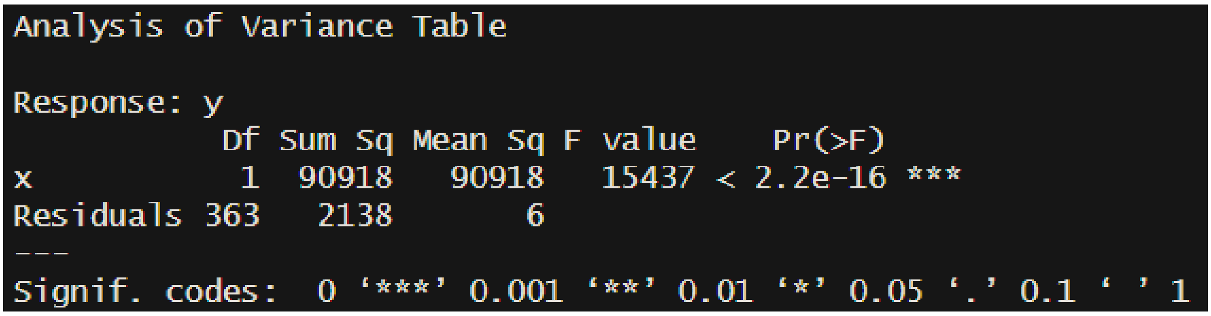
2. R-Squared and Correlation Coefficient:



The value of **R^2 is 0.977**. This indicates that temperature accounts for approximately 97.7% of the variation in ice cream profits. R^2 values range from 0 to 1. 0 indicates that no variation in ice cream sales can be attributed to temperature. 1 indicates the temperature accounts for all of the variability in ice cream sales. Our result of 0.977 indicates **a very high R^2 value**, suggesting a **strong linear association** between ice cream profits and temperature.

Moreover, the table above shows that the **correlation coefficient is 0.988** (nearly 0.99). The correlation coefficient can be between -1 to +1. Perfect negative correlation is denoted by a -1. This means that, in a perfectly linear manner, one variable decreases as the other grows. A value of 0 indicates no association, meaning that there is no linear relationship between the variables. Perfect positive correlation is denoted by +1. This means that, in a perfectly linear manner, one variable increases as the other grows. Our data suggests that there is a very strong positive linear relationship between temperature and ice cream profits, as indicated by our value of 0.988.

3. ANOVA Table and Hypothesis Tests:



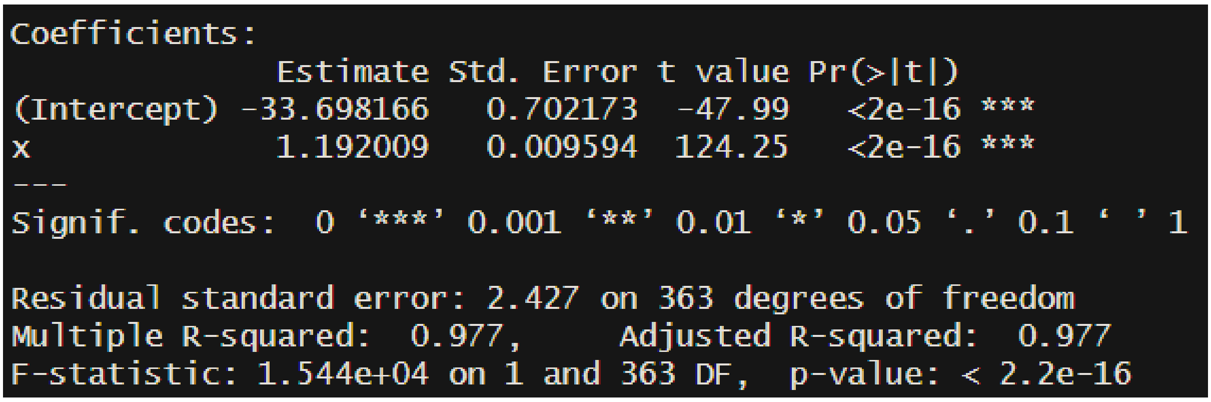
```
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value    Pr(>F)
x           1  90918   90918   15437 < 2.2e-16 ***
Residuals 363   2138       6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Null Hypothesis: The regression model does not explain significantly the variability in Ice cream profits

Alternative Hypothesis: The regression model explains significantly the variability in Ice cream profits

The ANOVA table provides information about the variance explained by the model and the residuals (unexplained variance). The p-value for the temperature predictor is **less than alpha value (0.05)**, which means that we can **reject the null hypothesis**. This indicates that **temperature is a significant predictor of ice cream profits**.



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -33.698166   0.702173  -47.99   <2e-16 ***
x             1.192009   0.009594  124.25   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.427 on 363 degrees of freedom
Multiple R-squared:  0.977,     Adjusted R-squared:  0.977
F-statistic: 1.544e+04 on 1 and 363 DF,  p-value: < 2.2e-16
```

