Data Analytics Sales Prediction Model

Avinash Dangwani
PhD Scholar
Department of Engineering (Computers)
Pacific University
Airport Road, Debari, Udaipur (Rajhasthan)

Dr. Chandansingh Rawat
Associate Professor
Dept of Electronics & Telecommunication
Engineering VESIT HAMC Collectors Colony,
Chembur Mumbai

Abstract

In every New financial year Company propose Advertisement Budget to improve their sales. Estimation of Advertisement Budget is not easy task as it involves financial parameters. Managers are always interested to know prediction model for sales which is function of Advertisement expenses.

This paper will develop Sales prediction model using simple linear regression. The model will be built using the training dataset to estimate the regression parameters. The method of Ordinary Least Squares (OLS) will be used to estimate the regression parameters using python. Regression model will be validated to ensure goodness of fit before it can be used for practical application. The single variable regression is the limitation of this model. In future multiple variables can be calculated using multiple linear regression model using python.

Keywords:

Simple linear regression, Ordinary Least Square (OLS), Training & validation data, Sum square regression (SST), Total sum of squares (SSR).

Introduction

This paper will develop sales prediction model using Simple Linear regression. sales prediction has two main methods(1) Qualitative method, (2) Quantitative method [3].Some of the Qualitative methods are Expert's Opinion Method, Sales Force Composite Method, Survey of Buyer's Expectations, Historical Analogy Method, Jury of Executive Opinions & Leading Indicators Method.

Some of the Quantitative methods areTest Marketing, Time Series Analysis, Moving Average Method, Exponential Smoothing Method, Regression Analysis&Econometric Models.

This paper will explore sales prediction using regression analysis due to its lower time complexity as compare to some of the other algorithm, Furthermore, these models can be trained easily and efficiently even on systems with relatively low computational power when compared to other complex algorithms. Building a regression model is an iterative process and several iterations may be required before finalizing the appropriate model [2]. Regression model is Organized in following sections.

ØSection – I: Simple Linear regression

ØSection – II: Ordinary least square(OLS) Method.

ØSection – III: Results& Model Diagnostics.

ØSection – IV: Conclusion

Simple Linear Regression

Simple linear regression (SLR) is a statistical technique which uses the existence of an association relationship between a dependent variable (outcome variable) and an independent variable(predictor/feature variable).

The functional form of SLR is as follows

$Yi = \beta0 + \beta1\ Xi + \varepsilon i$ (1)

Where

Yi = Value of the ith observation of the dependent variable

Xi = Independent variable of ith observation

εi = random error (residuals) in predicting thevalue of Yi

β0 & β1 = regression parameters

Ordinary least square (OLS) Method

Equation (1) can be re written as

$\varepsilon i = Yi - \beta0 - \beta1\ Xi$ (2)

The regression parameters β0& β1 are estimated by minimizing the sum of squared errors(SSE)

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \quad (3)$$

The estimated values of regression parameters are given by taking partial derivative of SSE with respect to β_0 & β_1 and solving resulting equation for the regression parameters. The estimated parameters are given by

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(Y_i - \bar{Y})}{(x_i - \bar{x})^2} \quad (4)$$

$$\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} \quad (5)$$

Where $\hat{\beta}_1$ & $\hat{\beta}_0$ are estimated values of the regression parameters β_1 & β_0 and $\bar{X}$, $\bar{Y}$ are mean values of X & Y.

A. Data Source

Sample Data is taken from Advertising Ratios & Budgets provided in annual report by Schonfeld & Associates, Inc [6]. which covers over 2,400 companies and 320 industries with information on fiscal 2018 and 2019 advertising& revenue spending.

For OLS Analysis total 52 sample companies data is taken from 12 different industries.

Table - 1 shows the sample percentage revenue growth & percentage advertisement growth for Electromedical & Electrotheraputic Appartus.

