Analysis of Ice Cream Sales
using R Programming

Arpit Rimza

1 Introduction

In this analysis, I will explore the ice cream sales in different cities of two countries and ways to increase them. Moreover, conclusions will be made through exploratory data analysis, hypothesis testing, modelling and prediction in order to gather new information about ice creams sales. The acquired data will be explored with the use of R programming language.

2 Exploratory data analysis

The dataset consists of 1,000 observations and six variables. None of the variables contains null values. Four out of the six variables are quantitative and contain continuous data; those are the ice cream sales in £, the average income of people in £, the price for serving ice cream in £ and the temperature in Celsius. Whereas the categorical variables are the season and the country to which the sales have taken place. The mean and median of the sales of ice cream is £1,200, the distribution is then not skewed but symmetrical (Fig. 1.a). The minimum and maximum value of sales are £272.8 and £2,127.2 respectively.

Furthermore, the ice cream sales present outliers both in country A and B for sales values over £2,000 and below £500 (Fig. 1.b). The same graph also reveals that the sales of ice creams in country A are almost identical to the sales in country B. This has also been analyzed numerically showing the small differences between the two countries. Specifically, the average of sales is £1,187 and £1,218 in country A and B respectively. The half of the observations (interquartile range) is concentrated between £975 and £1,399 of sales in country A. On the other hand, the half of the sales observations are concentrated between £1,014 and £1,413 in country B. Additionally, by comparing the sales per season, we can see outliers on spring and summer. The graph in Fig. 1.c shows that consumption of ice cream on Winter and Summer is slightly greater than in Spring and Autumn - leading in higher sales in the formers.

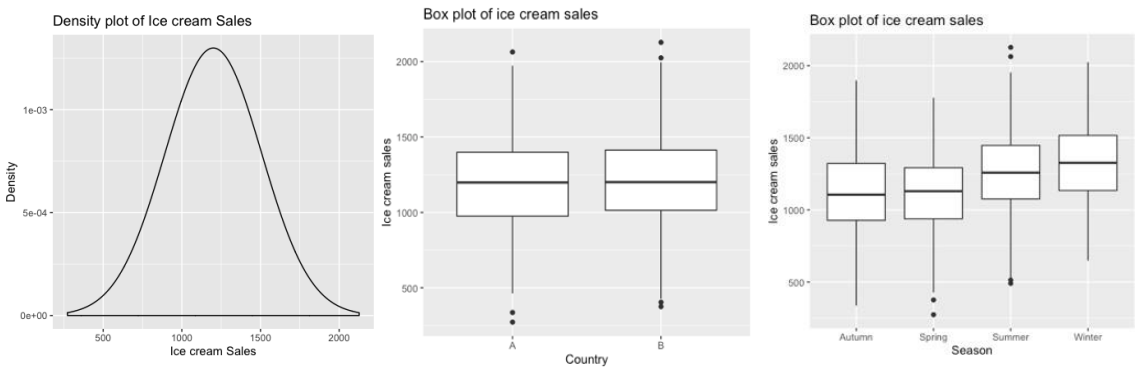

Figure 1: Analysis for the sales. Left (a), density plot of sales. Centre (b), box plot of sales per country. Right (c), box plot of sales per season.

The typical income observed in the data is around £28,084 and shows a distribution with positive skewness (Fig. 2.a), most of the values are in the range of £24,000 (1st quartile) to £30,000 (3rd quartile). Both countries, and in every season, present outliers for income over £40,000 (Fig. 2.b and 2.c) and, overall, similar distribution.

