M7S2 - Regression Line

Professor Jarad Niemi

STAT 226 - Iowa State University

November 15, 2018

Outline

- Regression line
  - Residual
  - Sample intercept and interpretation
  - Sample slope and interpretation

Interpreting a line

Suppose there is a line

$$y = m \cdot x + b$$

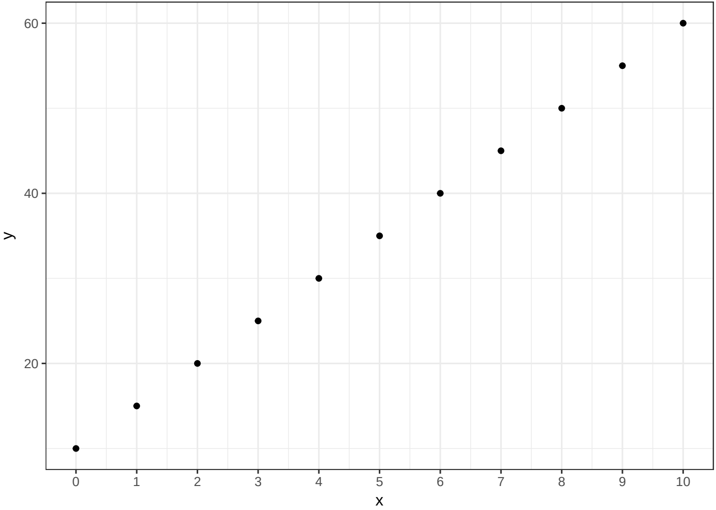
Interpret

- b: is the y-intercept, i.e. the value of y when $x = 0$
- m: is the slope, i.e. the change in y for each unit change in x

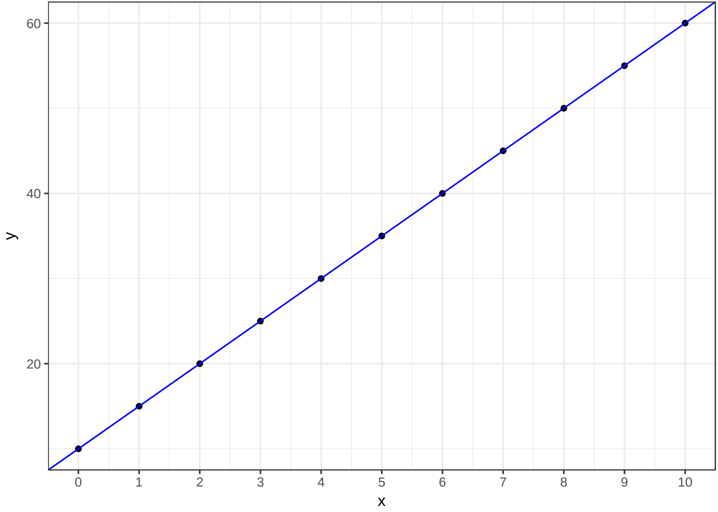
If x increases by one unit, then y changes by

$$\begin{aligned} & m \cdot (x+1) + b - (m \cdot x + b) \\ & = m \cdot x + m + b - m \cdot x - b \\ & = m \end{aligned}$$

