Linear Regression Analysis

B.M Njuguna

2022-09-02

Contents

1.0 Introduction

In some cases, the relationship between the response variable y and the independent or predictor variable or variables might not be linear. In a such a case, we cannot apply the linear regression analysis as the assumption of linearity is violated. Thus, the **Polynomial Regression** is a type of regression whereby the relationship between the response variable and the predictor variables is modeled as the n^{th} degree polynomial. The polynomial regression fits a non-linear relationship between the values of the independent variable X and the conditional mean of y which is denoted as $\mathbf{E}(y|x)$. Although the polynomial regression fits a nonlinear model to the data, as a statistical estimation problem it is linear in the sense that the conditional mean of y i.e $\mathbf{E}(y|x)$ is linear to the unknown to the parameters estimated from the data, hence it is referred to as a special case of multilple linear regression.

With polynomial regression, data is approximated using a polynomial equation of degree n written as;

$$\mathbf{f}(\mathbf{x}) = \alpha_0 + \alpha_1 x + \alpha_2 x_1 + \cdots + \alpha_n x^n$$

where α is the set of coefficients.

Now the polynomial regression equation can be written as;

$$y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_n x_i^m + \epsilon_i fori = 1, 2, 3, \ldots, n$$

The above model can be expressed matrix form in terms of a design matrix $\curvearrowleft$, response vector $\vec{y}$, the parameter vector $\vec{\beta}$ and the random vector $\vec{\epsilon}$, as follows;

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_2^2 & \ldots & x_n^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ 1 & x_3 & x_3^2 & \ldots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \\ \vdots \\ \beta_m \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Which can be written as;

$$\vec{y} = \mathbf{x}\vec{\beta} + \vec{\epsilon}$$

The polynomial coefficients β can be estimated using the **ordinary Least Square** as follows;

$$\vec{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}\vec{y}$$

It is often difficult to interpret individual polynomial regression coefficients since the underlying monomials are highly correlated for example if x is **Uniformly distributed**,then x and x^2 have a high correlation of 0.97. Therefore, it is generally informative to consider the fitted regression function as a whole.[1]

In r, we use the function **poly()** which is in the basic syntax to fit a polynomial regression model to data.

[1] monimal *in plural monomials* is an algebraic expression consisting of one term.