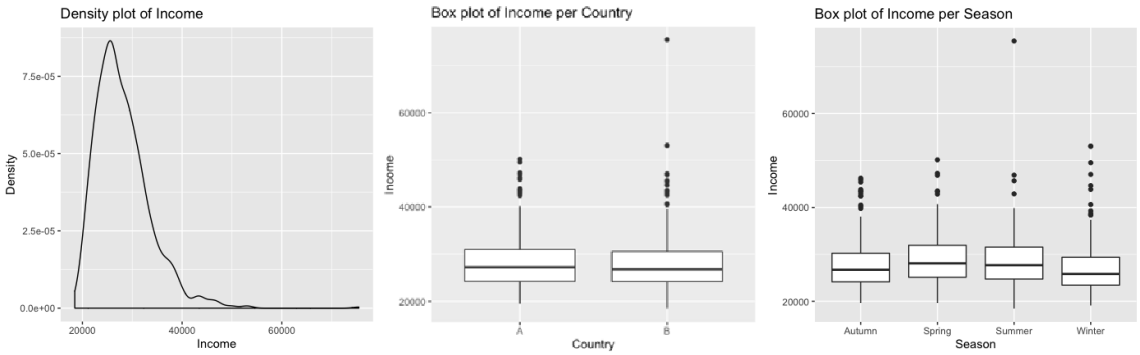

Figure 2: Analysis for the sales. Left (a), density plot of income. Centre (b), box plot of income per country. Right (c), box plot of income per season.

The distribution of the average price is similar for both country A and B (Fig. 3.a), although the median of the price for country A is slightly higher (Fig. 3.b); the distribution is negatively skewed. Half of the observations (interquartile range) is concentrated between £3.80 and £4.60. There are also some outliers for serving price under around £2.50.

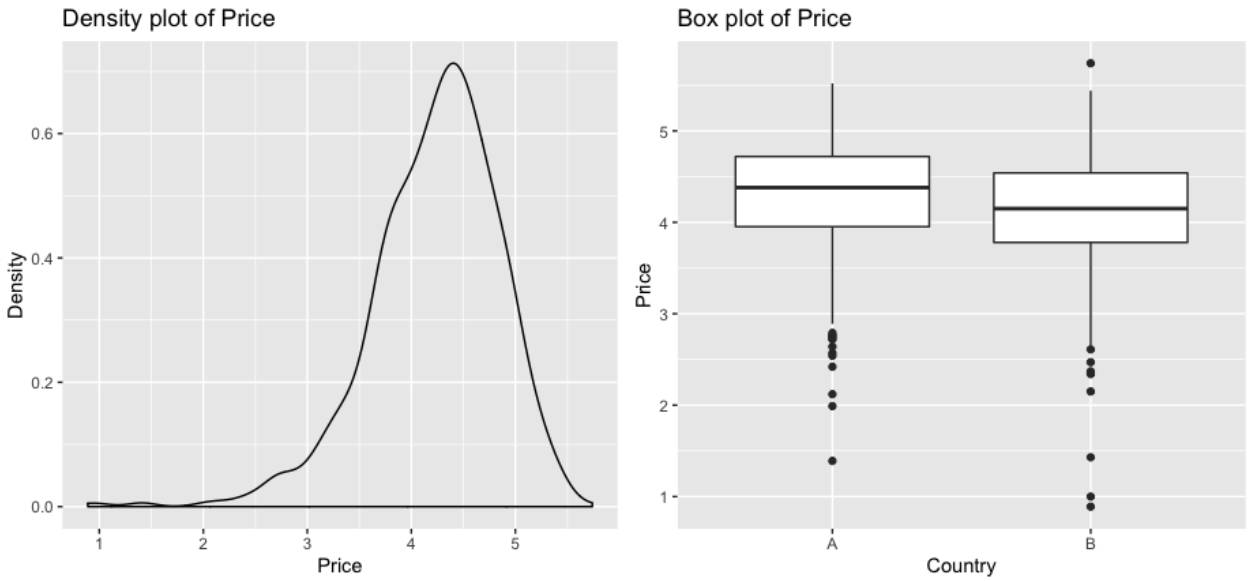

Figure 3: Analysis for the price. Left (a), density plot of price. Right (b), box plot of price per country.