Growth is taken from 2018 to 2019			Adv Grw & Rev Grw are in %
Sr.No	Company	Ad Grw	Rev Grw
1	Adm Tronics Unlimited, Inc.	-3.88	-19.58
2	Axogen, Inc.	-85.78	27.13
3	Biolife Solutions Inc	43.33	38.64
4	Cutera Inc	0	11.67
5	Digirad Corp	0	9.6
6	Edap Tms Sa	-2.7	8.51
7	Electromed, Inc.	21.52	10.57
8	Escalon Medical Corp	22.22	-15.57
9	Fonar Corp	-11.37	6.96
10	Iridex Corp	-40	1.99
11	Masimo Corp	-21.79	9.27
12	Medifirst Solutions, Inc.	-92.13	-2.73

Table – 1:Data Source[6]: June 2020 Sample data of Advertising Ratios & Budgets from Schonfeld-Associates-Inc-v417 of Market Research.com

We will develop an simple regression model to understand and predictpercentage sales revenue growth on the percentage advertisement growth.

B. Creating Feature Set(X) and Outcome Variable(Y) Using Python

The OLS model takes two parameters Y and X.In our example percentage advertisement growth will be X and percentage sales revenue growth will be Y.We will split data set into two sets, training & validation set. Trainng set will be used to train algorithm to predict output. Validation set will be used to test accuracy & efficiency.

C. Python for Building Regression Model

Python language is used as tool for building regression model for sales prediction. The statsmodel library is used in Python for building statistical models. OLS(Oridnary least square) API available in statsmodel.api is used for estimation of parameters for simple linear rgression model.

D. Splitting the Dataset into Training and Validation Set

The data is divided into two subsets training data set and validation data set. The proportion of training dataset is usually between 70% and 80% of the data and the remaining data is used for validation data. We have taken train_size = 0.8which implies that 80% of the data will be used for training the model and remaining 20% will be used for validating the model. The records that are selected for training and test set are randomly sampled using python functions which returns four variables as shown below.

train_X = feature values of the training set

train_Y = response values of the training set

test_X = feature values of the test set

test_Y = response values of the test set

E. Finding estimated parameters

Regression parameters $\hat{\beta}_1 \& \hat{\beta}_0$ are estimated from equations (4) & (5) using Python functions as tool.

F. Fitting the Model

Linear regression calculates an equation that minimizes the distance between the fitted line and all of the observed data points. Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals. In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased.

G. Co-efficient of Determination (R-Squared / R^2)

R-squared is a statistical measure of how close the data are to the fitted regression line. It is defined as

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}} \quad (6)$$

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2} \quad (7)$$

SSR = The *sum squared regression (SSR)* is the sum of the square residuals $(y_i - \hat{y}_i)^2$. Residual is the difference between observed value y_i & estimated value $\hat{y}_i$ as shown below in Fig - 1.

