****

9B17E002

linear regression: a high-level overview

Greg Zaric wrote this note solely to provide material for class discussion. The author does not intend to provide legal, tax, accounting, or other professional advice. Such advice should be obtained from a qualified professional.

*This publication may not be transmitted, photocopied, digitized or otherwise reproduced in any form or by any means without the permission of the copyright holder. Reproduction of this material is not covered under authorization by any reproduction rights organization. To order copies or request permission to reproduce materials, contact Ivey Publishing, Ivey Business School, Western University, London, Ontario, Canada, N6G 0N1; (t) 519.661.3208; (e)* [*cases@ivey.ca*](mailto:cases@ivey.ca)*;* [*www.iveycases.com*](http://www.iveycases.com)*.*

Copyright © 2017, Richard Ivey School of Business Foundation Version: 2017-01-27

Linear regression is one of the most commonly used tools in statistics and analytics and can be thought of as the foundation for all predictive analytics. Linear regression is an attempt to create a line of best fit in a data set and use that line to explain the relationship between two quantities and to predict future values.

APPLICATIONS

A sample of some applications of linear regression is listed below:

**Marketing:** predicting sales as a function of advertising spending.

**Real Estate:** estimating the value of a residential property as a function of size and distance to public transit.

**Education:** predicting student performance as a function of scores on standardized tests.

**Human Resources:** understanding the relationship between salary and qualifications.

**Scheduling:** predicting future workload based on the number of jobs scheduled two weeks in advance.

**Sports:** predicting stadium revenue as a function of a team’s win/loss record.

**Entertainment:** predicting the number of viewers for a television movie based on the day when it airs, the number of “A-list” stars it contains, and the subject matter.

**Automotive:** predicting fair values for used cars based on model, mileage, age, and features.

**Food and Beverage:** predicting future wine prices as a function of the average temperature and precipitation in France in the year when the grapes were grown.

TERMINOLOGY

**Dependent Variable:** In linear regression, you are often trying to predict or forecast something. This is called the dependent variable and is typically represented by the letter Y. In the marketing example described above, “sales” would be the dependent variable because this is what you are trying to predict.

**Independent Variables:** This information is what you have and will use to make your predictions and is typically represented by the letter X. Depending on the application and the background of the user, these variables are also called “features,” “factors,” and “attributes.” In the marketing example described above, “advertising spending” would be the independent variable because this is what you are using to try to predict sales.

**Regression Line:** This line of best fit describes how the dependent variable (Y) changes when the independent variables (X) change.

**Scatterplot:** This graph of your data set illustrates the relationship between the independent and dependent variables. Typically, the independent variable is shown on the horizontal axis and the dependent variable is shown on the vertical axis. An example is shown below.

**Coefficients:** These numerical measures describe the relationship between the dependent variable and the independent variable. In particular, the “slope coefficient” describes how much the dependent variable increases or decreases for a 1-unit change in the independent variable. In the marketing example described above, the slope coefficient would describe how much sales increase for each extra dollar invested in advertising.

**Goodness of Fit:** This formal measure expresses the quality of the model. The most commonly used measure is called “R-squared” (or R2), and it ranges from zero, representing no fit, to 100 per cent, representing a perfect fit.

All of these concepts are illustrated in the sample regression example in the next section.

A Sample Regression

The following example shows how regression could be used to explain and predict sales prices for converted office properties as a function of the size of the building.

$-

$50,000

$100,000

$150,000

$200,000

$250,000

$300,000

$350,000

0

1,000

2,000

3,000

4,000

5,000

6,000

7,000

Area in SF\*

Price vs. Area

\*SF = square feet

The scatterplot shows an upward trend, which makes sense—you expect larger buildings to have higher sales prices. The relationship between price and size is obvious, but it is also obvious that the relationship is not perfect. Again, this makes sense because there are several factors that contribute to the value of an office property, including location and condition.

You can be more formal about expressing the relationship between sales price and size by adding the line of best fit (the regression line), the equation for the line of best fit, and the measure for goodness of fit to the graph.

y = 37.347x + 98922

R² = 0.4604

$-

$50,000

$100,000

$150,000

$200,000

$250,000

$300,000

$350,000

0

1,000

2,000

3,000

4,000

5,000

6,000

7,000

Area in SF

Price vs. Area

The new graph shows the line of best fit. Note that approximately half the points are above the line and half are below. This is because the line has to explain all of the observations, on average. The box on the graph shows the equation of the regression line and the measure for goodness of fit.

The R-squared term is 0.4604, indicating that 46 per cent of the variability in sales prices can be explained by variability in area. Thus, one single factor explains a large amount of variability in price.

The regression line is stated in general terms with Y and X, the common notation for dependent and independent variables. To use the line, you need to translate it back to your original situation. Thus, you interpret the line as

The first coefficient— 98,922—is typically not interpreted on its own and is only useful for improving the predictive accuracy of your model. The second coefficient, 37.347, says that each additional square foot of space adds $37.35 to the sales price, on average.

How would you use the line? If you are trying to predict a fair price for a property and you know the size of the property, then you can substitute that value into the regression equation to predict the fair price. For example, a 4,715 square foot office property is predicted to cost

Generalization—More than One Explanatory Factor

In most cases, there is more than one factor that can be used to explain and predict the dependent variable. In the office property example above, the original model fit was not perfect. If available, you can add information about other independent variables and determine a new regression line. One example would be a model containing “area in SF,” “number of parking spots,” and “zoned for medical use (yes or no).” This results in a new regression line:

The model can be used in the same fashion as before; if you know the values for area and the number of parking spots, and if it is zoned for medical, you can substitute those numbers into the regression equation to determine a predicted fair price. Since you cannot use “yes” and “no” in an equation, you can use the number 1 if it is zoned for medical and 0 if it is not. The new model has an R-squared value of 0.6127, indicating that 61 per cent of the variability in sales price can be explained by this combination of factors. By adding terms to your model (number of parking spots and medical zoning), you have improved the fit of the model.

Linear Regression versus Other Techniques

A number of other modern techniques, such as neural networks, clustering algorithms, k-Nearest Neighbours algorithms, CART, and random decision forests are gaining in popularity. However, linear regression is often the benchmark against which these techniques are judged, for several reasons:

1. Linear regression has been around for more than 100 years and is well understood.
2. A linear regression model is easy to interpret and explain. The other methods can be used to make predictions, but often the methods are not nearly as transparent as in a regression, which can cause a lack of trust with the methods.
3. Specialized software is not needed to build and use regression models. Even Microsoft Excel has a number of tools for working with regression models.

Thus, an important consideration in using a more advanced technique is whether you expect the performance to exceed that of a standard linear regression model.