<u>For intercept:</u>

Null Hypothesis: The intercept does not significantly affect ice cream profits

Alternative Hypothesis: The intercept significantly affects ice cream profits

The p-value of the intercept is **less than the alpha value (0.05)**, which means that we can **reject the null hypothesis**.

<u>For Temperature:</u>

Null Hypothesis: Temperature does not significantly affect ice cream profits

Alternative Hypothesis: Temperature significantly affects ice cream profits

The p-value of temperature is **less than the alpha value (0.05)**, which means that we can **reject the null hypothesis**. This also means that **temperature is a significant predictor of ice cream profits.** Each unit of increase of the temperature will increase the ice cream sale 1.19 times.

Furthermore, the spread of the residuals (Residual standard error: 2.427) gives us an idea of the typical prediction error. This tells us that the regression model predicts the ice cream profits with an average error of about 2.427.

4. Interpreting and The Results:

The ice cream profits and the temperature are both having a **strong positive linear relation** based on the regression model. Profits that are made from ice cream sales are remarkably affected by temperature. The **R^2**, which is equal to **0.977**, shows the model's explanation of the significant amount of the variance in the profits of ice cream profits. Findings show the fact that the **temperature is a solid indicator of ice cream sales**, and the model that we created for the profits of ice cream sales based on temperature is **quite well-founded**.

Codes Used:

The codes used for this analysis are provided, including data loading, library loading, data visualization, and regression analysis.

```
datam<-read.csv('C:/Users/ASUS/Documents/IceCream.csv',header=TRUE, sep=',')

# Load necessary libraries
library(DT)
library(ggplot2)
library(corrplot)

# Load the dataset
datam <- read.csv('C:/Users/ASUS/Documents/IceCream.csv', header = TRUE, sep = ',')

# Display the first 10 rows of the dataset in a readable format
datatable(head(datam, 10), options = list(pageLength = 10), caption = "First 10 Rows of IceCream Dataset")

x<-datam$Temperature
```

```
y<-datam$Ice.Cream.Profits

my_reg<-lm(y~x,datam)

#linearity
ggplot(datam,aes(x=x,y=y))+
  geom_point()+
  geom_smooth(method="lm", col="red")

#independence of errors
#normality of errors
#equal variances
par(mfrow=c(2,2))
plot(my_reg)


# Select numeric columns only
numeric_columns <- datam[sapply(datam, is.numeric)]

# Calculate the correlation matrix
correlation_matrix <- cor(numeric_columns, use = "complete.obs")

# Print the correlation matrix
print(correlation_matrix)

# Visualize the correlation matrix using corrplot
corrplot(correlation_matrix, method = "circle", type = "upper", tl.col = "black", tl.srt = 45,
addCoef.col = "black")

#for r-squared and regression equation
summary(my_reg)

#for anova and hypothesis testing
anova(my_reg)
```