

# Fitting a Line, Residuals, and Correlation

Colby Community College

## Definition

**Linear Regression** is the statistical method for fitting a line to data where the relationship between two variables,  $x$  and  $y$ , can be modeled by a straight line with some error:

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

## Note

We don't use  $y = mx + b$  because the format  $y = b_0 + b_1 x$  can easily be expanded to include more variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots$$

This is used when performing a multiple regression.

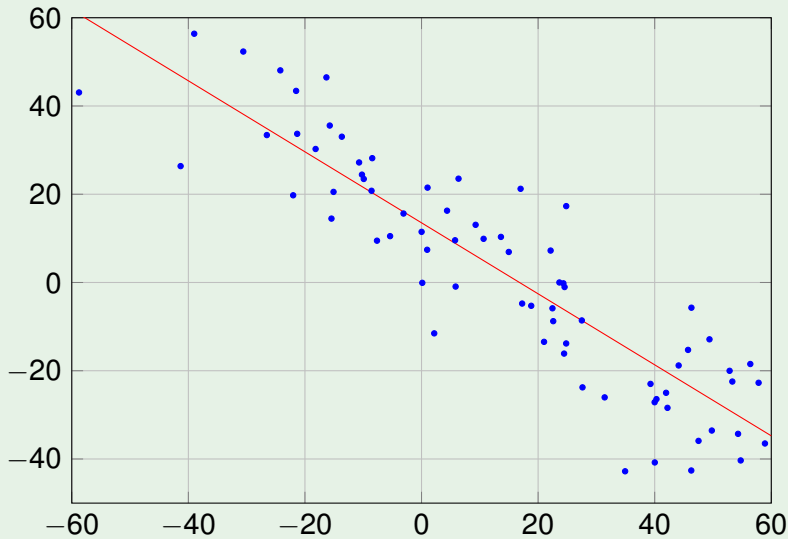
## Definition

We call  $x$  the **explanatory variable** or **predictor variable**.

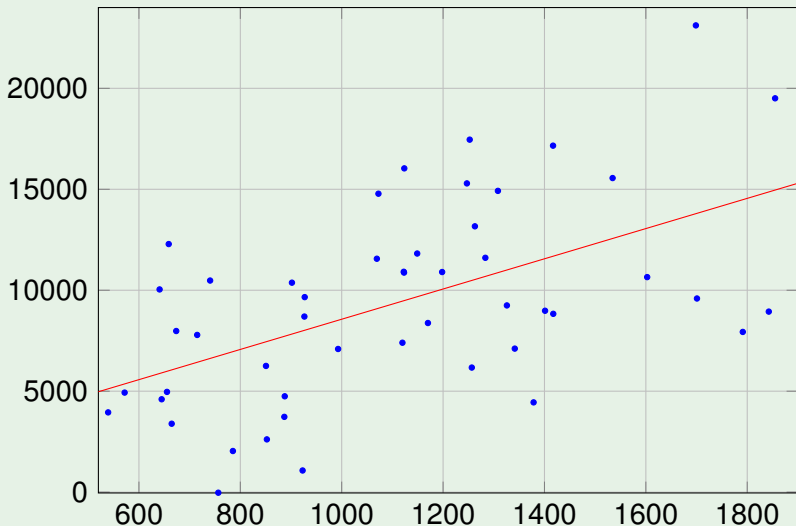
## Definition

We call  $y$  the **response variable**.

