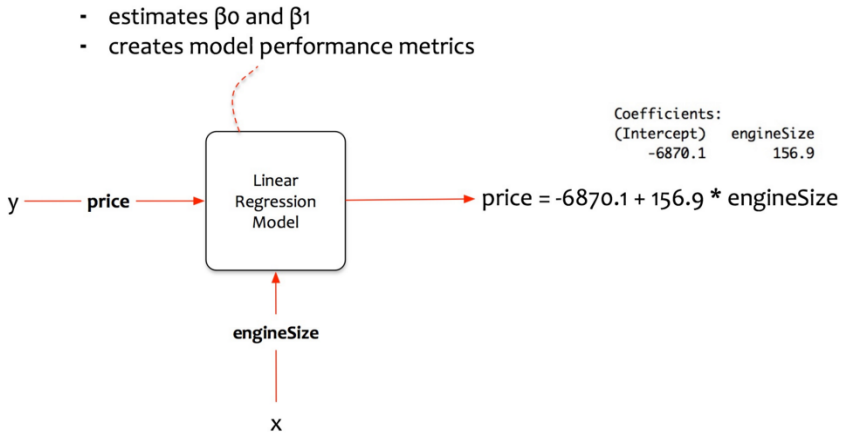
Pradeep Menon Follow
Director of #BigData and #AI Solution Architecture @ Alibaba Cloud. Impact driven. Executive-level interpersonal skills. Hands-On. #WorldTraveller. #Blogger
Aug 5, 2017 · 8 min read

Data Science Simplified Part 5: Multivariate Regression Models

In the last article of this series, we discussed the story of Fernando. A data scientist who wants to buy a car. He uses Simple Linear Regression model to estimate the price of the car.



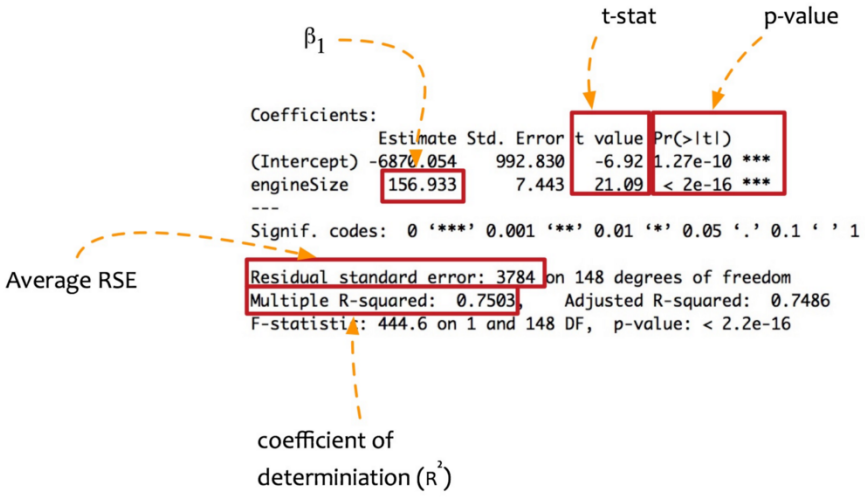
The regression model created by Fernando predicts **price** based on the **engine size.** *One dependent variable predicted using one independent variable.*

The simple linear regression model was formulated as:

price = β0 + β1 x engine size

The statistical package computed the parameters. The linear equation is estimated as:

price = -6870.1 + 156.9 x engine size



The model was evaluated on two fronts:

- Robustness- using hypothesis testing
- Accuracy- using the coefficient of determination a.k.a R-squared

Recall that the metric R-squared explains the fraction of the variance between the values predicted by the model and the value as opposed to the mean of the actual. This value is between 0 and 1. The higher it is, the better the model can explain the variance. The R-squared for the model created by Fernando is 0.7503 i.e. 75.03% on the training set. It means that the model can explain more than 75% of the variation.

However, Fernando wants to make it better.

He contemplates:

- What if I can feed the model with more inputs? Will it improve the accuracy?

Fernando decides to enhance the model by feeding the model with more input data i.e. more independent variables. He has now entered into the world of the multivariate regression model.