As far as the temperature, the median value for country A is about 8 points lower than country B (Fig. 4). Moreover, the hottest seasons are autumn and winter, whereas spring and summer present a lower temperature with median values close to 10.

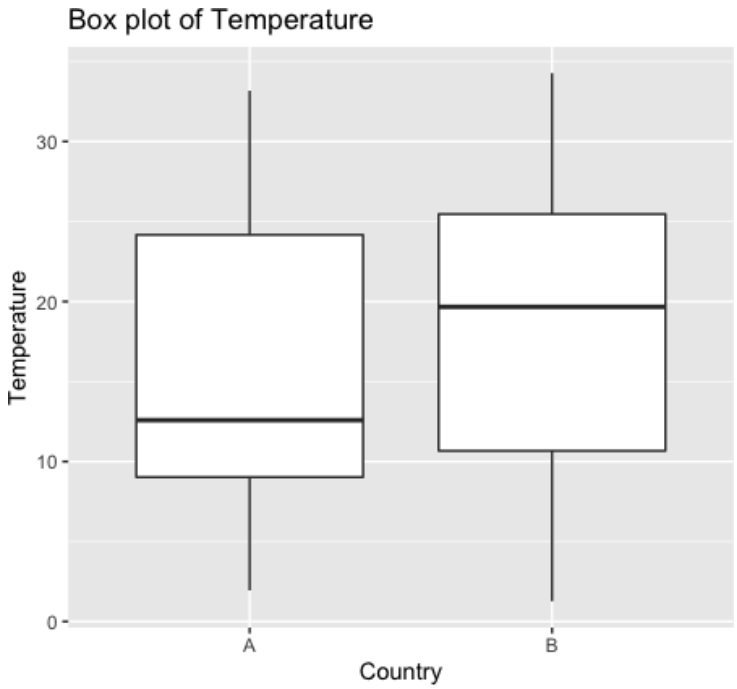

Figure 4: Analysis of the temperature for each country.

The dataset contains two different values A and B for the categorical variable country. There are 578 observations for country A and 422 for B. The fact that there is a similar number of observations for both countries help in investigating sales at a regional level instead of globally. Similarly, for the season variable, we have a balanced number of observations for each season; a minimum of 219 for Summer, and a maximum of 277 for Autumn.

3 Hypothesis testing

Assumptions (Satisfied):

- Sample randomly selected;
- Population normally distributed;
- Individual observations are independent.

Country A

- n_A = 578 observations
- μ_A = 1186.951

- σ_A = 298.528

Country B

- n_B = 422 observations
- μ_B = 1217.872
- σ_B = 297.526

Hypothesis:

- Null Hypothesis; H_0: $\mu_A = \mu_B$
- Alternative Hypothesis; H_a: $\mu_A \neq \mu_B$

Means are compared in Fig. 5 and 6. The t-statistic of the test is t = -1.6208, the p-value is 0.1058. Using a = 0.05 as the level of significance criterion, we fail to reject H_0 since the p-value is greater than 0.025. Hence, the means of sales for country A and B are not statistically different.

