Theory of Linear Regression - Learn and grow with Analytics

Notebook:	Tester and downloads
Created:	7/17/2017 19:36
URL:	https://equiskill.com/introduction-to-linear-regression-for-business/

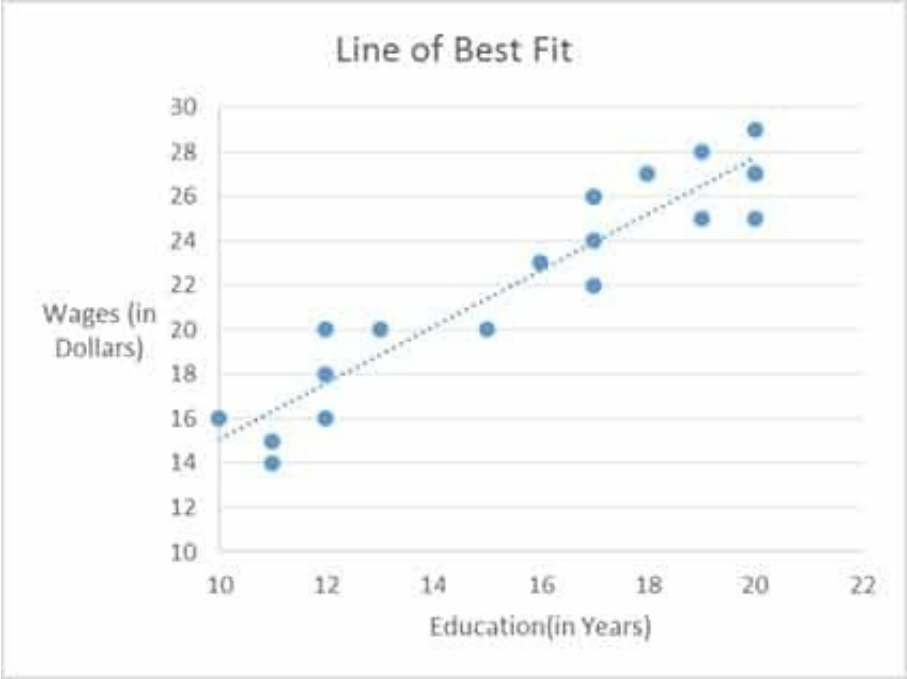
Theory of Linear Regression

Posted by
AMIT UPADHYAY

Categories
ANALYTICS, BUSINESS ANALYTICS, LINEAR REGRESSION

Date
MARCH 29, 2016

Comments
0 COMMENT



Summary

Linear regression is generally used to model the linear relationship between a continuous dependent variable and one or more independent (predictor) variables. When there is only one independent (predictor) variable, it is called “Simple Linear Regression” and when there are more than one independent variables, it is referred as “Multiple Linear Regression”. The general form of the linear regression model is given below. Where b0 is the Y intercept, e is the error in the model, b1 is the coefficient (slope) for independent factor x1, and b2 is the

coefficient (slope) for independent factor x2 and so on.

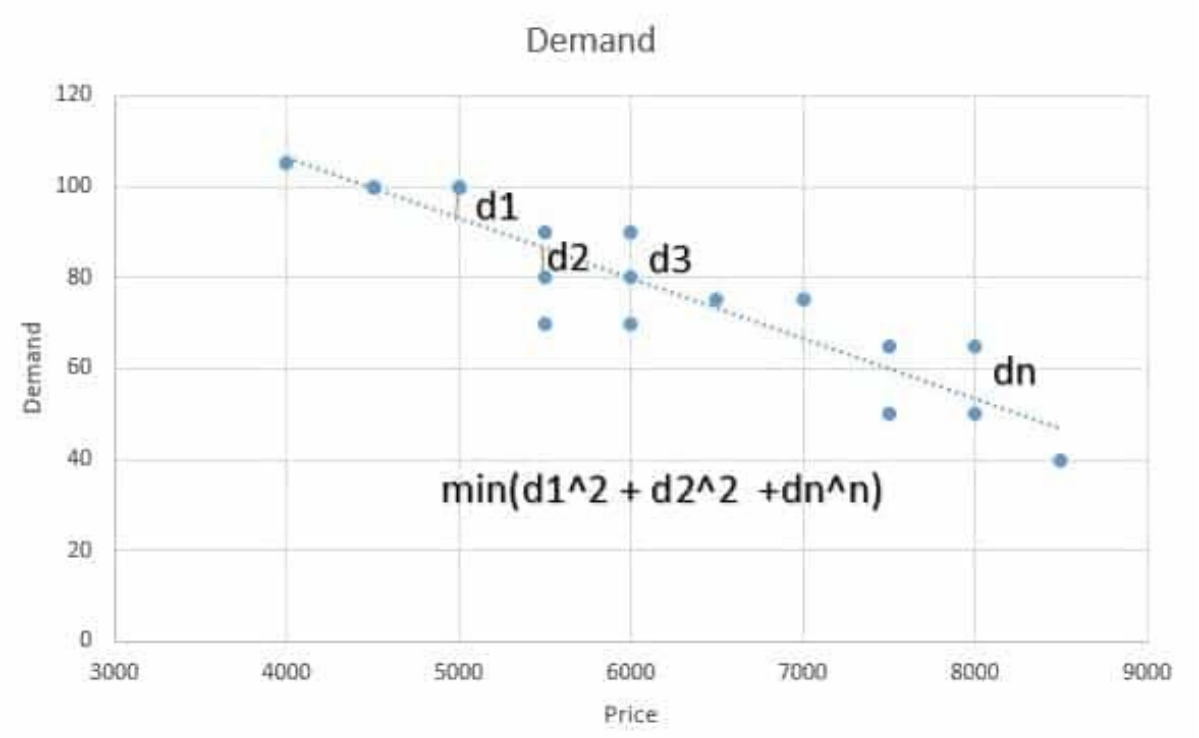
$$y = b0 + b1 * x1 + b2 * x2 + e$$

Applications of Linear Regression

Linear regression can be used for establishing the relationship (coefficient estimation) between independent and dependent variables, testing hypothesis and prediction of dependent variable for a certain combination of independent variables. Examples- 1. Prediction of income of a customer based on the zip code where customer is living, spend pattern, # of loans, payment patterns etc. 2. Prediction of sales for an e-commerce firm 3. Prediction of unemployment rate of a country etc. 4. What-if Analysis based on model established relationship between dependent and independent factors.