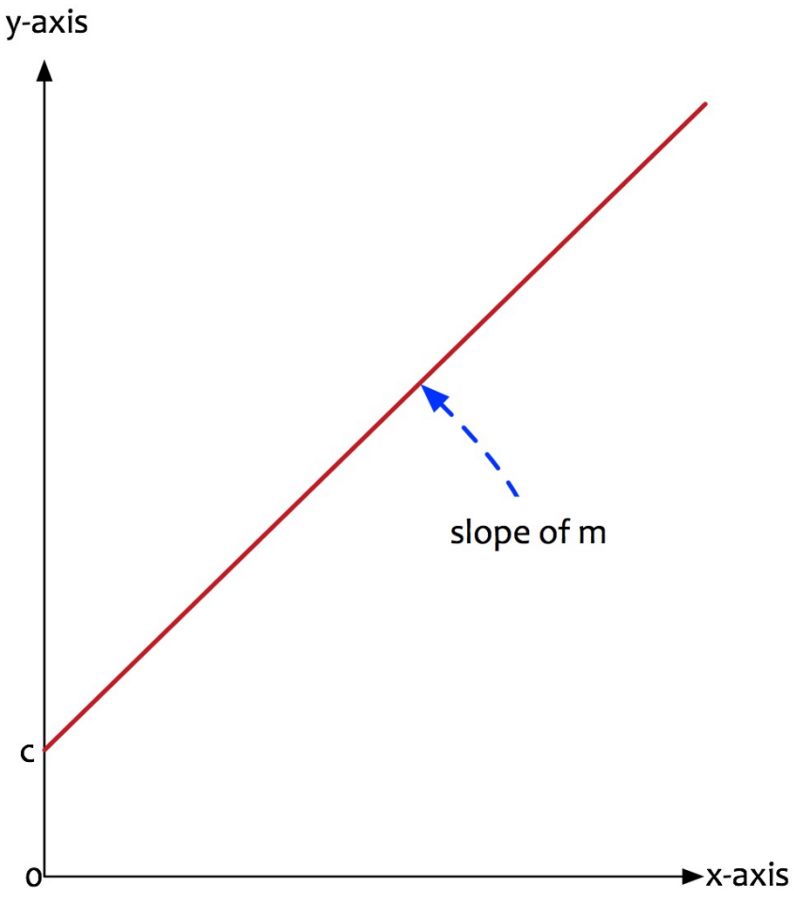
The Concept:

Linear regression models provide a simple approach towards supervised learning. They are simple yet effective.

Recall that linear implies the following: arranged in or extending along a straight or nearly straight line. Linear suggests that the relationship between dependent and independent variable can be **expressed in a straight line.**

The equation of the line is y = mx + c. One dimension is y-axis, another dimension is x-axis. It can be plotted in a two-dimensional plane. It looks something like this:

The equation of line is **y = mx + c.** One dimension is y-axis, another dimension is x-axis. It can be plotted in a two-dimensional plane. It looks something like this:



The generalization of this relationship can be expressed as:

y = f(x).

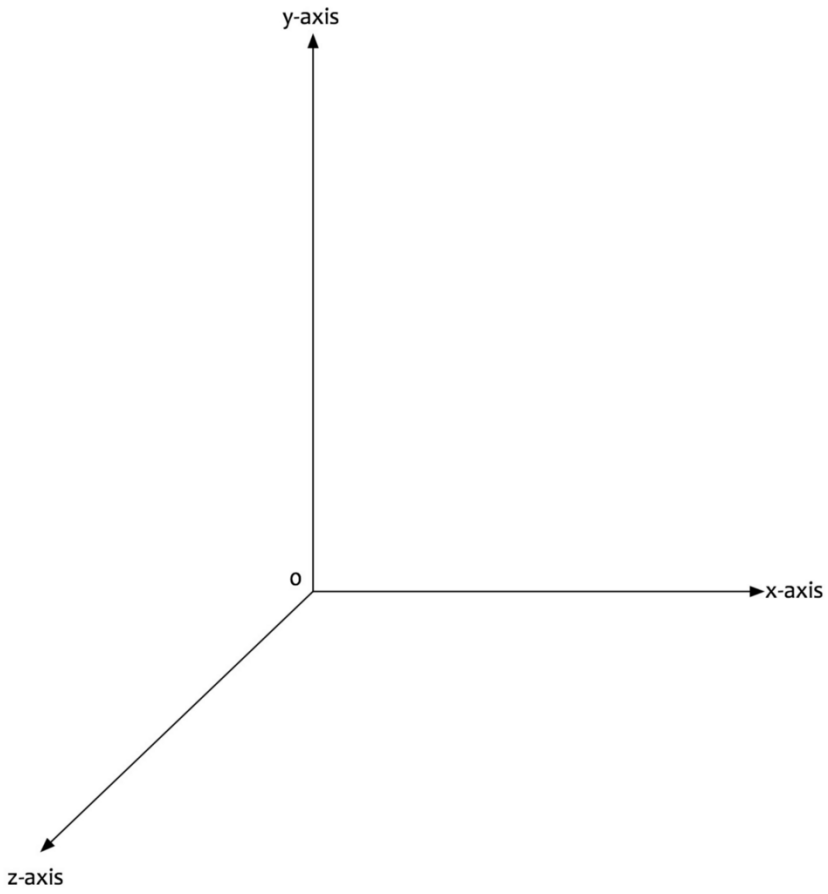
It doesn't mean anything fancy. All it means is:

Define y as a function of x. i.e. define the dependent variable as a function of the independent variable.

What if the dependent variable needs to be expressed in terms of more than one independent variable? The generalized function becomes:

y = f(x, z) *i.e. express y as some function/combination of x and z.*

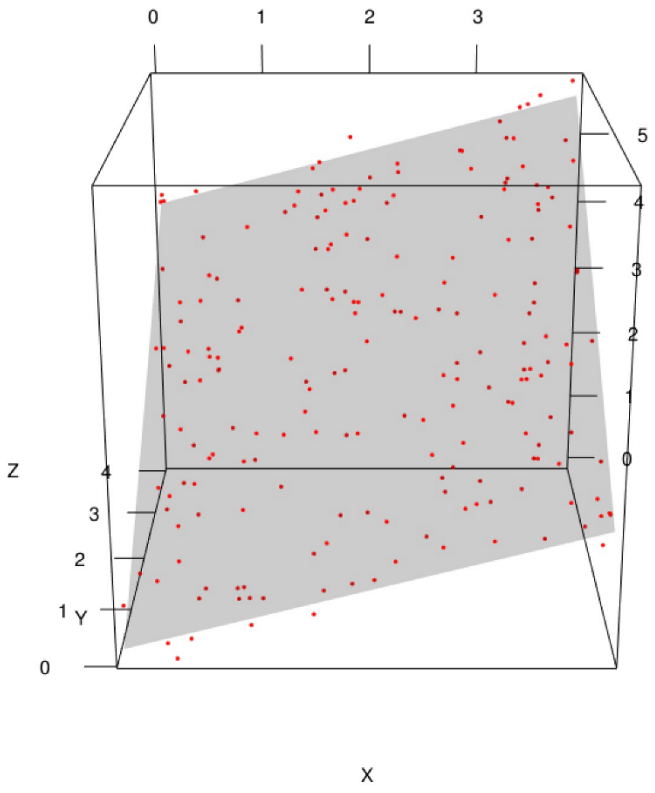
There are three dimensions now y-axis, x-axis and z-axis. It can be plotted as:



Now we have more than one dimension (x and z). We have an additional dimension. We want to express y as a combination of x and z.

For a simple regression linear model a **straight line** expresses y as a function of x. Now we have an additional dimension (z). What will

happen if an additional dimension is added to a line? **It becomes a plane.**



The plane is the function that expresses **y as a function of x and z.** Extrapolating the linear regression equation, it can now be expressed as:

y = m1.x + m2.z+ c

- **y** is the dependent variable i.e. the variable that needs to be estimated and predicted.
- **x** is the first independent variable i.e. the variable that is controllable. It is the first input.
- **m1** is the slope of x1. It determines what will be the angle of the line (x).
- **z** is the second independent variable i.e. the variable that is controllable. It is the second input.

- **m2** is the slope of z. It determines what will be the angle of the line (z).
- **c** is the intercept. A constant that determines the value of y when x and z are 0.