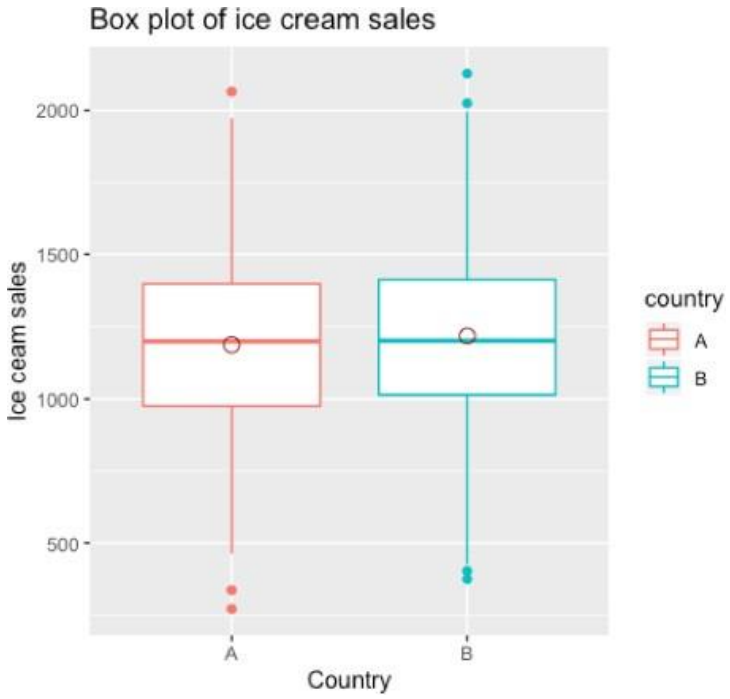

Figure 5: The box plot highlights the relationship between ice cream sales in country A and B of the data set. It shows how the medians (lines) and the mean (small circles) of the two distributions compare.

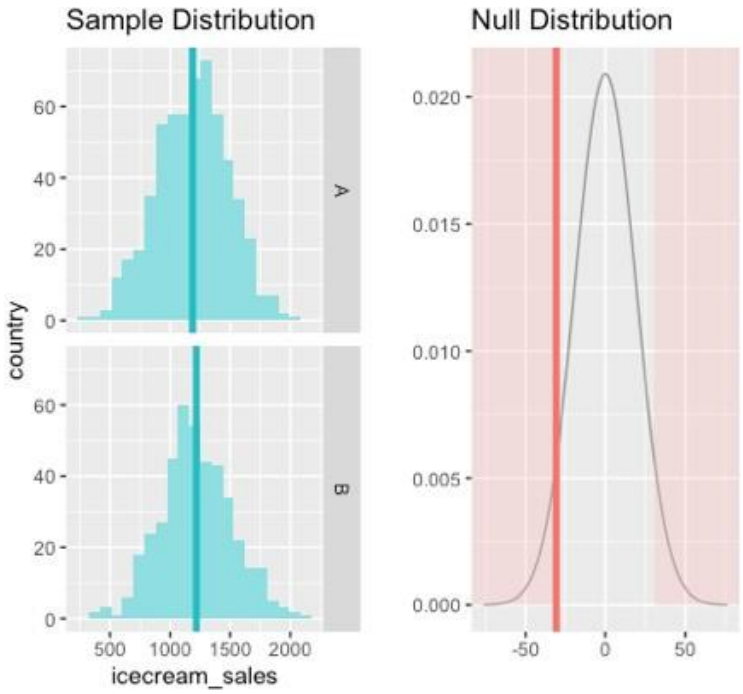

Figure 6: Results obtained from the hypothesis test.

4 Modelling

In this study, for the given data set, the focus will be on one particular technique: Multiple Linear Regression. The model consists of two categorical variables (country and season) and three continuous explanatory variables (income, price, and temperature). The outcome variable y is the sales of ice cream.

The correlation between y and the continuous explanatory variables is investigated in thecorrelation matrix in table 1. The correlation between ice cream sales and temperature is 0.574, indicating a strong positive linear association (Fig. 7). In other words, this means as the temperature increases, the sales of ice cream also increase, and vice versa. The income is positively correlated with the ice cream sales and the correlation equal to 0.053, however,the relationship does not seem so strong as the temperature with the sales (Fig. 7).

Furthermore, it would be important to mention that outliers – extreme scores in the data – influence the correlation. When there are outliers, the correlation is weaker and the Pearson's R is smaller than there are no outliers.

Lastly, we see a negative linear association between the prices of ice cream and the sales - which is -0.253 - meaning that as the prices of ice cream increase, the sales decrease. The relationship is linear as the best fitting line to the data is straight (Fig. 7).

