STATISTICS FOR DATA SCIENCE- PROJECT REPORT

TITLE: SIMPLE LINEAR REGRESSION

Performed By:
3836 Disha Morye
3853 Aparajita Sarkar

OBJECTIVE:

The following project helps to build a linear regression model to predict the sales of a TV model using a predictor variable.

ABSTRACT:

Simple linear regression is a statistical technique used to analyze the relationship between a dependent variable and a single independent variable. It is commonly used in data analysis to study the relationship between two variables and to make predictions based on that relationship. In this report, we will apply simple linear regression to the TV sales dataset, which contains data on TV sales and advertising expenditure.

THEORY OF SIMPLE LINEAR REGRESSION

The simple linear regression model can be expressed in the form of an equation:

$y = \beta 0 + \beta 1 * x + \varepsilon$

where y is the dependent variable, x is the independent variable, β0 and β1 are the coefficients to be estimated, and ε is the error term or residual. The goal of simple linear regression is to estimate the values of β0 and β1 in order to make predictions about the dependent variable y.

FORMULAS FOR SIMPLE LINEAR REGRESSION

The formula for estimating the coefficient β1 in simple linear regression is given by:

$\beta 1 = \Sigma(xi - \bar{x})(yi - \bar{y}) / \Sigma(xi - \bar{x})^2$

where xi is the value of the independent variable, x̄ is the mean of the independent variable, yi is the value of the dependent variable, and ȳ is the mean of the dependent variable.

The formula for estimating the coefficient β0 is given by:

$\beta 0 = \bar{y} - \beta 1 * \bar{x}$

Once the values of β0 and β1 have been estimated, we can use them to make predictions about the dependent variable y. The predicted value of y, denoted ŷ, can be calculated using the formula:

$\hat{y} = \beta 0 + \beta 1 * x$

The error or residual for each observation can be calculated as:

$\varepsilon = yi - \hat{y}$

ERROR CALCULATIONS

The two most common measures of error in simple linear regression are the sum of squared errors (SSE) and the mean squared error (MSE). The SSE is calculated as the sum of the squared errors or residuals:

$SSE = \Sigma(yi - \hat{y})^2$

The MSE is calculated by dividing the SSE by the degrees of freedom, which is equal to the number of observations minus the number of parameters estimated (in this case, 2: β0 and β1):

MSE = SSE / (n - 2)

The square root of the MSE is the root mean squared error (RMSE), which is a measure of the standard deviation of the errors:

RMSE = sqrt(MSE)

The coefficient of determination, denoted R^2, is another measure of the fit of the regression model. It represents the proportion of the variance in the dependent variable that is explained by the independent variable. It is calculated as:

$R^2 = 1 - SSE / SST$

where SST is the total sum of squares, which is the sum of the squared differences between each observation and the mean of the dependent variable:

$SST = \Sigma(yi - \bar{y})^2$

DATA ANALYSIS:

The dataset contains two variables: TV Advertising Budget (in thousands of dollars) and Sales of the product (in thousands of units). The aim of the analysis is to investigate the relationship between TV Advertising Budget and Sales, and to build a linear regression model to predict sales based on the advertising budget.

MODEL:

The Simple Linear Regression model for TV Sales dataset can be written as:

Sales = β0 + β1 * TV_Advertising + ε

where,

- Sales: Sales of the product (in thousands of units)
- TV_Advertising: TV Advertising Budget (in thousands of dollars)
- β0: Intercept of the regression line
- β1: Slope of the regression line
- ε: Error term

By evaluating the regression model, we get equation for regression line as:

Sales = 7.31 + 0.046 * TV_Advertising

RESULTS:

Sales = 7.03 + 0.05 * TV_Advertising

This equation implies that for each additional thousand dollars spent on TV advertising, the sales increase by 50 units, on average.

The coefficient of determination (R^2) for the model is 0.611, which means that 61.1% of the variation in Sales can be explained by the variation in TV Advertising Budget. The Root Mean Squared Error (RMSE) for the model is 3.26, which indicates that the model's predictions are typically off by 3.26 thousand units.

- $\beta 0 = 7.310$
- $\beta 1 = 0.0464$
- Sales = 7.310+0.0464×TV_Advertising
- MSE = 7.5
- R^2 error = 0.72

CONCLUSION:

There is a positive relationship between TV Advertising Budget and Sales, and the Simple Linear Regression model provides a reasonable fit to the data. The model suggests that increasing the advertising budget can increase sales. However, the model's predictions are not perfect and have some degree of error, which should be taken into account while making decisions based on the model's predictions.

Overall, the Simple Linear Regression model can be used to predict sales based on the TV Advertising Budget, but further research and analysis are required to develop a more robust and accurate model.