Media Sales Analysis Theory

Linear Regression

Jessica Christine Erasmus
April 2023

Table of Contents

Linear Regression

Linear regression is a statistical method used to analyse the relationship between more than one continuous variable. It uses a linear equation to fit the observed data and represent the connection between a dependent variable and one or more independent variables. Linear regression seeks the line with the greatest accuracy to reflect the connection between variables.

In terms of equations, a simple linear regression model has the equation y = mx + b. This means that y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the intercept. The slope (m) is the change in the dependent variable (y) for each unit of variation in the independent variable (x), whereas the intercept (b) is the value of the dependent variable when the independent variable is zero.

A multiple linear regression will be employed in this scenario. This entails fitting a linear equation with several independent variables. The equation would be y = b0 + b1.x1 + b2.x2 +... + bn.xn, where y is the dependent variable, x1, x2,..., xn are the independent variables, and b0, b1, b2,..., bn are the independent variable coefficients. The coefficients reflect the change in the dependent variable for a unit change in the associated independent variable if all other independent variables remain constant.

Introduction to dataset

The chosen dataset is a dataset which will be used to compare sales within different advertising platforms(TV, Newspaper and Radio).

This dataset can be found on Kaggle with the following link:

https://www.kaggle.com/datasets/ashydv/advertising-dataset

Within this project, the dataset will be found within the 'input' folder.

Why is the data set appropriate for linear regression?

Linear Regression is commonly used in datasets for sales prediction. The referred dataset is a model example of linear regression. In this case, the dataset will be used to analyse the relationship between sales and the regarded advertising platforms to identify the strongest one.

This provided dataset will use multiple linear regression with the independent variables, based on the advertising platform types, becoming the values of x1, x2, and x3 and their coefficients represented as b1, b2 and b3 with the intercept of all these variables being b0 and sales being represented as y.

The first part of this code, the data visualisation will consist of box plots and scatter plots of the initial data before it is split to the training and test set. This analysis for outliers and its initial suitability towards linear regression based on the fit towards the regression line.

Analysis which will be performed on the dataset.

The following factors will be used to analyse the dataset.

Plan to verify results.

Data visualisation will be used to verify results such as plot graphs. These graphs will be compared to the output results of linear regression to check the dataset's reliability on each advertising platform. Hence, the plan is to use residual plots, analyse the dataset's outliers and develop a linear regression model summary so that the results of the dataset can become verified.

What features need to be extracted?

There are three main features which need to be extracted by the model, namely:

1. Dependent variable: This is the variable that will be predicted or analysed. This will refer to the sales column.
2. Independent variable(s): These are the predictor variables, which are used to predict or analyse the dependent variable. Hence, our independent variables refer to our advertising platforms which as was stated earlier is TV, Newspaper and Radio.
3. Data: This refers to the dataset, 'advertising_data.csv' provided which contains both the dependent variable and the independent variables.

Furthermore, before data can be fit to the model, the data must be cleaned and pre-processed to ensure the validity of data values, should there be any null values.

How the data will fit and considerations for over and under fitting

After the data is cleaned, the data will be split to training and testing sets to assess its performance within the linear regression model and check if it is a good fit. This can help you determine if your model is overfitting, indicating it performs well on the training set but poorly on the test set, or underfitting, implying it underperforms on both the training set and the testing set.

Libraries which will be imported and used.

- NumPy - provides support for multidimensional arrays, matrices, and mathematical functions that operates on array data.
- pandas - provides data structures and functions for working with structured data to manipulate, clean and analyse data.

- Matplotlib Pyplot - provides a sufficient interface for creating different types of plots, such as line charts, scatter plots, histograms, and bar charts.
- seaborn - includes functions for creating informative visualisations of data distributions, relationships, and comparisons.
- Scikit-learn - provides tools for classification, regression, model selection and evaluation.

Which predictions will be made?

The main prediction will be made on sales based on the advertisement platforms. Given the data provided from the dataset for the sales and predictor variables, future sales will be predicted for the different types of advertisement platforms.

How the data will be visualized?

With the use of the seaborn library, the data will be most likely be visualised in a scatterplot with a regression line. The sales data will be represented on the y-axis, while the predictor variables will be plotted on the x-axis. Each data point represents a single observation and appears on the graph as a point. A regression line may also be added to the scatter plot to point out the linear relationship between the predictor variable and the sales outcome. This line depicts the best-fit line, which minimizes the difference between predicted and actual values. The regression line will be a straight line if the connection between the predictor variable and sales is linear.

Further Instructions

1. Ensure all package and libraries are imported when running. Install where necessary.
2. Run on Jupyter Notebook.
3. Ensure all files are attached and in the referenced folder. The dataset should be in the input folder.

References

Binkhonain, M. & Zhao, L., 2019. A review of machine learning algorithms for identification and classification of non-functional requirements. *Expert Systems with Applications: X.*

Hastie, T., Tibshirani, R. & Friedman, J., 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. NY: Springer.

Huilgol, P., 2023. *Bias and Variance in Machine Learning – A Fantastic Guide for Beginners!.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2020/08/bias-and-variance-tradeoff-machine-learning/
[Accessed 17 Apr 2023].

Michigan Technological University, 2022. *Statistical Analysis and Data Analysis: What's the Difference?.* [Online]
Available at: https://onlinedegrees.mtu.edu/news/statistical-analysis-vs-data-analysis
[Accessed 4 Mar 2023].

Muller, A. C. & Guido, S., 2018. *Introduction to Machine Learning with Python.* Forth ed. Sebastopol, CA: O'Reilly Media, Inc.