To build a multiple linear regression model to predict an outcome variable y (sales) based on multiple distinct predictor variables x (TV, radio, and newspaper).

We want to build a model for estimating sales based on the advertising expenditure on TV, Radio, and newspaper, as follow:

sales = b0 + b1\* TV + b2\* radio + b3\* newspaper

we can get the model coefficients in R as follow:

Text

Description automatically generated

R-squared is the percentage of the dependent variable variation that is explained by the linear model. Adjusted R-squared incorporates the model’s degrees of freedom. F is a test of the whole model fit. We can see that p-value of the F-statistic is 2.2e-16 (last line of the model summary), which is highly significant. It means that, at least, one of the predictor variables (TV, radio, and newspaper) is significantly related to the outcome variable (sales).

The estimated model equation can be written as:

**sales = 2.938889 + 0.045765\* TV + 0.188530\* radio - 0.001037\* newspaper**

**Q1**:

Given everything else unchanged, if “TV” advertising increases one unit (e.g., one dollar), the sales units will increase by approximately 0.045765. TV is a significant explanatory variable due to its p-value (2e-16) < 0.05.

**Q2**:

The radio coefficient suggests that for every 1 unit (e.g., one dollar) increases in the radio advertising, holding all other independent variables the same, we can expect an increase of 0.188530 sales units. Radio is a significant explanatory variable due to its p-value (2e-16) < 0.05.

**Q3**:

The newspaper coefficient suggests that for every 1 unit (e.g., one dollar) increases in newspaper advertising, holding all other predictors constant, we can expect a decrease of0.001037 sale units.

I also found that newspaper is not statistically significant in this model because of p-value (0.86) > 0.05. This means that, holding all other independent variables the same, changes in the newspaper advertising will not significantly affect sales units. The independent variable (newspaper) is not a significant predictor for "sales” in this model. Newspaper is not a significant explanatory variable due to its p-value.

**Q4**:

Chart, scatter chart

Description automatically generated

**Linearity**:

Chart, scatter chart

Description automatically generated

The residuals vs fitted plot indicates the presence of non-linear (curvilinear) relationship between an outcome variable and independent variables. We can see that all these different observations are not equally spread along the horizontal line. And, the red line is a smooth fit to the residuals, intended to make it easier to identify the trend (non-linearity in the data). The red line has curved pattern and is not approximately horizontal at zero. There is clearly curvature association with the residual or error terms. Therefore, the linearity assumption does not hold for the given observations that we used to estimate our model.

**Normality:**

Chart, line chart

Description automatically generated

The Q-Q plot of the residuals shows that the residuals are not approximately normally distributed. In the plot, although the middle portion of the residual points fall roughly on the diagonal, all other points on tails are not well distributed on the straight line. Also, the plot is slightly curved at both tails rather than straight. The overall residuals or error terms are not as close to a normal distribution, and they do not approximately follow a normal distribution. Therefore, the normality assumption may not hold in this case.