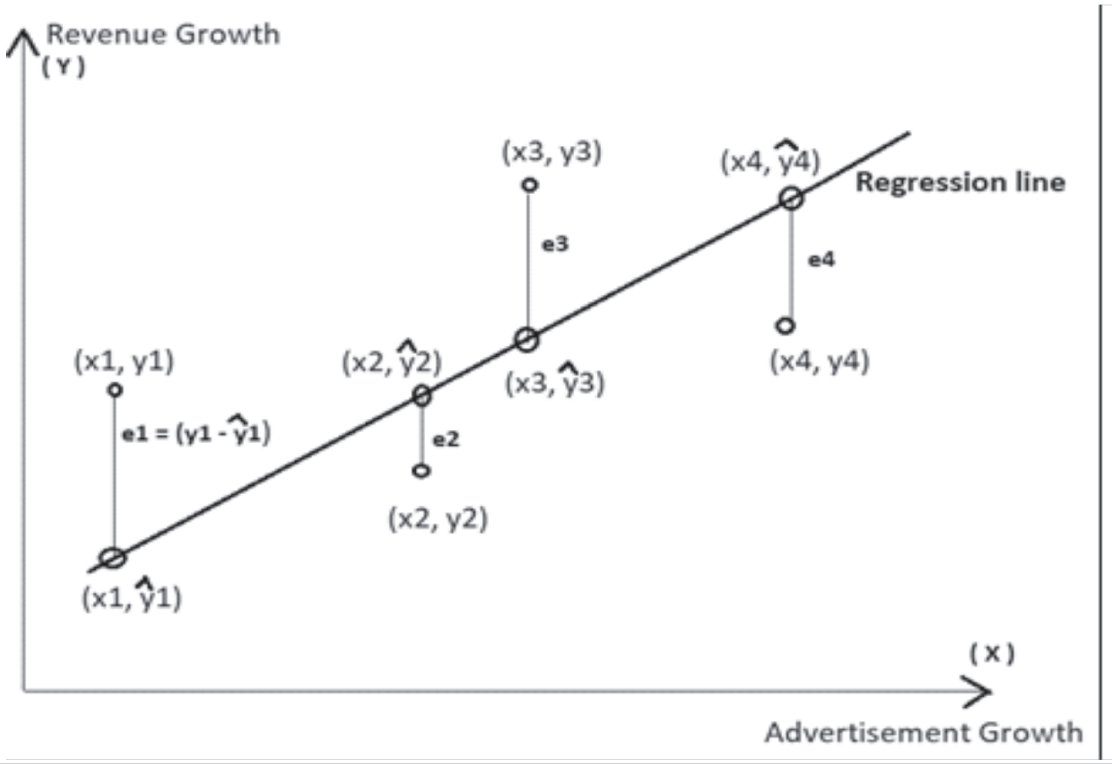

Fig – 1: Residuals as function of Actual value & estimated value

$$SSR = \sum(y_i - \hat{y}_i)^2 = e_1^2 + e_2^2 + e_3^2 + e_4^2 \quad (8)$$

= square sum of variations w.r.t to estimated value

SST = *total sum of squares* is the sum of the distance the data is away from the mean (central tendency) all squared as shown below in Fig - 2.

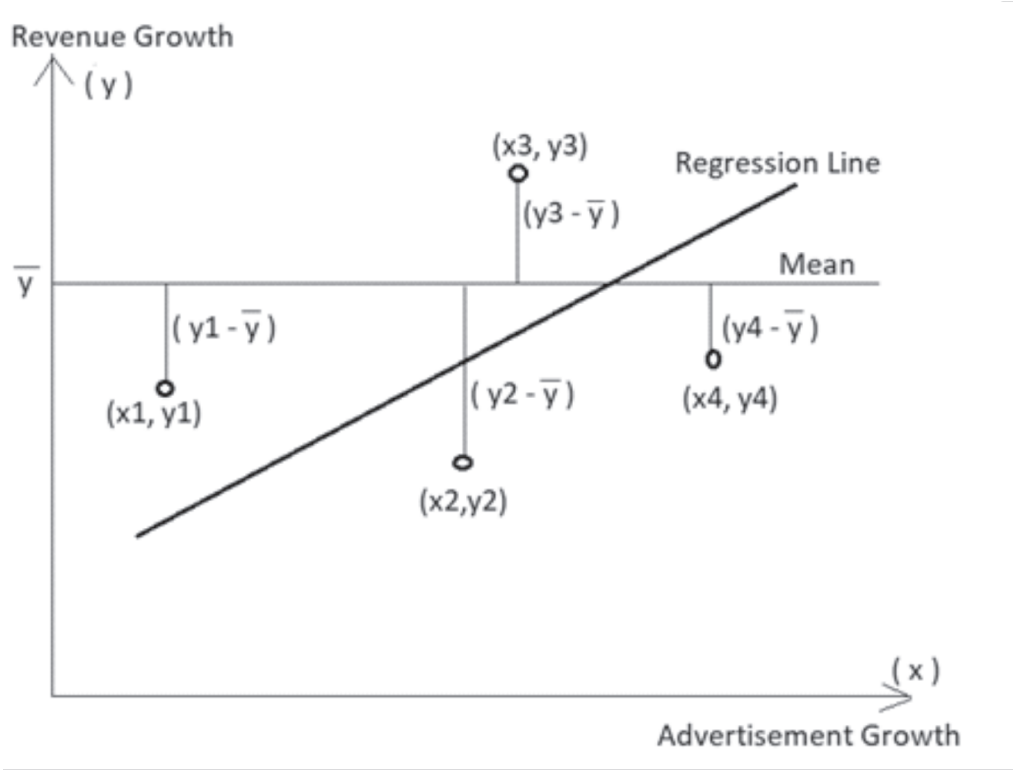

Fig – 2: Residuals as function of Actual value & Mean value

$$SST = \sum(y_i - \bar{y})^2 = (y1 - \bar{y})^2 + (y2 - \bar{y})^2 + (y3 - \bar{y})^2 + (y4 - \bar{y})^2 \quad (9)$$

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}$$

$$R^2 = \frac{SST - SSR}{SST}$$

$$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (10)$$

The above equation indicates that R2 is directly proportional to difference between the square sum of variations in y w.r.t mean and square sum of variations in y w.r.t estimated value.