	Ice Cream Sales	Income	Price	Temperature
Ice Cream Sales	1.000			
Income	0.053	1.000		
Price	-0.253	-0.101	1.000	
Temperature	0.574	-0.279	-0.126	1.000

Table 1: Correlation matrix of the outcome variable with the numerical explanatory variables.

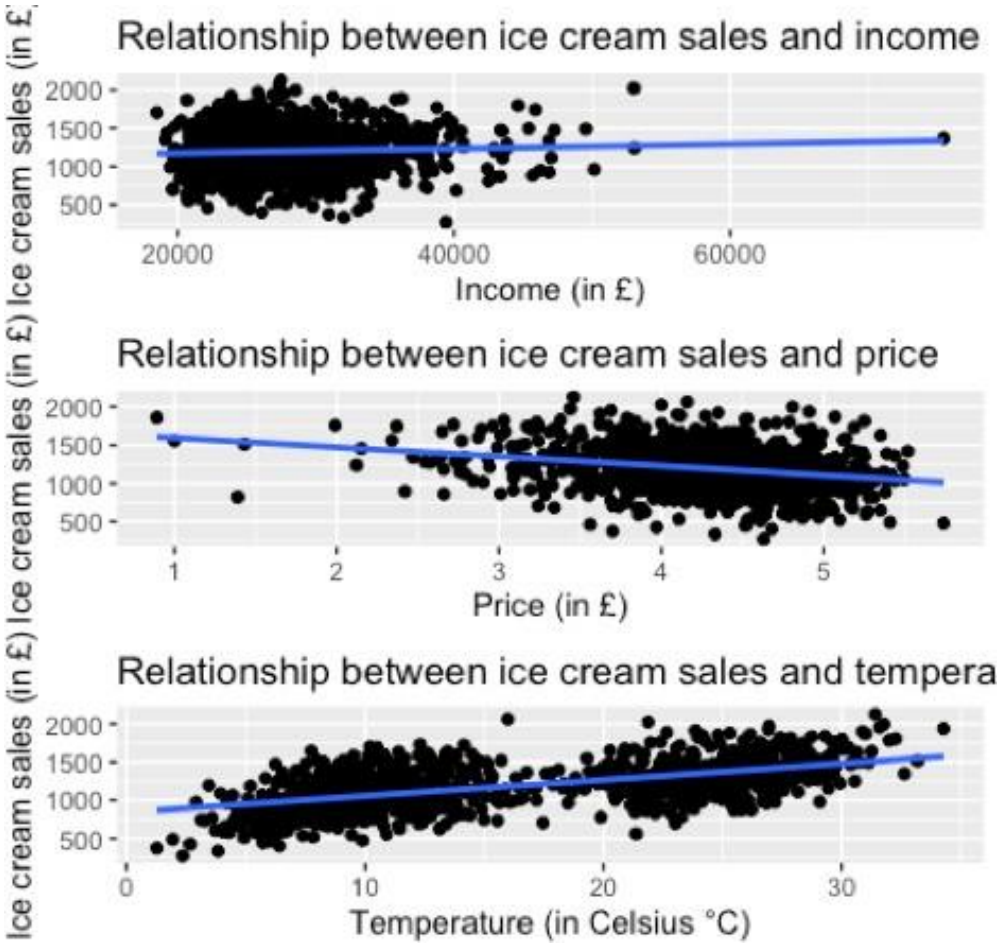

Figure 7: Linear association of the outcome variable with the numerical explanatory variables: Income (top), Price (middle) and Temperature (bottom).

The relationship of the outcome variable with the explanatory variables is visualised belowon the plots (Fig. 8), it also shows how this varies by country.

Figure 8: Scatter plots of the ice cream sales versus the numerical explanatory variables: Income (left), Price (center) and Temperature (right).

