Linear Regression

Intro

Linear regression is an algorithm used to determine the relationship between a continuous dependent variable and one or more independent variables. This is done through fitting the line of best fit which uses beta coefficients that minimize the sum of squared residuals.

Definitions

Beta Coefficient: the degree of change in the outcome variable for each unit of change in the predictor variable(s)

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

where β_0 is the intercept, β_n is the coefficient, and X_n is the independent variable

Residuals: the difference between the predicted value of a dependent variable and the actual value of that dependent variable (loss function)

$$(y - \hat{y})$$

where y is the actual value and $\hat{y}$ is the predicted value

Sum of Squared Residuals: the sum of all residuals squared (cost function)

$$RSS = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

Derivation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Solving for ε_i:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$$

letting E stand for the sum of squared errors:

$$E = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

To find the line of "best fit", we need to find a pair of $(\beta_0 \,\&\, \beta_1)$ which will minimize E

Solving for β_1's: derivative

$$\frac{\partial E}{\partial \beta_1} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1)(x_i)$$

$$0 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(x_i)$$

$$0 = \sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2$$

Let $v = \sum_{i=1}^{n} x_i$ and $u = \sum_{i=1}^{n} x_i^2$

(1) $$0 = \sum_{i=1}^{n} y_i x_i - \beta_0 v - \beta_1 u$$

Solving for β_0's: derivative

$$\frac{\partial E}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} x_i$$

Let $\sum_{i=1}^{n} 1 = n$ and $v = \sum_{i=1}^{n} x_i$

(2) $$0 = \sum_{i=1}^{n} y_i - n\beta_0 - v\beta_1$$

Using equations ② and ①

$$② \quad \sum_{i=1}^{n} y_i = n\beta_0 + v\beta_1$$

$$① \quad \sum_{i=1}^{n} y_i x_i = v\beta_0 + U\beta_1$$

If we multiply equation ② by v, and equation ① by n, we can isolate β_1

$$v \sum_{i=1}^{n} y_i = nv\beta_0 + v^2\beta_1$$

$$n \sum_{i=1}^{n} y_i x_i = nv\beta_0 + nU\beta_1$$

Subtracting results in:

$$v \sum_{i=1}^{n} - n \sum_{i=1}^{n} y_i x_i = v^2\beta_1 - nU\beta_1$$

$$\frac{v \sum_{i=1}^{n} y_i - n \sum y_i x_i}{(v^2 - nU)} = \beta_1 \quad ③$$

Using equation ② we can substitute equation ③ for β_1, and solve for β_0

$$\beta_0 = \frac{\sum_{i=1}^{n} y}{n} - \frac{v\beta_1}{n}$$

$$\beta_0 = \frac{\sum_{i=1}^{n} y}{n} - \frac{v\left(v\sum_{i=1}^{n} y - n\sum_{i=1}^{n} yx\right)}{n(v^2 - nU)}$$

Simplifying and reducing results in:

$$\beta_0 = v^2\sum_{i=1}^{n} y - nu\sum_{i=1}^{n} y - v^2\sum_{i=1}^{n} y + vn\sum_{i=1}^{n} y_i x_i$$

$$\beta_0 = \frac{-nu\sum_{i=1}^{n} y + vn\sum yx}{nv^2 - nu} \quad (4)$$

β_0 can be re-written using prior definitions of v and u :

$$\beta_0 = \frac{-n \sum_{i=1}^{n} x_i^2 \sum y + \sum_{i=1}^{n} x_i \, n \sum_{i=1}^{n} y_i x_i}{n\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2}$$

Multiplication of $-1/n^2$ to numerator and denominator can further reduce the equation

$$\beta_0 = \frac{\sum_{i=1}^{n} x_i^2 \, \bar{y}_n - \sum_{i=1}^{n} y_i x_i (\bar{x}_n)}{\sum x_i^2 (1/n) - \bar{x}\left(\sum_{i=1}^{n} x_i\right)}$$

Following the same steps above for β_1 :

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - n \sum_{i=1}^{n} y_i x_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2}$$

Using $1/n^2$ rather than $-1/n^2$:

$$\beta_1 = \frac{\bar{x}_n \bar{y}_n - \left(\sum_{i=1}^{n} y_i x_i\right)(1/n)}{\bar{x}_n^2 - \left(\sum_{i=1}^{n} x^2\right)(1/n)}$$