What is Regression Analysis

Edris Safari

Bellevue University, Nebraska U.S.A.

Abstract

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean)(Wikipedia). In the 20th century, “regression” is defined in very much the same way, but with statistical tools and methods. It has evolved into becoming a process whereby we can make predictions and choices with high confidence. We will examine the statistical concept of regression, and introduce tools and techniques used in regression analysis.

What is Regression Analysis

Introduction

The basic concept of regression is to estimate the relationship or correlation among variables. Much in the same was as Francis Galton's description, variables that relate to each other, tend to contribute to one another in positive or negative way. For example, tall fathers tend to have tall sons, or tall mother tend to have tall daughters. In these cases, the height of the ancestor can be used to reliably predict the height of the descendant. The variables involved in this use case are the parent's height, and the height of their descendants. And the data is based on a sample of a population which consists of a parent's height and their descendant's height. Since we want to estimate the descendant's height, we will call it the dependent variable. We will call the parent's height the independent variable or the predictor since it will be used to predict the descendant's height. The relationship between these two variables can be shown with the classic linear equation:

$$Y = mX + b$$

Where X is the independent variable and Y is the dependent variable. If we were to graph this equation, m would be the slop of the linear line and b the intercept. In statistics, and particularly in regression analysis, X and Y are a set of numbers representing a distribution (i.e. data gathered from 1000 males. In this case, the equation above becomes:

$$Y_i = m\, X_i + b + \varepsilon$$

, where the subscript 'i' denotes the data for 'ith' respondent and ε is error. The graph of this equation is the all too familiar linear regression graph shown in Figure 1. The blue line inside the gray envelope is the best-fit line such that the residuals are minimized and the dots around it are the actual observations obtained by say a survey-it's in fact the visualization of the data set, but only two columns of it.

