STATISTICS FOR DATA SCIENCE

Satya Vani NL
Department of Science & Humanities

STATISTICS FOR DATA SCIENCE

Uncertainties in Least Squares Coefficients
The More Spread in the x Values , the Better

Satya Vani NL
Department of Science & Humanities

Consider Bivariate data $(x_{i,} y_i)$ for i=1,2,3,…………….n

$$y = \beta_0 + \beta_1 x$$

The line $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i is the error, that best fits the data in the sense of minimizing the sum of the squared errors. It is called the least squares regression line

$\widehat{\beta_0}$, $\widehat{\beta_1}$ are estimates of β_0 , β_1 .

$\widehat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$ = Fitted line

If ε_i tend to be large then $(x_{i,} y_i)$ are widely scattered around the line.

If ε_i tend to be small then $(x_{i,} y_i)$ are tightly clustered around the line.

$\widehat{\beta_0}$, $\widehat{\beta_1}$ are called Least Squares Coefficients and defined as

$$\widehat{\beta_1} = \sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{\sum_{i-1}^{n}(x_i - \overline{x})^2}\right] y_i$$

$$\widehat{\beta_0} = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\bar{x}\,(x_i - \bar{x})}{\sum_{i-1}^{n}(x_i - \overline{x})^2}\right] y_i$$

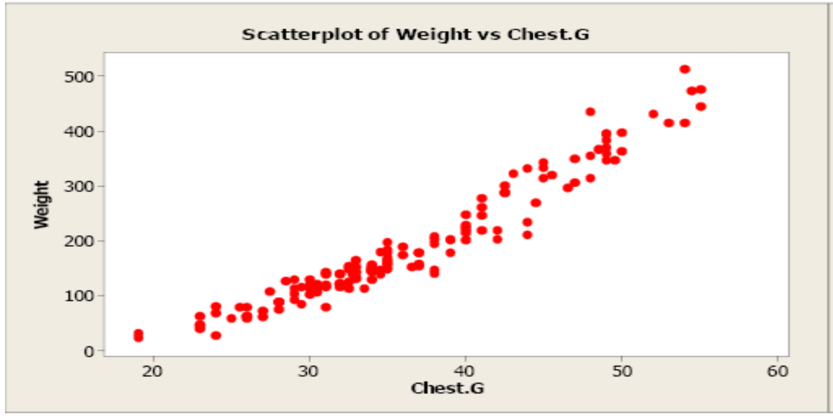
This indicates that $\widehat{\beta_0}$, $\widehat{\beta_1}$ are linear combination of y_i .

$$s_{\widehat{\beta_0}} = s\sqrt{\left[\frac{1}{n} + \frac{\overline{x}^{\,2}}{\sum_{i-1}^{n}(x_i - \overline{x})^2}\right]} \qquad s_{\widehat{\beta_1}} = s\sqrt{\left[\frac{1}{\sum_{i-1}^{n}(x_i - \overline{x})^2}\right]}$$

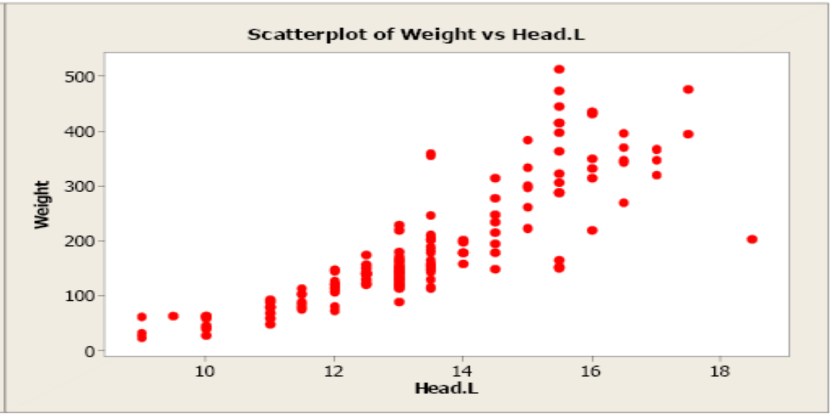
$$s_{\widehat{\beta_1}} \;\; \alpha \;\; \frac{1}{\sum_{i-1}^{n}(x_i - \overline{x})^2}$$

If x – values are more spread then the uncertainty of estimates $\widehat{\beta_0}$, $\widehat{\beta_1}$ are Smaller.

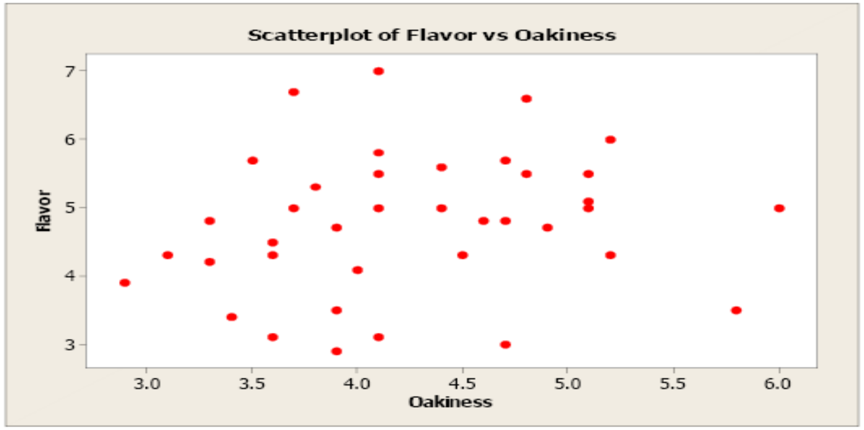
The standard deviation of x is more.



Strong positive relationship
r = 0.96



Moderate positive relationship
r = 0.67



Very weak positive relationship
r = 0.07

Problem: Two engineers are conducting independent experiments to estimate spring constant for a particular spring. The first engineer suggests measuring the length of the spring with no load, then applying loads of 0,1,2,3,& 4 lb. The second engineer suggests using loads of 0, 2, 4, 6 & 8 lb. Which will be more precise?

Sol: X ----- 0, 1, 2, 3, 4 Y-------- 0, 2, 4, 6, 8

σ_y is twice as great as σ_x .

Uncertainty of X is twice as large as the uncertainty of Y.

Hence, the Engineer, Y 's estimate is twice as precise.

THANK YOU

Satya Vani NL

Department of Science & Humanities
sathyavaninl@pes.edu

+91 80 66186410