This is the genesis of the multivariate linear regression model. There are more than one input variables used to estimate the target. A model with two input variables can be expressed as:

$$y = \beta 0 + \beta 1.x1 + \beta 2.x2$$

Let us take it a step further. What if we had three variables as inputs? Human visualization capabilities are limited here. It can only visualize three dimensions. **In machine learning world, there can be many dimensions.** A model with three input variables can be expressed as:

$$y = \beta 0 + \beta 1.x1 + \beta 2.x2 + \beta 3.x3$$

A generalized equation for the multivariate regression model can be:

$$y = \beta 0 + \beta 1.x1 + \beta 2.x2 + \ldots\ldots + \beta n.xn$$

Model Formulation:

Now that there is familiarity with the concept of a multivariate linear regression model let us get back to Fernando.

Fernando reaches out to his friend for more data. He asks him to provide more data on other characteristics of the cars.

make	fuelType	nDoors	driveWheels	engineSize	horsePower	peakRpm	cityMpg	price
alfa-romero	gas	two	rwd	130	111	5000	21	13495
alfa-romero	gas	two	rwd	130	111	5000	21	16500
alfa-romero	gas	two	rwd	152	154	5000	19	16500
audi	gas	four	fwd	109	102	5500	24	13950
audi	gas	four	4wd	136	115	5500	18	17450
audi	gas	two	fwd	136	110	5500	19	15250
audi	gas	four	fwd	136	110	5500	19	17710
audi	gas	four	fwd	136	110	5500	19	18920
audi	gas	four	fwd	131	140	5500	17	23875
bmw	gas	two	rwd	108	101	5800	23	16430
bmw	gas	four	rwd	108	101	5800	23	16925
bmw	gas	two	rwd	164	121	4250	21	20970
bmw	gas	four	rwd	164	121	4250	21	21105
bmw	gas	four	rwd	164	121	4250	20	24565
bmw	gas	four	rwd	209	182	5400	16	30760
bmw	gas	two	rwd	209	182	5400	16	41315
bmw	gas	four	rwd	209	182	5400	15	36880

The following were the data points he already had:

- make: make of the car.
- fuelType: type of fuel used by the car.
- nDoor: number of doors.
- engineSize: size of the engine of the car.
- price: the price of the car.

He gets additional data points. They are:

- horsePower: horse power of the car.
- peakRPM: Revolutions per minute around peak power output.
- length: length of the car.
- width: width of the car.
- height: height of the car.

Fernando now wants to build a model that predicts the price based on the additional data points.

The multivariate regression model that he formulates is:

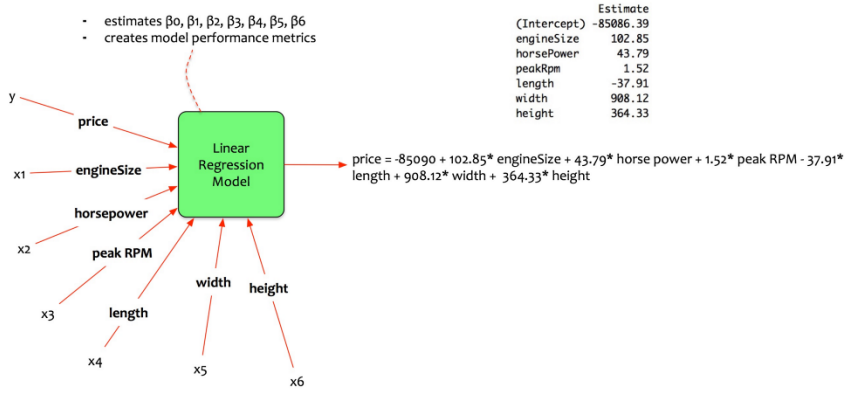
> *Estimate* ***price*** *as a function of* ***engine size, horse power, peakRPM, length, width and height.***

> ***=> price = f(engine size, horse power, peak RPM, length, width, height)***

> ***=> price = β0 + β1. engine size + β2.horse power + β3. peak RPM + β4.length+ β5.width + β6.height***

Model Building:

Fernando inputs these data into his statistical package. The package computes the parameters. The output is the following:



	Estimate
(Intercept)	-85086.39
engineSize	102.85
horsePower	43.79
peakRpm	1.52
length	-37.91
width	908.12
height	364.33

The multivariate linear regression model provides the following equation for the price estimation.