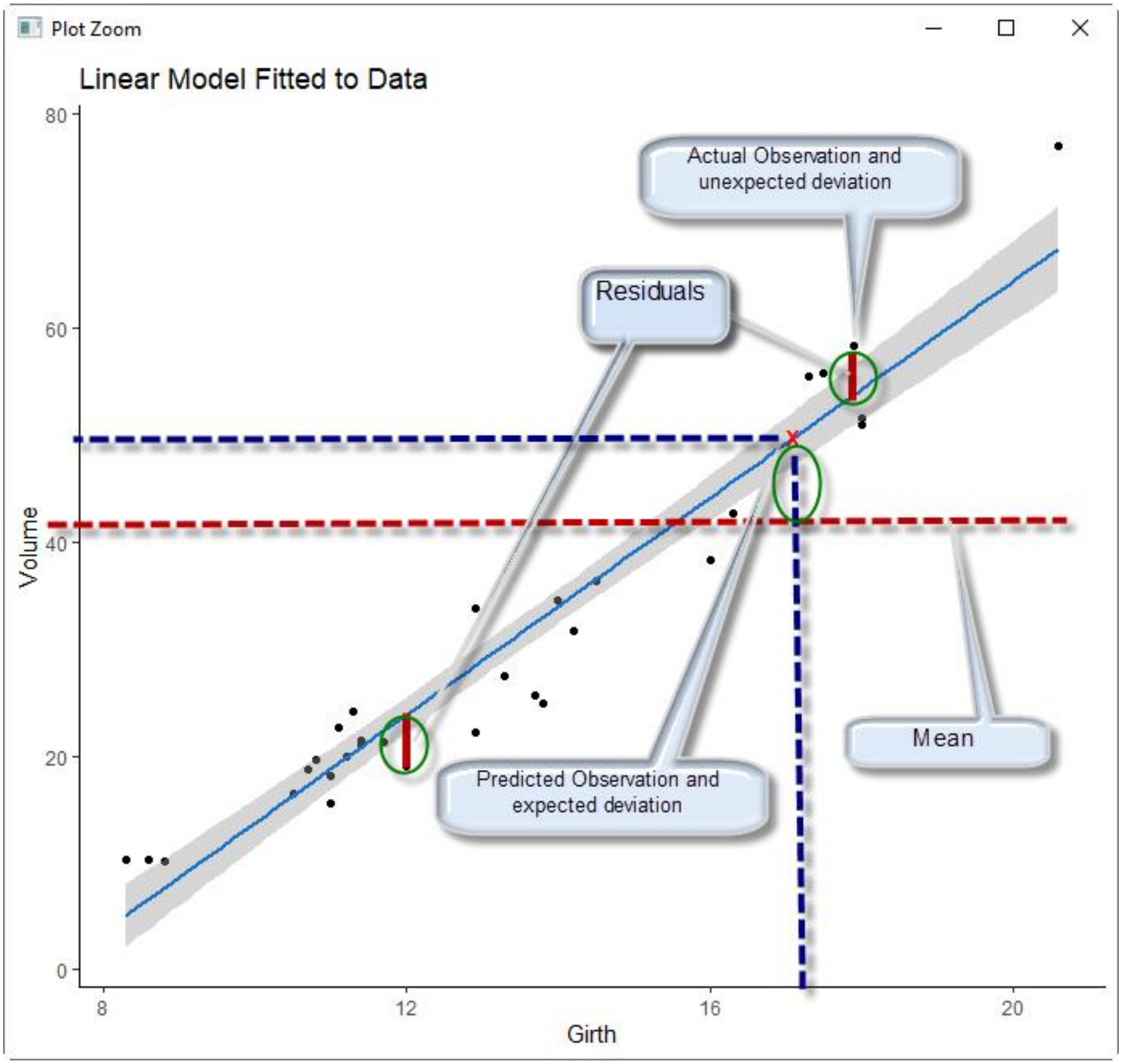

Figure 1- Graph of a linear equation

The minimization of the residuals and thus coming up with the best fit line is the process of regression. We will explain this process in the following section. We will then describe the process of regression when dealing with multiple independent variables that may or may not help us predict the dependent variable.

Simple Linear Regression

The annotations in Figure 1, describe the process. To minimize the residuals, we would want to add up all the residuals and determine from its magnitude whether the line is a bet fit. This calculation results in zero. On the other hand, if the line is indeed a good fit, the dots would be much closer (almost on the line). In this case the values are normally distributed which means the prediction will be more accurate (when we draw a line from the horizontal line to intersect the line). Also shown in Figure 1, the mean value of the dependent variable play an important role in the minimization process. Let's call the horizontal line (Girth) in Figure 1, **X** and the vertical line (Volume), **Y** and describe the process in mathematical terms.

$$expected\ deviation\ = \tilde{Y}_i\ -\ \bar{Y}$$

$\tilde{Y}_i\ is\ the\ estimated\ or\ predicted\ value\ of\ the\ dependent\ variable\ based\ on\ X_i\ and\ \bar{Y}$ is the mean value of the dependent variable.

$$unexpected\ deviation\ =\ e_i\ = Y_i\ -\ \tilde{Y}_i$$

From these deviations we can compute the sum of squared residuals for both expected and unexpected residuals as follows:

$$SSR = \sum \left(\tilde{Y}_i - \bar{Y}\right)^2\ ;\ this\ is\ sum\ of\ expected\ residuals\ to\ to\ regressiob$$

$$SSE = \sum \left(Y_i - \tilde{Y}_i\right)^2\ ;\ this\ is\ sum\ of\ unexpected\ residuals\ to\ to\ errors$$

, and the total sum of squares is:

$$SST\ =\ SSR\ +\ SSE\ =\ \sum (Y_i -\ \bar{Y})^2\ ;\ this\ is\ total\ sum\ of\ sqaures$$

For the model to fit perfectly SSR and SST must be equal and thus their ratio must equal to 1. This is denoted by R^2as follows:

$$R^2 = SSR/SST = \sum\left(\tilde{Y}_i - \bar{Y}\right)^2 / \sum(Y_i - \bar{Y})^2$$

If the observations are scattered such that SSE is high and SSR is low, R2 is lower and thus a bad fit(X cannot reliably predict Y). On the other hand, if SSE is low and SSR is high, R2 is high, a better fit and X can(possibly) reliably predict Y. $R^2$is the proportion of the variation in the dependent variable that is being explained by the independent variable. Note that as the number of points (or observations) increase R^2(whose value by the way goes from 0 to 1) approaches 1. However, this does not mean that X and Y are related. It may mean that X is significant at best. This is remedied by the adjusted R^2 which is calculated using R^2 and the degree of freedom which is $df = n - k - 1$ where n is the number of independent variables(in this case 1 but greater than 1 for multiple regression and k is the number of observations. The equation for adjusted R2 is shown below:

$$\bar{R}^2 = 1 - (1 - R^2)n - 1/(n - k - 1)$$

In this formula, as k increases and thus df decreases, $\bar{R}^2$ decreases. This is when useless variables are included. $\bar{R}^2$ increases when useful variables are included. While it seems $\bar{R}^2$is the statistic to rely on, it is not. It is one statistic used in hypothesis testing. We will describe them after introduction to multiple linear regression.

Multiple Linear Regression

As the name implies multiple linear regression deals with multiple variables-namely multiple independent variables, but still one dependent variable. The equation would change to:

$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + B_3X_{3i} + \ldots + + B_nX_{ni} + \varepsilon_i$$

With two independent variables shown in Figure 2 , we can try to fit a plane (as opposed to line) into the model and calculate $\bar{R}^2$ or we can use the 'ols' method from the python 'statsmoddel' package.