The modelling equation is presented below:

$$Sales = \beta_0 + \beta_1 * Income + \beta_2 * Price + \beta_3 * Temperature + \beta_4 * Country|B + \beta_5 * Season|Spring +$$

$$\beta_6 * Season|Summer + \beta_7 * Season|Winter$$

The regression intercept is 582.74. The estimate for the slope of income is 0.01, the slope of the price is -59.82 and the estimated coefficient of temperature is 27.02. Country A is treated as the baseline for comparison. The estimate for the slope of country B is -50.15, which is the average difference in ice cream sales that take place in country B relative to the baseline of country A. Accordingly, the intercept for country A is 582.74 and for country B is 532.59. Both countries A and B have the same slope for income, price, and temperature. Similarly, the analysis for the seasons is contiguous with the variable of country, since both variables are categorical, i.e. Autumn is treated as the baseline for comparison.

As for the interpretation of the regression coefficients, the intercept shows that when there is no income, price, temperature, and the categorical variables are indifferent as per country or which season, the sales are positive 582.74. The intercept is usually not meaningful in many contexts but has an important role to adjust the height of the regression line. The slopeof income informs that all else being constant, as income increases by £1, sales of ice cream increase by 0.01 units. In other words, when the income budget increases by £1000, sales increase by approximately 12 ice creams. The slope of price indicates that as ice cream priceincreases by £1, sales decrease by 59.83 units, all else held constant. The slope of temperature advises that when the temperature increases by 1° C, the sales are also increasing by 27.02 units with all remaining variables to be held constant. The regression model becomes:

$$Sales = 582.74 + 0.01 * Income - 59.83 * Price + 27.02 * Temperature - 50.15 * Country|B +$$

$$170.93 * Season|Spring + 207.16 * Season|Summer + 67.47 * Season|Winter$$

All else being equal, the predicted value of ice cream sales in Country A with an average income of £20,000 is 1247.32 and the predicted value of ice cream sales in Country B withan average income of £30,000 is equal to 1315.30. This leads to a difference of 67.98 units of sales. Instead, by increasing the price by £0.50 and the temperature by 2 degrees, it is observed an increase in the sales for £24.12.

The R square shows that 47% in the variation in the sales can be explained by these explanatory variables accounted for the model. We cannot measure the rest of the variation with the current data, we would need additional variables explaining changes in ice cream sales.

Furthermore, as for the inference for the model as a whole, we use F-statistic. As the p-value (2.2e-16) is less than the significance level 0.05, the model is statistically significant. This means that one of the beta coefficients is non-zero. In order to test the significance of each coefficient estimated in the model, we can look at the t-values or p-values. Assuming we are interested in significance tests at 5% significance level, we compare the calculated t-values with t-critical, which is around 1.96. Since the calculated t-values are greater than the t-critical value, we reject the null hypothesis of no significance and conclude that these coefficients are significantly different from zero.

The confidence intervals of coefficients on explanatory variables at a 90% confidence level are shown below on the table 2:

	5%	95%
(Intercept)	452.586	712.898
Income	0.009	0.013
Price	-79.584	-40.072
Temperature	25.104	28.929
Country—B	-73.769	-26.538
Season—Spring	137.004	204.860
Season—Summer	174.268	240.045
Season—Winter	33.660	101.275

Table 2: Confidence intervals of coefficients on explanatory variables at a 90% confidencelevel.

The conditions for multiple regression are:

- Linearity between the continuous or numeric dependent and continuous explanatoryvariables

- Having nearly normally distributed error terms
- A constant variance of the error
- Independence of the observations in the data

Firstly, the residual plot will be used to test linearity, which accommodates other variables in the model. Consequently, this is to see the trend in the relationship between the dependent and explanatory variables. The Fig. 9 show the residuals plots, where a random scatter around zero is observed.