> ***price = -85090 + 102.85 * engineSize + 43.79 * horse power + 1.52 * peak RPM - 37.91 * length + 908.12 * width + 364.33 * height***

Model Interpretation:

The interpretation of multivariate model provides the impact of each independent variable on the dependent variable (target).

> *Remember, the equation provides an estimation of the* ***average value of price. Each coefficient is interpreted with all other predictors held constant.***

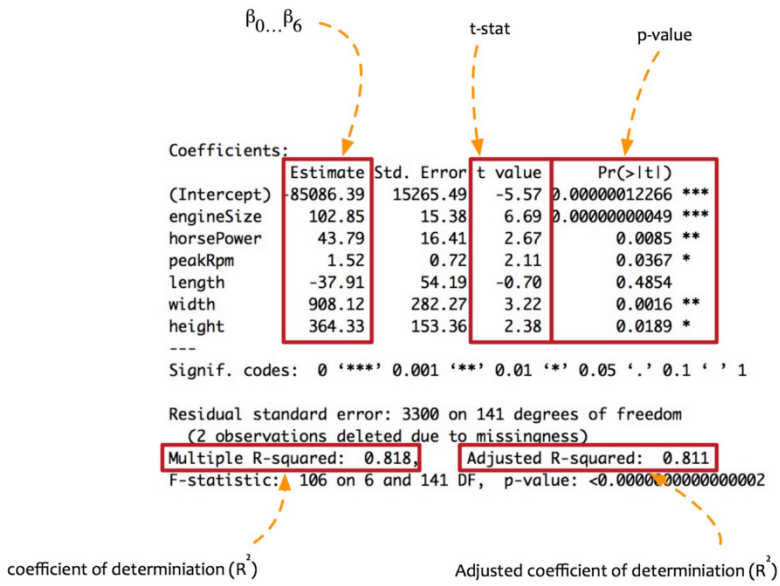
Let us now interpret the coefficients.

- Engine Size: With all other predictors held constant, if the engine size is increased by one unit, the average price **increases** by $102.85.
- Horse Power: With all other predictors held constant, if the horse power is increased by one unit, the average price **increases** by $43.79.
- Peak RPM: With all other predictors held constant, if the peak RPM is increased by one unit, the average price **increases** by $1.52.
- Length: With all other predictors held constant, if the length is increased by one unit, the average price **decreases** by $37.91 (length has a -ve coefficient).
- Width: With all other predictors held constant, if the width is increased by one unit, the average price **increases** by $908.12
- Height: With all other predictors held constant, if the height is increased by one unit, the average price **increases** by $364.33

Model Evaluation

The model is built. It is interpreted. Are all the coefficients important? Which ones are more significant? How much variation does the model explain?

The statistical package provides the metrics to evaluate the model. Let us evaluate the model now.



Recall the discussion on the definition of t-stat, p-value and coefficient of determination. Those concepts apply in multivariate regression models too. The evaluation of the model is as follows:

- **coefficients:** All coefficients are greater than zero. This implies that all variables have an impact on the average price.
- **t-value:** Except for length, t-value for all coefficients are significantly above zero. For length, the t-stat is -0.70. It implies that the length of the car may not have an impact on the average price.
- **p-value:** The probability of observing the p-value purely by chance is quite low for all of the variables except for length. The p-value for length is 0.4854. This implies that probability that the observed t-stat is by chance is 48.54%. This number is quite high.

Recall the discussion of how R-squared help to explain the variations in the model. When more variables are added to the model, the r-square will not decrease. It only increases. However, there has to be a balance. Adjusted R-squared strives to keep that balance. The **adjusted R-squared** is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared compensates for the addition of variables and only increases if the new term enhances the model.

- **Adjusted R-squared:** The r-squared value is 0.811. This implies that the model can explain 81.1% of variations seen in training data. It is better than the previous model (75.03%).

Based on these evaluations, Fernando concludes the following:

- All variables except for the *length* of the car has an impact on the price.
- The length of the car does not have the significant impact on price.
- The model explains 81.1% of the variation in data.

Conclusion:

Fernando has a better model now. However, he is perplexed. He knows that length of the car doesn't impact the price.

He wonders:

> *How can one select the best set of variables for model building? Is there any method to choose the best subsets of variables?*

In the next part of this series, we will discuss variable selection methods.