Not good fit:

Smaller R2 value indicates that SSR value is large and close to SST which indicates that variation in y w.r.t estimated value is large & close to variation in y w.r.t mean, which is not good fit.

Good fit:

Large R2 value indicates that SSR value is very small (actual values of y are close to estimated values of y) and not close to SST, which indicates that variation in y w.r.t estimated value is not close to variation in y w.r.t mean, which is a good fit.

Results & Model Diagnostics

A. Estimated Model

Using python as tool parameters of regression model are calculated as shown below.

Using 80% training data set

- ✓ Constant $\hat{\beta_0}$ = 6.101
- ✓ Regression coefficient $\hat{\beta_1}$ = 0.160

The estimated model can be written as

$Y_i = \beta_0 + \beta_1 X_i (11)$

Rev Grw(%) = 6.101 + 0.160 * (Ad Grw(%))

B.Interpretation of Estimated Model

Model estimates that 1% Ad Growth will increase Revenue by 0.160 %. For example, if the sales revenue was 2 Million in year 2018 then according to our model sales revenue in year 2019 will increase by 0.0032 million i.e. estimated sales revenue can be 2.0032 millions that is rise of 3200/- in revenue.

C.Model Diagnostics (Validation)

Before using regression model in practical applications, it should be validated & tested for goodness of fit. We will be using Co-efficient of determination (R-squared) method to determine goodness of fit. Using python as a tool following value of R2 is calculated

R2 = 0.208

According to Cohen – 1992 [9] r-square value 0.12 (12%) or below indicate low, between 0.13 (13%) to 0.25 (25%) values indicate medium& 0.26 (26%) or above values indicate high. Our model explains 20.8% of the variance in the validation set, so it is reasonably good fit.

Conclusion

The simple linear regression model using ordinary least square (OLS) method shows functional relationship between the outcome variable (Sales revenue growth in %) and the feature (advertisement growth in %). The model validation is investigated using R2 technique to ensure goodness of fit.while an R-square as low as 10% is generally accepted for studies in the field of arts, humanities and social sciences because human behavior cannot be accurately predicted, therefore, a low R-square is often not a problem in studies in the arts, humanities and social science field. There are various other control parameters which affects the value of R-square. Therefore, in order to extend scope of this research various social science characteristics like age, gender, motivation towards product and festive season should be included as control variables in analysis.

References:

[1] Core Python Programming by Dr.R.Nageswara Rao second edition dreamtechPress.

[2] Machine Learning Using Python by Manaranjan Pradhan & U Dinesh Kumar first reprint edition Wiley publications

[3] Sales prediction types available online at URL: https://www.economicsdiscussion.net/sales/sales-forecasting-methods/32270

[4] Advantages of Linear regression model available online at URL: https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/

[5] Will Koehrsen Article how to setup your-machine learning problem can be found at following URL:https://towardsdatascience.com/prediction-engineering-how-to-set-up-your-machine-learning-problem-b3b8f622683b

[6] Data source of Ratios & Budgets can be found at following URL:

https://www.marketresearch.com/Schonfeld-Associates-Inc-v417/Advertising-Ratios- Budgets-13373044/

[7] https://www.keboola.com/blog/linear-regression-machine-learning.

[8] https://internal.ncl.ac.uk/ask/numeracy-maths-statistics/statistics/regression-and-correlation/coefficient of-determination-r-squared.html#:~:text=R2%3D1%E2%88%92sum%20squared,from%20the%20mean%20all%20squared.

[9] Cohen's Conventions for Small, Medium,and Large R2 values can be found on following URL http://core.ecu.edu/psyc/wuenschk/docs30/EffectSizeConventions.pdf

[10] Small is beautiful. The use and interpretation of R2 in social Research by Ferenc Moksony Pages 6 & 7