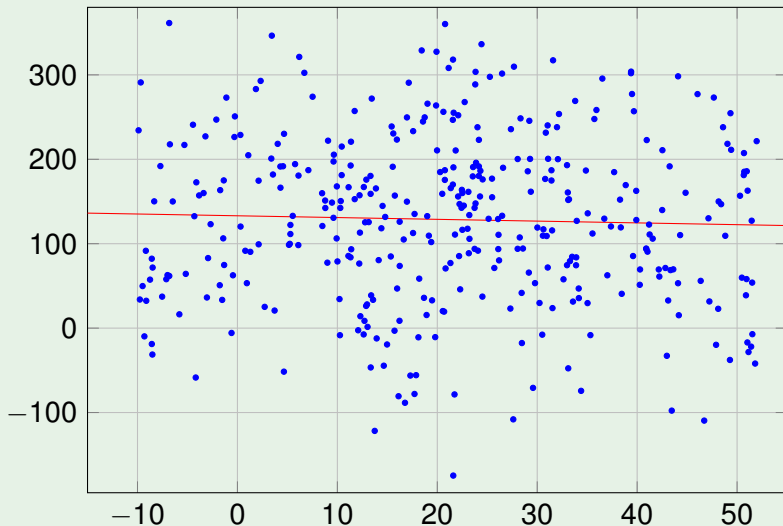
## Example 1



## Example 2

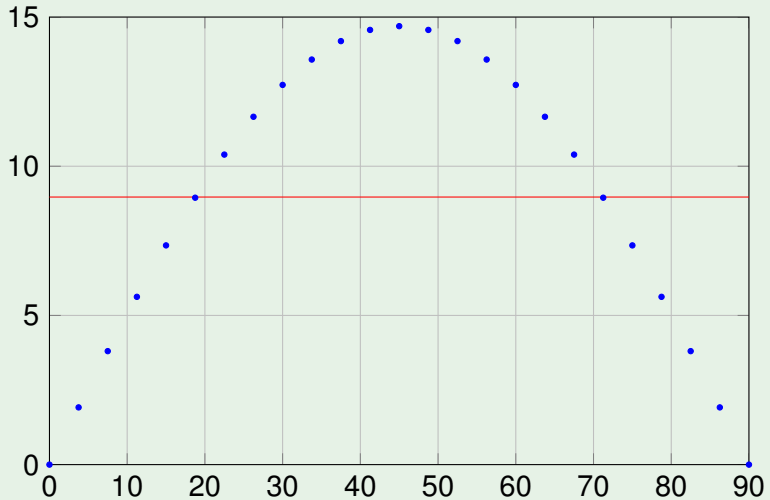


### Example 3



Even though this looks like just a cloud, the linear model may be useful.

## Example 4



Because, there is a clear non-linear pattern, the linear model is a poor choice for this data.

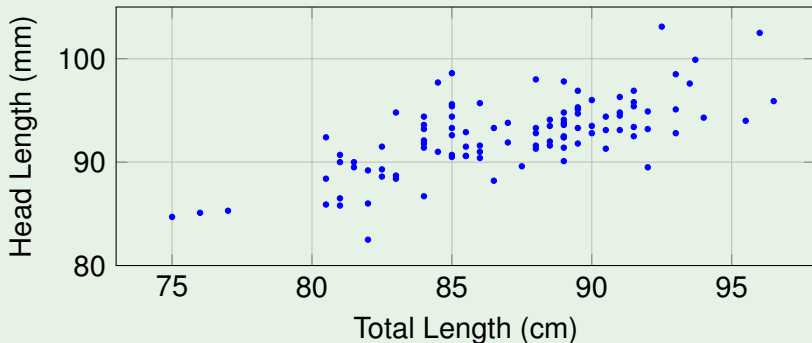
## Example 5

Brushtail possums are a marsupial that lives in Australia.

Researchers captured 104 of these animal and took body measurements before releasing the animals back into the wild.

We will consider two measurements:

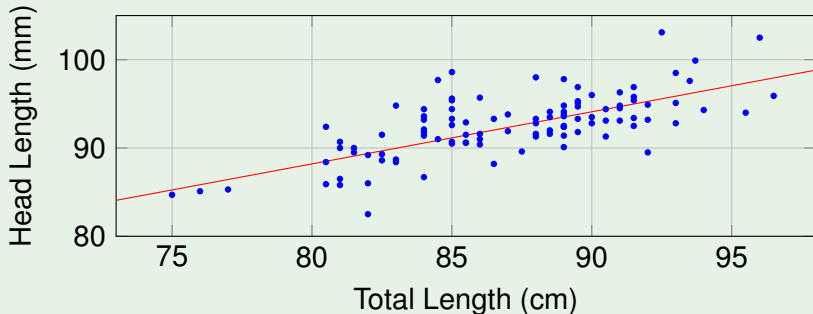
- The length of each possum from head to tail.
- The length of each possum's head.



## Example 5 (Continued)

We could fit the linear relationship by eye, giving the equation:

$$\hat{y} = 41 + 0.59x$$



This allows us to make estimates of the possum population.

$$\hat{y} = 41 + 0.59(80) = 88.2$$

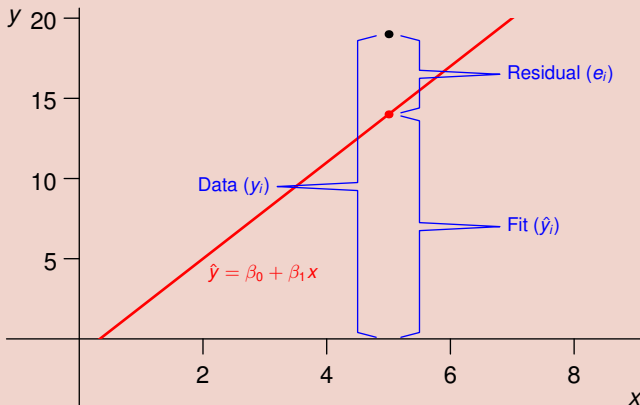
We expect that a possum with a total length of 80cm would have a head length of about 88.2mm.