Finding the line

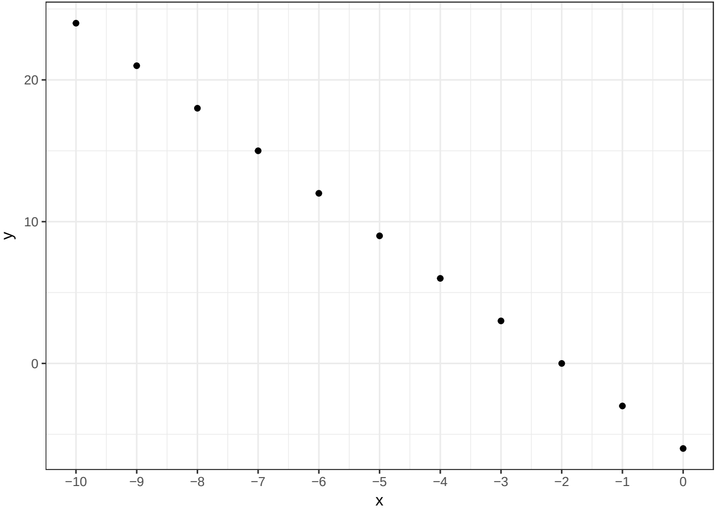


Finding the line

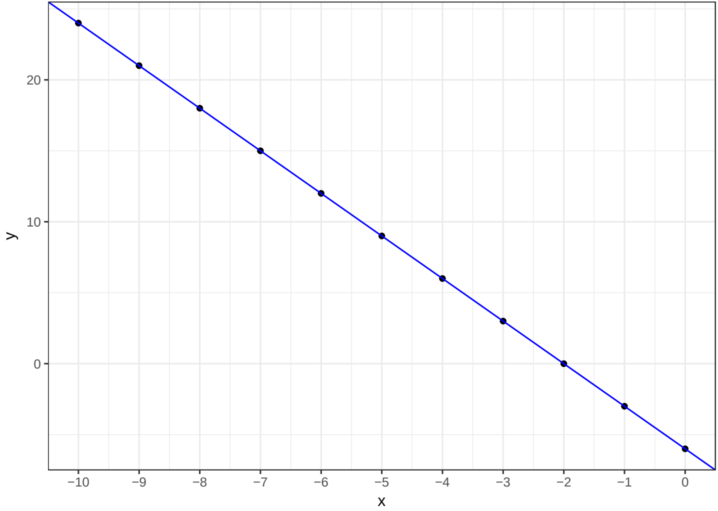


$$y = 5x + 10$$

Finding the line



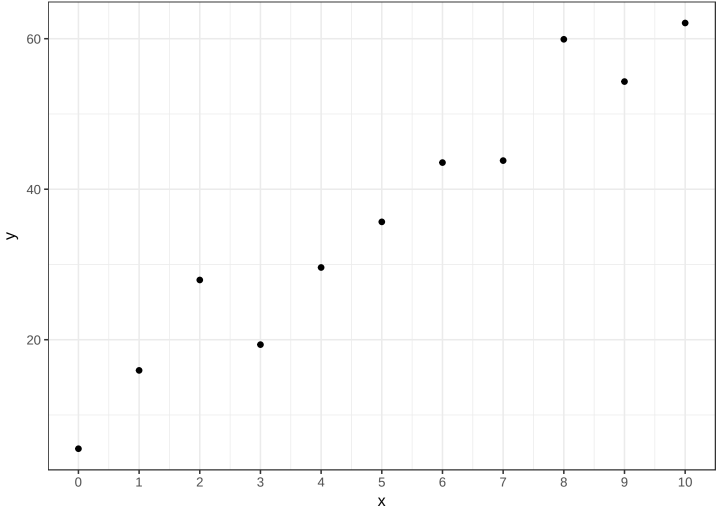
Finding the line



$$y = -3x + -6$$

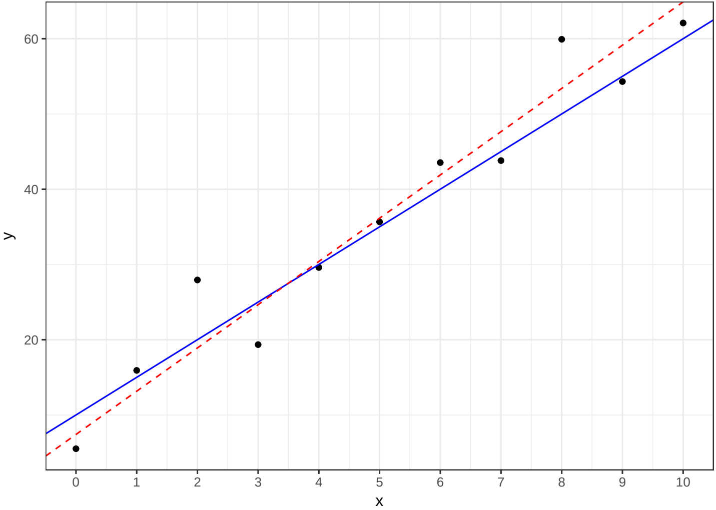
Noisy data

When the data are noisy, finding the line is not so easy



Noisy data

When the data are noisy, finding the line is not so easy



Residuals

Definition

A prediction equation is given by

$$\hat{y} = b_0 + b_1 \cdot x$$

where $\hat{y}$ is the predicted value of y for a specified value of x for some intercept b_0 and slope b_1. For a collection of observations (x_i, y_i) for $i = 1, \ldots, n$, we can calculate the predicted value for each observation, i.e.