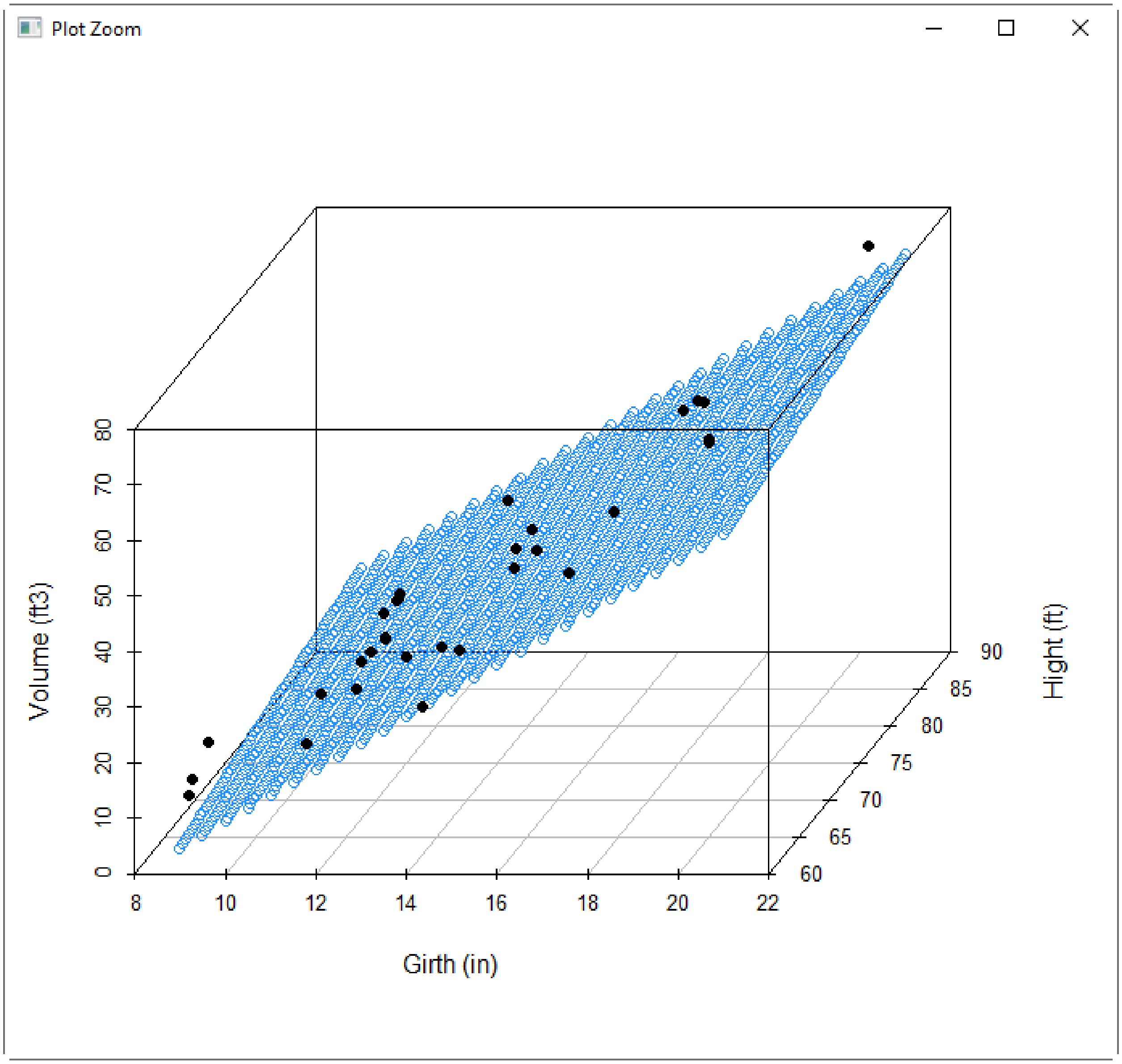

Figure 2- Multiple linear regression

The call to the old method is as follows:

Result = smf.ols(formula, data=dataset)

The formula is written as ‘dependent variable’ ~ ‘independent variable (1)’ + ‘independent variable (2)’ + .. + ‘independent variable (n)’. The result will include a lot of information. Included with the $\bar{R}^2$, the coefficients $B_1, B_2, \ldots, B_n$ will also be provided because the method tries to find the best fit. Other values such as p-value t-stat, f-state and standard error are provided as well. When multiple variables are introduced, issues related of multicollinearity must be addressed. Multicollinearity occurs when the independent variables themselves are related and either don’t contribute or contribute inaccurately. Examination of the coefficients (B values) and the variances of the variables are good indicators from the result of the ols method.

Conclusion

Examination of all the parameters and statistics that are part of linear regression is part of the hypothesis testing and the overall exploration. The outcome of the regression analysis is a set of independent variables that are deemed significant and reliable predictors of the dependent variable. Once the independent variables are selected the model is run through several samples and predictions made. The sample results are then compared and analyzed until validated and the model deemed production worthy.

References

https://en.wikipedia.org/wiki/Regression_analysis

Downey, Allen B.. Think Stats: Exploratory Data Analysis . O'Reilly Media. Kindle Edition.

Field, Andy. Discovering Statistics Using R (p. 318). SAGE Publications. Kindle Edition.

Regression I: What is regression? | SSE, SSR, SST | R-squared | Errors (e vs. e) - https://www.youtube.com/watch?v=aq8VU5KLmkY

Regression II: Degrees of Freedom EXPLAINED | Adjusted R-Squared - https://www.youtube.com/watch?v=4otEcA3gjLk

Regression III: Understanding regression output - https://www.youtube.com/watch?v=VvlqA-iO2HA&t=17s

ANOVA: One-way analysis of variance - https://www.youtube.com/watch?v=9cnSWads6oo