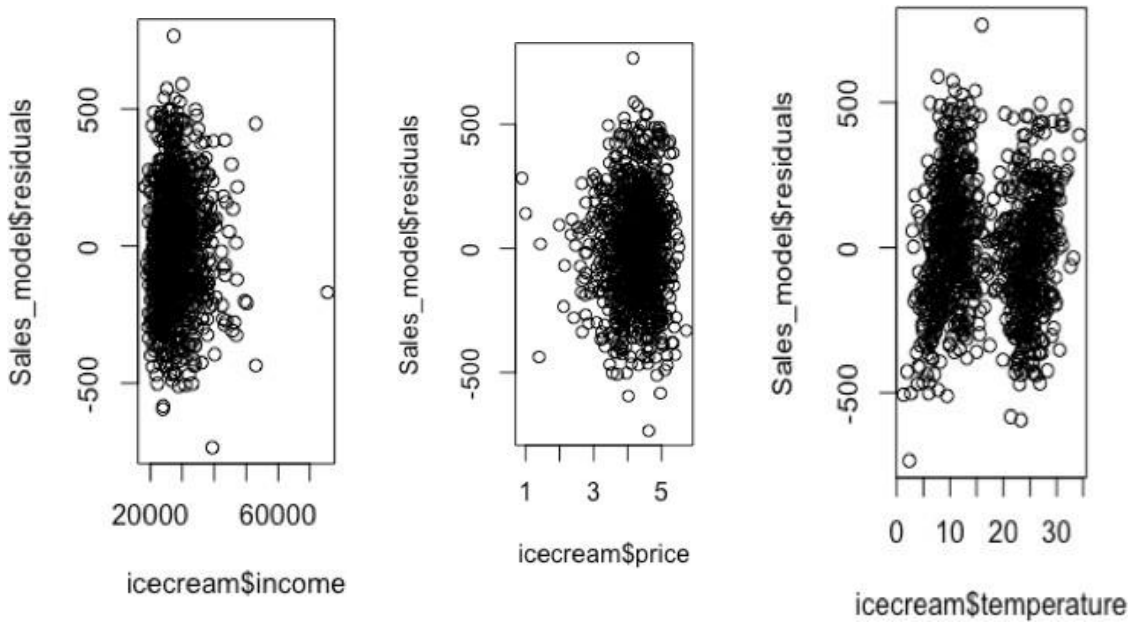

Figure 9: Residual plots of the numerical explanatory variables: Income (left), Price (center) and Temperature (right).

Secondly, the condition of nearly normal residuals around a mean zero seems to be satisfied, this is investigated through the histogram of residuals as shown below in Fig. 10, wherea positive skew is observed. Moreover, some data points are not in the straight line at the lower and upper tail and there're not huge deviations from the mean.