## Definition

**Residuals** are the leftover variation in the data after accounting for the model fit:

$$\text{Data} = \text{Fit} + \text{Residual}$$

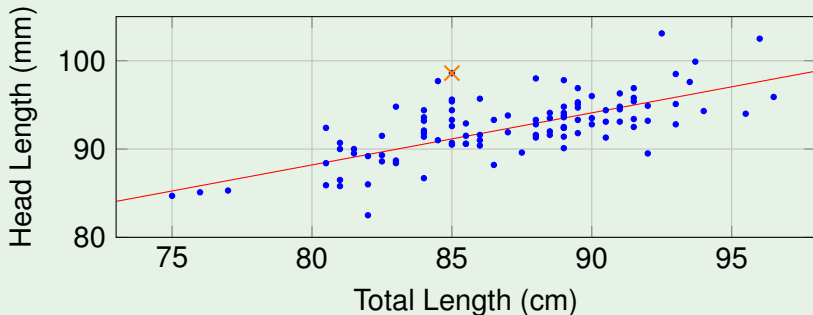


The residuals are calculated as:

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

## Example 5 (Continued)

Let's calculate the residual for the observation (85.0, 96.6).



We first need to find  $\hat{y}$ :

$$\hat{y}_{\times} = 41 + 0.59(85) = 91.15$$

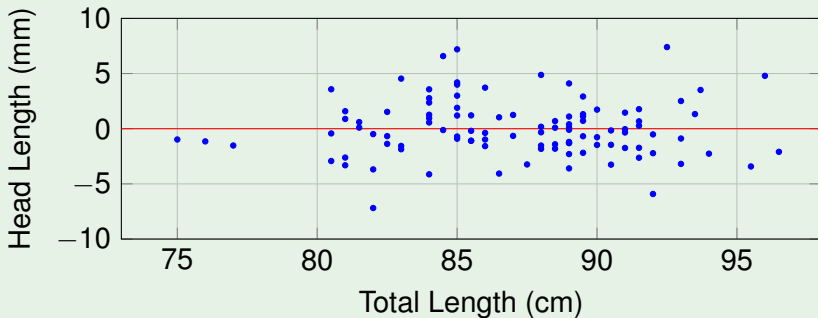
Next, the residual:

$$e_{\times} = y_{\times} - \hat{y}_{\times} = 96.6 - 91.15 = 7.45$$

## Definition

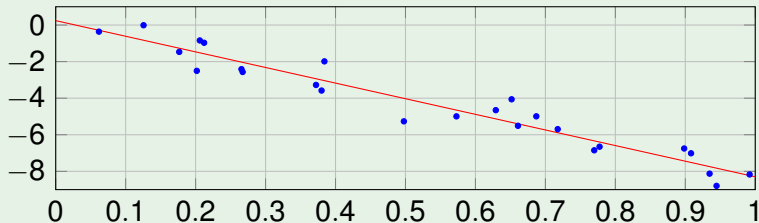
If the residual for each point is calculated, the corresponding graph is called a **residual plot**.

## Example 5 (Continued)

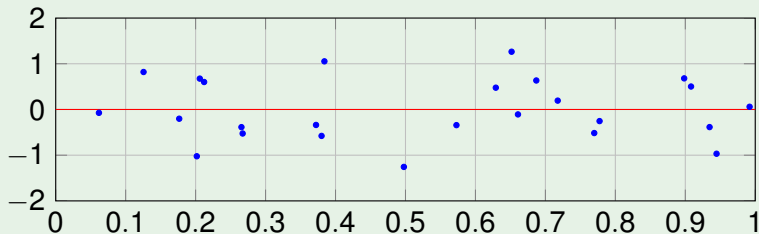


## Example 6

Scatter plot with linear regression:

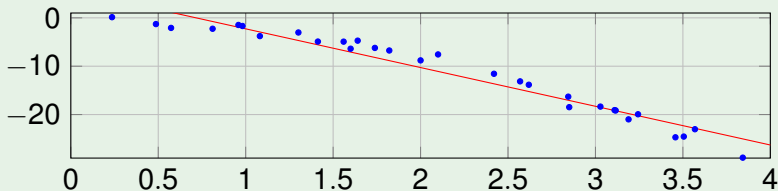


Residual plot:

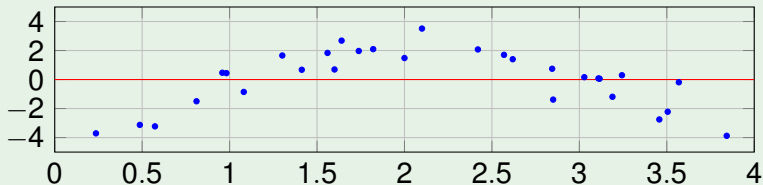


## Example 7

Scatter plot with linear regression:



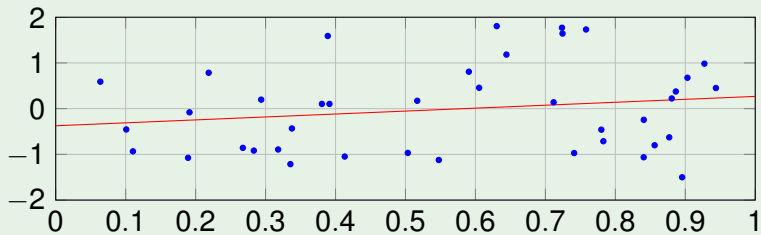
Residual plot:



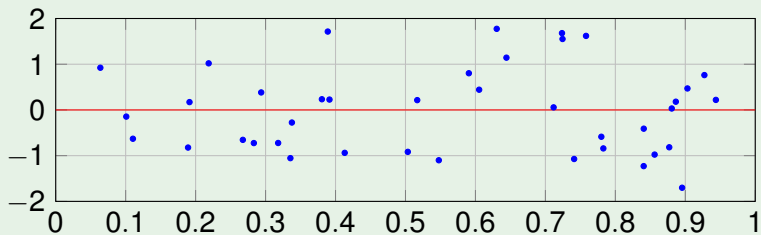
Since there is a clear curve in the residual plot, we should not use a linear model. A more advanced method is needed.

## Example 8

Scatter plot with linear regression:



Residual plot:



## Definition

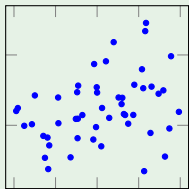
**Correlation**, which is always between -1 and 1, describes the strength of the linear relationship between two values. We denote the correlation by  $R$ .

## Note

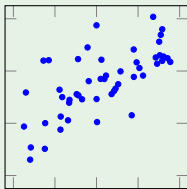
While technology is often used, the formula for correlation is:

$$R = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y} \right)$$

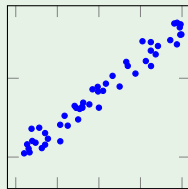
## Example 9



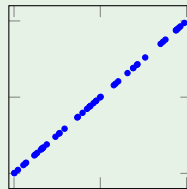
$$R = 0.33$$



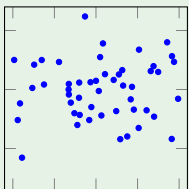
$$R = 0.69$$



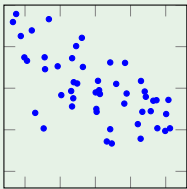
$$R = 0.98$$



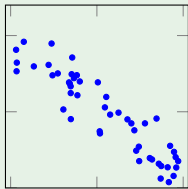
$$R = 1.00$$



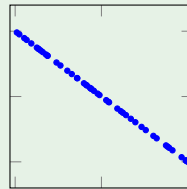
$$R = 0.08$$



$$R = -0.64$$



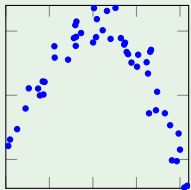
$$R = -0.92$$



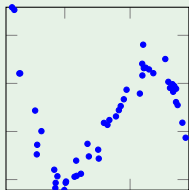
$$R = -1.00$$



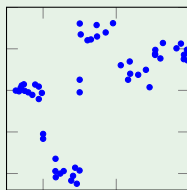
## Example 10



$$R = -0.23$$



$$R = 0.31$$



$$R = 0.50$$

Since each of these scatter plots has a clear non-linear pattern, a linear model is not appropriate and correlation shouldn't have been calculated.

### Note

Given a table of  $x$  and  $y$  values, a computer will happily compute correlation. It is your job to determine if a linear model makes sense.