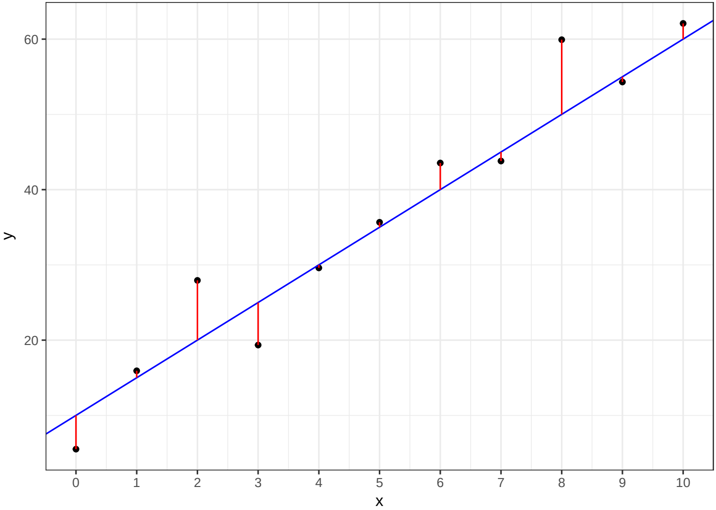
$$\hat{y}_i = b_0 + b_1 \cdot x_i$$

The residual, r_i, for an observation is the observed value minus the predicted value, i.e.

$$r_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 \cdot x_i) = y_i - b_0 - b_1 \cdot x_i$$

The residual is the vertical distance from the observation to the line.

Residuals graphically



Regression line

Definition

The (least squares) regression line is the value for b_0 and b_1 in the prediction equation that minimizes the sum of the squared residuals, i.e. minimizes

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 \cdot x_i)^2$$

We call

- b_0 the sample intercept and
- b_1 the sample slope.

Sometimes the regression line is referred to as the prediction line.

`https://gallery.shinyapps.io/simple_regression/`

Speed vs stopping distance of cars

We run an experiment where we record

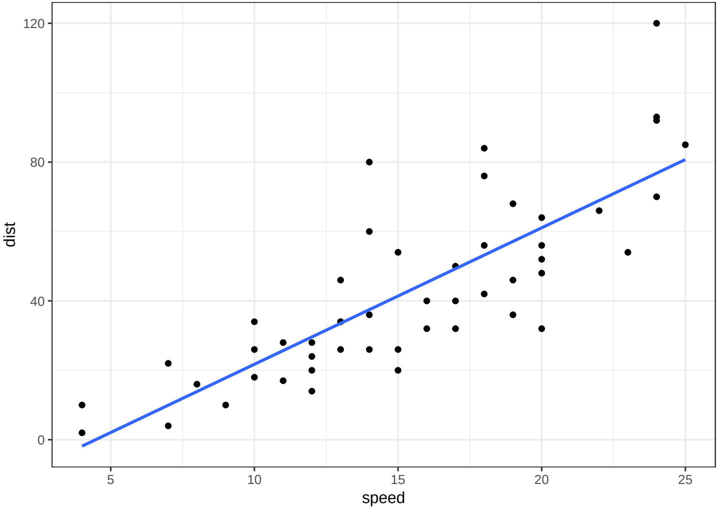
- the speed (mph) a car is going and
- the distance (ft) it takes for the car to stop.

We are interested in constructing a regression line to understand the relationship between speed and distance.

Let

- the explanatory variable be the speed and
- the response be the distance.

Speed vs stopping distance graphically



Estimated intercept and slope

```
Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)        speed
    -17.579        3.932
```

Thus the regression line is (approximately)

$$\hat{y} = -18 + 4 \cdot x$$

where

- x represents speed (mph) and
- y represents distance (ft).

Interpretation

Definition

The sample intercept (b_0) is the predicted value of the response, i.e. $\hat{y}$, when the explanatory variable (x) is zero, i.e. $x = 0$. The sample slope (b_1) is the predicted change in the response when the explanatory variable increases by one unit.

Notes:

- The intercept may not be meaningful.
- A positive slope, $b_1 > 0$, indicates a positive direction ($r > 0$).
- A negative slope, $b_1 < 0$, indicates a negative direction ($r < 0$).

Speed vs stopping distance of cars

Thus the regression line is (approximately)

$$\hat{y} = -18 + 4 \cdot x$$

where

- x represents speed (mph) and
- y represents distance (ft).

Thus

- The predicted stopping distance of a car at 0 mph is -18 ft. This is not meaninful!
- For each additional mile per hour the car is traveling, the predicted additional distance to stop is 4 ft.