**Analysis of Sales data**

From the sample dataset of Global superstore data, analysis is done to find out the total sales of different products in different parts of US. For that, analysis is done to find out how the proposed independent variables - Shipping Cost, Profit and Quantity of products are related to the dependent variable, Sales. To start with, first the **correlation matrix** is calculated as shown in the table below to find the relation between variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Sales** | **Shipping Cost** | **Profit** | **Quantity** |
| **Sales** | 1 |  |  |  |
| **Shipping Cost** | 0.9628176 | 1 |  |  |
| **Profit** | 0.6743871 | 0.710757594 | 1 |  |
| **Quantity** | 0.21495 | 0.233861806 | 0.069683 | 1 |

In the above table, the first column shows the correlation of Sales, the dependent variable with all the other independent variable. All the variables have a positive correlation with Sales. Shipping Cost and Profit have strong correlation while Quantity has the weakest correlation. There is a strong collinearity between Profit and Shipping Cost. The value is 0.711 which is greater than 0.7 and hence it is significant.

Computing the multi-regression equation using all the independent variables. The excel output is shown below.



**Interpretation:**

The coefficient of variation, R² is 0.9274 which means 92.74% of the variance in Sales is because of the independent variables. The regression equation for predicting Sales would be

**ŷ = -27.46 – 0.13 (Profit) – 6.35 (Quantity) + 15.05 (Shipping Cost)**

From the above equation, Quantity and Profit are inversely related to Sales and Shipping Cost is directly related to Sales. So, if the Shipping Cost increases by 100 dollars, Sales increases by 1505 dollars. But if Quantity or Profit increases, Sales decreases according to the equation.

**The global hypothesis test** is conducted to see if any of the regression coefficients are different from 0 at 0.05 significance level.

H₀: β₁=β₂=β₃

H₁: β₁≠β₂≠β₃

Decision rule is to reject H₀ if p value is less than 0.05. From the ANOVA table above, significance of F is 0 and so **H₀ is rejected** which means at least one of the regression coefficients is not equal to 0. The individual correlation coefficients for each independent variable are given in the column P-value. From the values, only **Shipping Cost has a p-value less than 0.05 which means it is the only significant predictor of Sales**. **So, the variables** **Quantity and Profit can be removed as they are not significant predictors of Sales**.

The regression analysis output for the new regression equation is shown below



R² and Adjusted R² have changed slightly after removing the variables Profit and Quantity which again indicates that they are not the significant predictors of Sales. The p-value for Shipping cost is less than 0 which indicates the significance of Shipping Cost.

So, **the final regression equation is ŷ = 14.75 (x) - 47.75** Or **Sales = 14.75 (Shipping Cost) – 47.75**

The above scatterplot of Sales vs Shipping Cost shows an increasing trend. Sales increases as the Shipping Cost increases and so the relation is linear.

The above scatterplot shows the plot of residuals (difference between actual Sales and predicted Sales). The distribution shows an equal proportion of positive and negative residuals which indicates that the assumption of linear relation is reasonable.

The histogram above also shows that the data is normally distributed as majority of observations are in the middle.

The above diagram shows the normal probability plot for Sales. For the distribution to follow normal distribution, the dots which shows the cumulative distribution of residuals, should be closer to the line. In the above diagram, most of the plots are closer to the line.

**Summary**

The conclusion is that Sales data is normally distributed with a linear relation with Shipping Cost. As the Shipping Cost increases the Sales also increases and the linear regression equation to predict Sales is **Sales = 14.75 (Shipping Cost) - 47.75**

Reference dataset and the work sheet is attached here