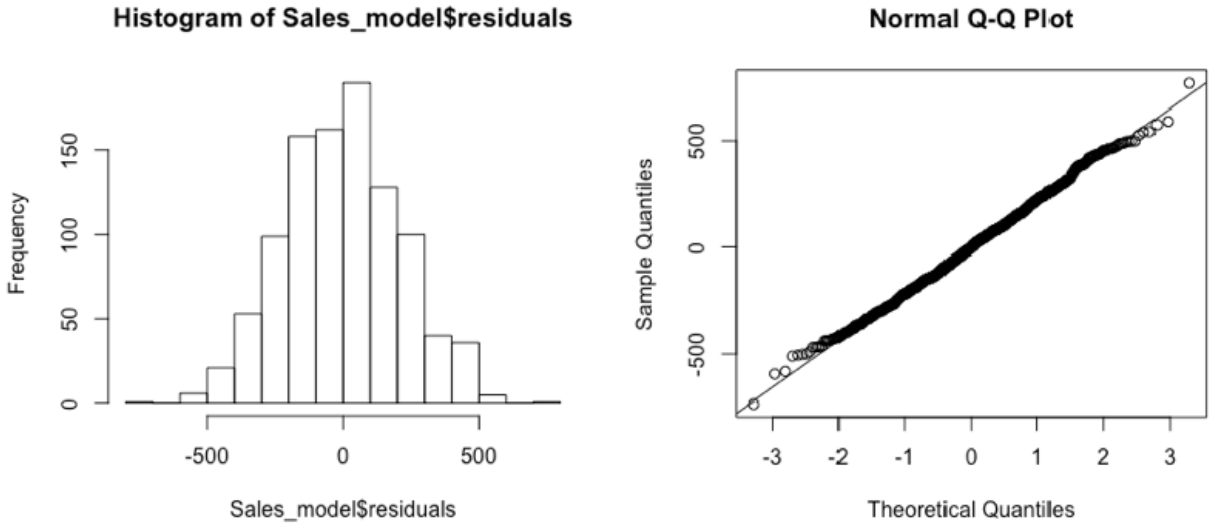

Figure 10: Nearly normal residuals around a mean zero.

The third condition, constant variability of residuals or homoskedasticity is satisfied sincethere is the same variability for lower and higher values of the predicted outcome variable, which is shown in the Fig 11 below.

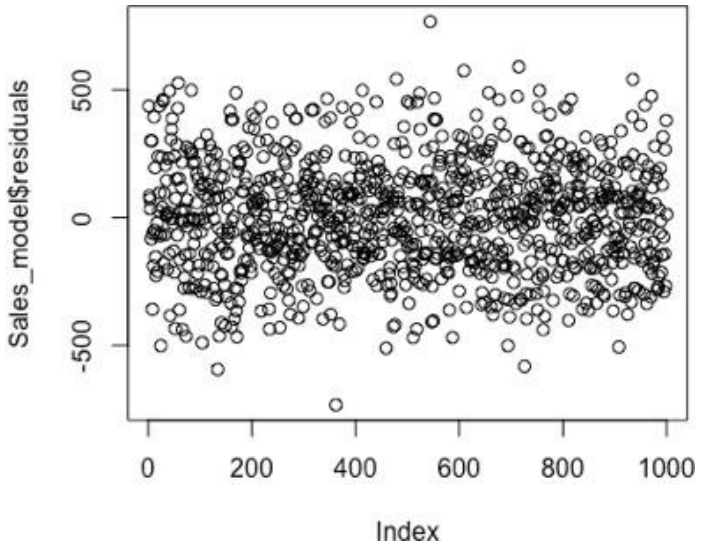

Figure 11: Variability of the residuals.

Finally, the independent residuals, which is the last condition, is also satisfied since no increasing or decreasing pattern appears in Fig. 12.

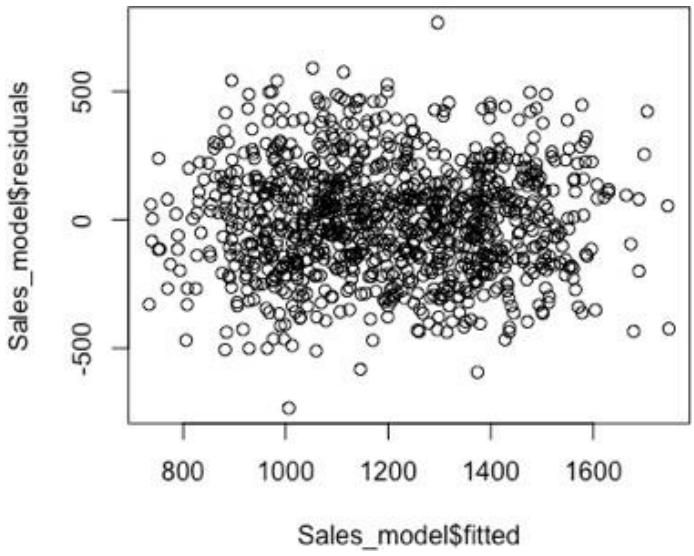

Figure 12: Independence of the observations in the data.

5 Prediction

The predicted value of ice cream sales at 95% confidence interval for the explanatory variables: income = £30,000, price = £3.00, temperature = 23° C, country = A and season = Spring;is expected to be £1,549.97. The measure of uncertainty around this prediction is provided with the prediction interval (£1120.11, £1979.82).