Simple Linear Regression

- simple approach to predict quantitative value (response) Y on the basis of a predictor variable (X)

$$\hat{Y} \approx \hat{\beta}_0 + \hat{\beta}_1 X \quad \text{(prediction)}$$

- We intend to find / estimate the coefficient, $\hat{\beta}_0$ and $\hat{\beta}_1$ so that we can represent the patter emerged by the data, as close as possible.

We use "MSE" to measure "closeness" / accuracy of the model.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, represent the prediction for Y based on the ith value of X.

We define

$e_i = y_i - \hat{y}_i$ to be the i th residual

We define,

$$RSS = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n$$

(Residual-sum of squares)

Derivation

$$L = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\frac{\delta L}{\delta \hat{\beta}_0} = \frac{\delta}{\delta b_0} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{\delta}{\delta b} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

$$= \frac{1}{n} \sum_{i=1}^{n} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$= \sum_{i=1}^{n} \frac{y_i}{n} - \sum_{i=1}^{n} \frac{\hat{\beta}_1 x_i}{n} - \sum \frac{\hat{\beta}_0}{n} = 0$$

$$= \boxed{\bar{Y} - \hat{\beta}_1 \bar{X} = \frac{\cancel{n}\beta_0}{\cancel{n}} = \beta_0}$$

$$\frac{\delta L}{\delta \beta_1} = \frac{\delta}{\delta \beta_1} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0$$

$\longrightarrow Y_i - \bar{Y} + \beta_1 \bar{X} - \hat{\beta}_1 x_i$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{\delta}{\delta \beta_1} (Y_i - (\bar{Y} - \beta_1 \bar{X}) - \hat{\beta}_1 X_i)^2 = 0$$

$$= \frac{1}{n} \sum_{i=1}^{n} 2(Y_i - (\bar{Y} - \beta_1 \bar{X}) - \hat{\beta}_1 X_i) . (-X_i + \bar{X}) = 0$$

$$= \sum_{i=1}^{n} 2 (Y_i + \hat{\beta}_1 \bar{X} - \bar{Y} - \hat{\beta}_1 X_i) . (\bar{X} - X_i) = 0$$

$$= \sum [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})] (X_i - \bar{X}) = 0$$

$$= \sum (Y_i - \bar{Y})(X_i - \bar{X}) = \hat{\beta}_1 \sum (X_i - \bar{X})^2$$

$$\Rightarrow \boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

→ β_0 is the value of Y when X = 0

→ β_1 represents the increase in Y associated w/ one unit increase in X.

The concept of bias (Sample variance)

→ An unbiased estimator is one that, on average, equal to the true population estimate.

For eg → an unbiased mean $\hat{\mu}$ of the sample would :

- In one sample – overestimated.
- In one sample – understimated.

Basically, the sample estimate might not be exact in terms of the population, but it will not systematically overestimate or

Standard Error ($Var(\hat{\mu})$)

→ We know that averages of $\hat{\mu}$ over different samples will be very close to μ. However, we don't know how [illegible] be a single estimat $\hat{y}$. To do this, we use standard Error

$$Var(\hat{\mu}) = SE(\hat{y})^2 = \frac{\sigma^2}{n}$$

$$\Rightarrow \boxed{SE(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}} \longrightarrow \text{standard deviation}$$

Similarly, for regression coefficients

$$SE(\hat{\beta}_0)^2 = \underbrace{\sigma^2}_{\text{variance of errors.}} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$\sigma^2 = \sqrt{(RSS) / (n-2)}$

Residual Standard Error

RECAP

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}}$$

Standard Error

$$SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

$$Var(\varepsilon) = \frac{\sum_{i=1}^{n} \varepsilon_i}{n-2}$$

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

→ Assumption : $\varepsilon \sim N(0, \sigma^2)$

This implies that for a given value of $x = x_i$, the value of y follows $N \sim (\hat{\beta}_0 + \hat{\beta}_1 x_i, \sigma^2)$

We use standard error to estimate the confidence intervals.

For LR :

$\boxed{\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)}$ ⇒ there is 95% chance that the interval $[\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)]$ will contain the true value of the parameter.

Hypothesis Testing

and the outcome

Null Hypothesis : Typically assumes no relation b/w any of the features, & For LR (single feature) :

H_0 : There's no relationship b/w X and Y i.e $\beta_1 = 0$

Alternate Hypothesis : Assumes there's some relationship b/w X and Y.

H_a : $\beta_1 \neq 0$

To test the hypothesis we calculate t-statistic

$$\downarrow t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)\uparrow}$$ $\longrightarrow$ represents the difference b/w the estimated slope and the hypothesized value of $\hat{\beta}_1$

the larger the abs. value of t, the more evidence we have against H_0. If $SE(\hat{\beta}_1)\uparrow \Rightarrow t\downarrow$, giving evidence for H_0.

Accuracy of the model

RSE
(Residual Standard Error)

$$RSE = \sqrt{\frac{1}{n-2} . RSS}$$

$$RSE = \sqrt{\frac{1}{n-2} . \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

R^2
This is b/w 0 and 1, and independent of the scale of Y.

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \text{ (variance)}$$

Created with Notewise

3.2 Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

$$RSS = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon)\right)$$

1. Is there a relationship b/w the Response and Predictors?

Null Hypothesis (H_0) : $\hat{\beta}_1 = \hat{\beta}_2 = \cdots = \beta_p = 0$
(There's no relationship b/w predictors and response.

Alternate (H_a): atleast one β_i is non-zero ; $1 \leq i \leq p$

Bias - Variance TradeoFF

While its important to minimise training MSE, it's equally important to get a minimized MSE on test points.

$$E\left(y_0 - f(x_0)\right)^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))] + Var(\varepsilon)$$
(Exp. test MSE)

Variance refers to the amount by which $\hat{f}$ would change if we estimated it using different sets. Essentially, if a method has higher variance, it would mean that a small change in the data can get large changes in f.

Bias refers to the inability of the model to accurately represent the patterns in the dataset.

→ easy to obtaing low bias and higher variance OR high variance and low bias

→ The challenge is to find an ideal midway b/w the two, which should ensure low bias and variance.