Underlying Algorithm and Assumptions

The underlying algorithm is called Ordinary Least Square (OLS). The algorithm tries to fit a straight line that passes through the points in a way that it minimizes the sum of squared distance of the points from the line. In other words, it minimizes the sum of squared error in predictions.



Some of the key assumptions for linear regression models are- **Linear Relationship-** The

predictor (Xs) variables and dependent variables have a linear relationship. This can be easily verified by plotting the X vs Y in scatter plot. If the linear relationship doesn't exist, either the variables need to be transformed or some other technique should be used. **No Heteroscedasticity-** The error between the predicted values and the actual values should be randomly distributed for all values of independent factors. This can be easily verified by plotting the error (residual) terms against each X. If there is no pattern there is Homoscedasticity, otherwise there is Heteroscedasticity (lack of constant variance). If Heteroscedasticity is present, this needs to be fixed prior to finalizing the model. **No or little Multicollinearity**– The independent factors should not be correlated to each other. If they are collinear, some of these need to be excluded from the final model to provide stability to the model and estimated coefficients. **Normality and Independence**– Residuals (errors), i.e. predicted minus actual data, should be normally distributed with mean of zero and constant standard deviation, and the residuals of independent factors should not be correlated to each other. **Independence-** Observations are independent from each other. Y (X+1) should not be correlated to Y (X)

Tools to Build Linear Regression

Excel- "Data Analysis" tool pack in Excel has a tool for building multiple linear regression models **R-** Function "lm" or "glm" are frequently used for building linear regression models **SAS-** Proc REG in SAS achieves the same objective.

Key Metrics and Interpretation

There are several metrics generated in the multiple linear regression output. Key ones are- **R^2 (R square) –** This tells the percent of variance in the dependent variable that can be explained by the model and the independent variables in the model. R^2= Explained Variation / Total Variation. The range for this metric is 0 to 1 or from 0% to 100%. If the R^2 is 0% that means the model explain 0% of the variation in dependent variable. On the other hand, 100% signifies a perfect model, i.e. explains 100% of variations. R^2 should be as to 1 as possible for a good model. **F Statistics and Related 'p' or significance value**– The F test measures the lift in the model with predictor variables versus a model with only intercept. 'p' value is giving the significance of rejecting the null hypothesis that all model coefficients for predictor variables are zero. F stat should be as high as possible and the associated p value of significance should be as low as possible for a good model. For example, a p value of 0.002 means that we are 1-0.002 or 99.8% confident that some coefficients of the

independent variables are non-zero in the model. In other words, some independent variables have good explaining power for the dependent variable. **Coefficients Estimate-** Coefficient estimates are the multiplicative term for each of the independent factor to derive the regression equation. In the example below, we are modelling for Sales of an eCommerce company. The equation for this can be derived as-

*Sales= 624.69704+ 0.18184*Marketing Budget- 0.556408*Price*

Coefficients:

	Estimate	Std. Error	T value	Pr(>|t|)
(Intercept)	624.69704	5.29087	118.071	<2e-16 ***
Marketing Budget	0.18184	0.97023	0.187	0.851
Price	-0.556408	0.06547	-84.982	<2e-16 ***

Std. Error can be used for constructing the confidence interval around the coefficient estimates. **t value** is the T-test stat for the null hypothesis that the coefficient for this independent factor is zero. Associated p value is shown for the same T-test. T value should be as high and p value should be as low as possible for a good model. A p value of greater than 0.05 signifies that the variable is not a good predictor. Similarly an absolute (T value) of less than 2 shows a weak predictor. In above case for example, Marketing Budget has a p value of 0.851 and T value of 0.187, and hence this is not a good predictor and should be removed from the regression equation. So the equation should like this- Sales= 624.69 – 0.56*Price (with rounding). The interpretation of the coefficient of the independent factor is really simple. This shows how a unit level change in the independent factor will change the dependent factor. For example, if the Price of an item goes up $ 1, it will cause a $ -0.56 drop in sales. For a more detailed video on the above technique, assumptions validations, data treatment, hands-on exercise, and codes for running linear regression in Excel, SAS, R and Knime, please Click here to view a video on Linear Regression. We offer instructor led online webinars (Free and Paid) on all Analytics and Data Science topics. If you would like to see our courses please click Courses Thanks for reading through. The Writer of the article is Ratnakar Pandey and has vast and rich experience in Analytics across Target, Citibank and Texas Instruments. Please ask any questions using the comment box below and we will respond in 24 hours.

Amit Upadhyay

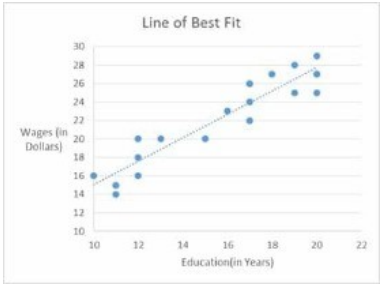
Previous post

Influencing Skills & Social Styles workshop
March 29, 2016

Next post

A beginners guide to Linear Regression
23 April, 2016

YOU MAY ALSO LIKE



A beginners guide t...

23 April, 2016