Start-Tech Academy

Simple Linear Regression

Simple linear regression is an approach for predicting a quantitative response Y on the basis of a single predictor variable X. It assumes that there is approximately a linear relationship between X and Y .

Introduction

Model Equation

$$\boldsymbol{Y} \approx \beta_0 + \beta_1 X$$

β_0 is known as Intercept

β_1 is known as slope

Together β_0 and β_1 known as the model *coefficients* or *parameters*.

For House Price data

- X will represent Room_num
- Y will represent Price

$$\text{Price} \approx \beta_0 + \beta_1 \times \text{Room\_num}$$

From our training data we will get $\hat{\beta}_0$ **and** $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$

Start-Tech
ACADEMY

Simple Linear Regression

Estimating the Coefficients

- Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_{\mathbf{1}}$ such that the linear model fits the available data well
- Total number of rows (Data Point) $\Rightarrow$ n = 506
- Data $\Rightarrow$ $(x_1, y_1), (x_2, y_2), (x_3, y_3),$ …………………(x_{506}, y_{506})
- Lets call calculated y value as $\hat{y}$

$$\widehat{y_1} = \hat{\beta}_0 + \hat{\beta}_{\mathbf{1}} x_1$$
$$\widehat{y_2} = \hat{\beta}_0 + \hat{\beta}_{\mathbf{1}} x_2$$
$$\widehat{y_{506}} = \hat{\beta}_0 + \hat{\beta}_{\mathbf{1}} x_{506}$$

- The difference between residual the *i*th observed response value and the *i*th response value that is predicted by our linear model is known as residual

$$e_i = y_i - \hat{y}_i$$

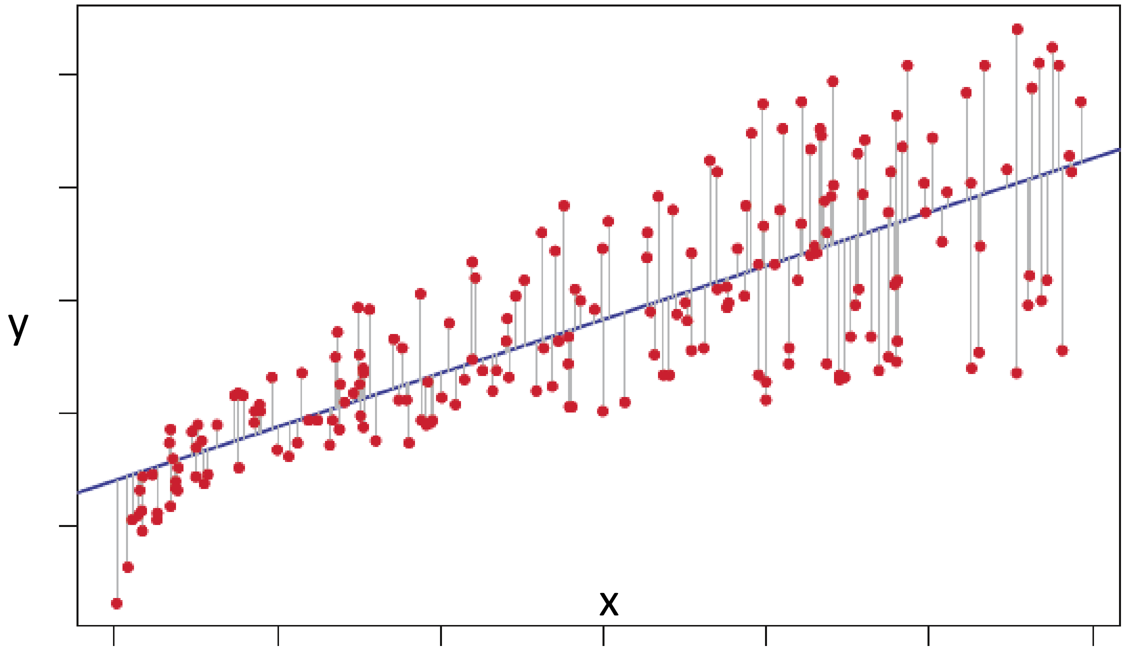
Simple Linear Regression

Residual

Residual –

The difference between residual the *i*th observed response value and the *i*th response value that is predicted by our linear model is known as residual

$$e_i = y_i - \hat{y}_i$$



Simple Linear Regression

RSS

Residual sum of squares **(RSS)**

$$RSS = e_1^2 + e_2^2 \ldots\ldots. + e_n^2$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_{\mathbf{1}}$ to minimize the RSS

Using some calculus, one can show that the minimizers are

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

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Simple Linear Regression

Model

For our Model

```
Residuals:
    Min      1Q  Median      3Q     Max
-23.336  -2.425   0.093   2.918  39.434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -34.6592     2.6421  -13.12   <2e-16 ***
room_num      9.0997     0.4178   21.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.597 on 504 degrees of freedom
Multiple R-squared:  0.4848,    Adjusted R-squared:  0.4838
F-statistic: 474.3 on 1 and 504 DF,  p-value: < 2.2e-16
```