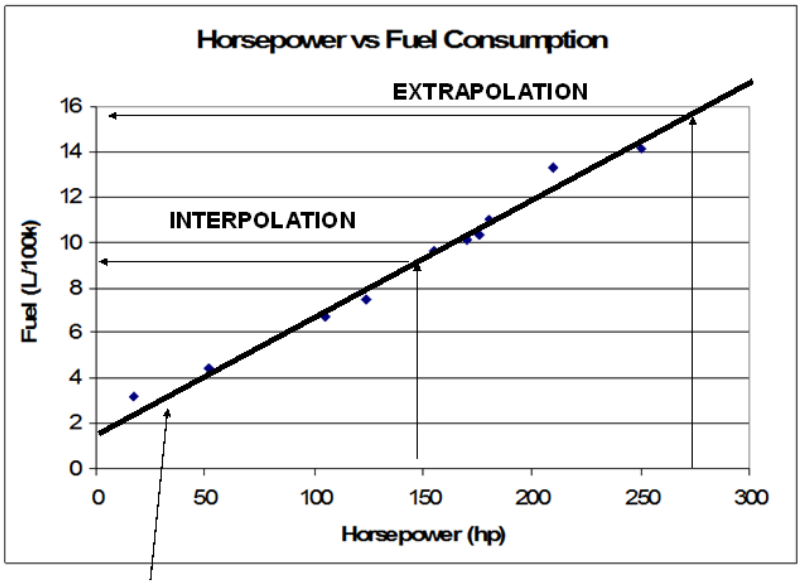
3.2 LINEAR REGRESSION

- Regression is an analytical technique for modeling the relationship between two variables.
- When the relationship between the two variables is **linear**, a simple mathematical model can be developed by finding **the line of best fit**.
- The equation of this line of best fit can be used to make predictions by **Interpolation** (estimates within the given data range) and **Extrapolation** (estimates outside of the given data range)

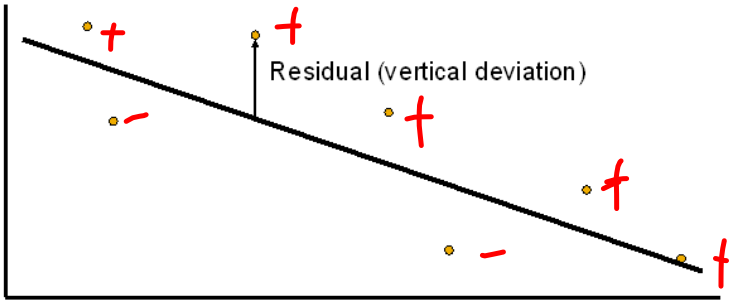
3.2 LINEAR REGRESSION



LINE OF BEST FIT: $y = 20x + 1.75$

3.2 LINEAR REGRESSION

- An analytic method for determining the line of best fit can be determined by the ***method of least squares***
- In order to understand the least squares method, we must first define the term ***residual***
- In regression analysis, a ***residual*** is defined to be the vertical distance from a particular data point to the line of best fit

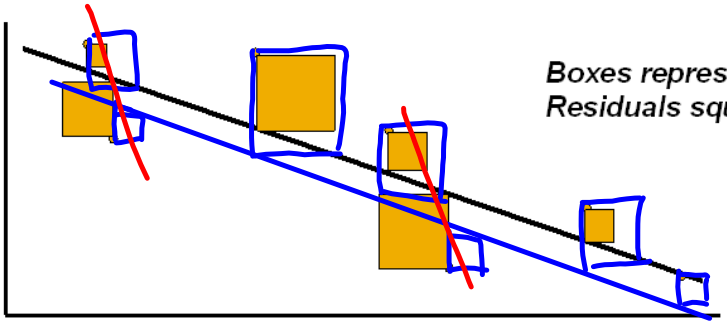


3.2 LINEAR REGRESSION

For the line of best fit in the ***method of least squares***

•The sum of the residuals is zero (sum of the distance above the line is equal to the sum of the distance below the line)

•The sum of the squares of the residuals has the least possible value. (Boxes shown below are the smallest possible)



Boxes represent the Residuals squared

3.2 LINEAR REGRESSION

Statisticians have developed the following formula to determine the equation of the line of best fit using the least squares method

The equation of a line is given by: $y = ax + b$ $y = mx + b$

WHERE... m $a = \dfrac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$ (slope)

and

$b = \bar{y} - a\bar{x}$ or $b = \dfrac{\sum y - a\sum x}{n}$ (m)

$\bar{y} = \dfrac{\sum y}{n}$ $\bar{x} = \dfrac{\sum x}{n}$ $b = \dfrac{\sum y}{n} - m\dfrac{\sum x}{n}$

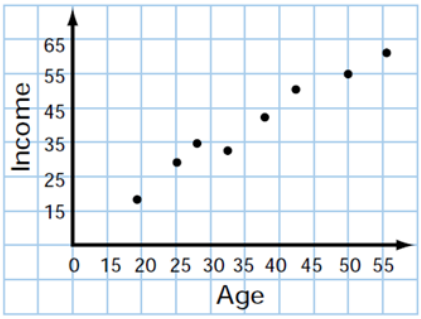
$= \dfrac{\sum y - m\sum x}{n}$

3.2 LINEAR REGRESSION

The table and scatter plot show data for the full-time employees of a company:

x y x^2 xy

Age (years)	Annual Income ($000)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35



3.2 LINEAR REGRESSION

In order to calculate the line of best fit using the least squares method, the following table and calculations are set up.

Age, x	Income, y	x^2	xy
33	33	1089	1089
25	31	625	775
19	18	361	342
44	52	1936	2288
50	56	2500	2800
54	60	2916	3240
38	44	1444	1672
29	35	841	1015
$\sum x = 292$	$\sum y = 329$	$\sum x^2 = 11\,712$	$\sum xy = 13\,221$

$(\sum x)^2$

$$\bar{x} = \frac{\sum x}{n} \qquad \bar{y} = \frac{\sum y}{n}$$

$$a = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$
$$= \frac{8(13\ 221) - (292)(329)}{8(11\ 712) - (292)^2}$$
$$= \frac{9700}{8432}$$
$$\doteq 1.15$$

$$\bar{x} = \frac{\Sigma x}{n} \qquad \bar{y} = \frac{\Sigma y}{n} \qquad b = \bar{y} - a\bar{x}$$
$$= \frac{292}{8} \qquad = \frac{329}{8} \qquad = 41.125 - 1.15(36.5)$$
$$= -0.85$$
$$= 36.5 \qquad = 41.125$$

$$y = ax + b$$
$$= 1.15x - 0.85$$

Therefore, the equation of the line of best fit is $y = 1.15x - 0.85$.

3.2 LINEAR REGRESSION

- The slope a indicates only how y varies with x on the line of best fit
- The slope a does NOT tell anything about the strength of the correlation between the two variables (the correlation coefficient r does)
- It is possible to have a weak correlation with a large slope or a strong correlation with a small slope

3.2 LINEAR REGRESSION

Outliers

•an observation that is numerically distant from the rest of the data

Why Do We Have Outliers?

•Measurement Error
•Miscoding
•Misinterpretation
•Entered incorrectly
•Relationship is non-linear
•etc.

3.2 LINEAR REGRESSION

How to deal with outliers

•Most common method is to ignore outliers or not even consider them when doing regression analysis

•May make sense to leave outliers in dataset if they make a meaningful cluster and/or there are many of them

•Apply a different regression model (non-linear)

Good rule of thumb:

When in doubt, present results with and without outliers.