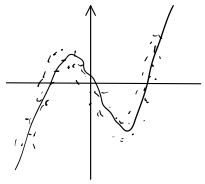
Association

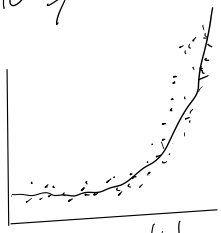
Two variables, (X, Y) where each value of one variable is associated with another (due to being from the same individual)

are associated if values of one variable tend to correspond to specific values of the other variable.

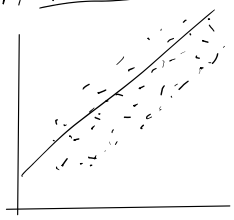
We are interested in linear association



non-linear (quadratic)



exponential (non-linear)



We're interested in Linear association which we can measure using the linear correlation coefficient

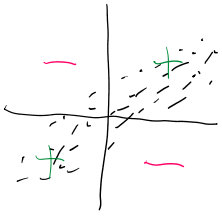
$R \in [-1, +1]$

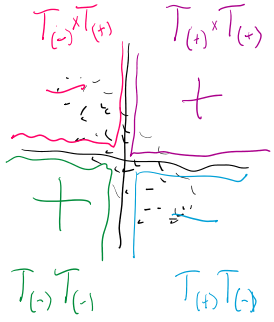
$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S_x}\right)\left(\frac{Y_i - \bar{Y}}{S_y}\right)$$

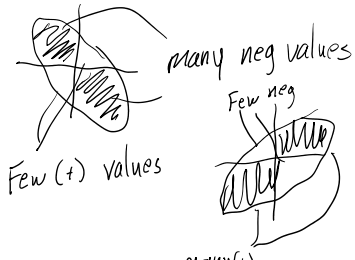
$$R = \frac{1}{n-1}\sum_{i=1}^{n} T_{X_i} \times T_{Y_i}$$

ID	X	Y
1	X_1	Y_1
2	X_2	Y_2
3	X_3	Y_3
4	X_4	Y_4
...	...	...
n	X_n	Y_n

Predictor ↗ Response ↖



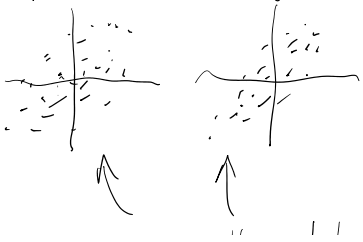


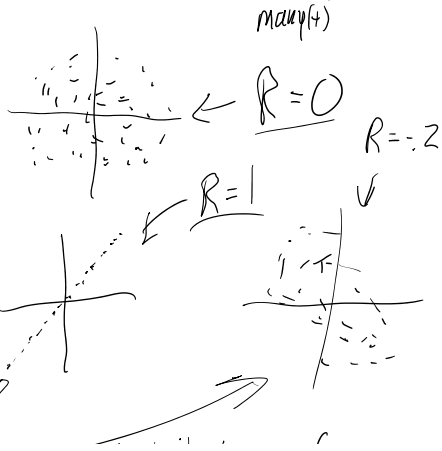


+/− indicates the direction

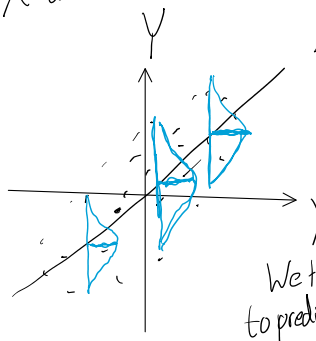
|R| indicates the strength

Lower |R| Higher |R|





Scatterplots show the joint distribution of X and Y values



We want to describe this relationship with a line

$\hat{y} = b_0 + b_1 X$ (Regression Coefficients)

We try to predict the average value of y_i values for specific values of x_i

$\hat{y}$

b_0 Intercept term

$b_1 X$ Slope term

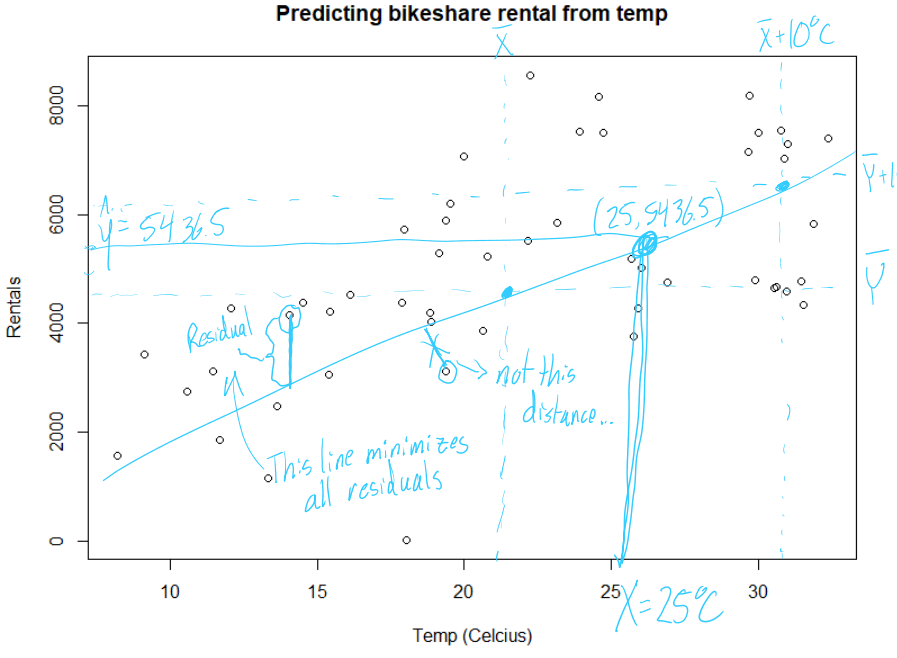
It's relativatively simple to find this line for two-variables

$$\hat{y} = b_0 + b_1 X$$

Consider Bikeshare rentals → how do rental rates (# of bikes rented) relate to temperature?

Y - # of bikes rented

X - temperature avg (°C) → Describes a day



Overall the trend appears fairly linear, can we find an equation to describe it?

Variables needed for finding the regression Eq.

$\bar{x} = 20.91°C$ $\quad \bar{y} = 4696.0$ bikes

$s_x = 7.429°C$ $\quad s_y = 2032.4$ bikes

$R = 0.6616$

The line will always pass through the point of averages $(\bar{x}, \bar{y})$, (20.91°C, 4696.0 bikes)

$\Rightarrow \hat{y} = \hat{b}_0 + \hat{b}_1 \bar{x}$ $\quad \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$ ($\hat{b}_0$: intercept)

$\hat{b}_1 = \frac{\text{Rise}}{\text{Run}} = \frac{s_y}{s_x} \cdot R$ (slope)

$\hat{b}_1 = \frac{2032.4 \text{ bikes}}{7.429°C} \cdot 0.6616$

$= 181.0 \text{ bikes}/°C$

Interpret: For a day that is 1 degree hotter, we expect 181 more bike rentals on average.

$\hat{b}_0 = 4696 \text{ bikes} - \frac{181.0 \text{ bikes}}{°C} \times 20.91°C$

$= 911.5$ bikes

Interpret (if it makes sense...) this is the expected # of rentals for a day with temp 0°C (not here)

This is outside the range of data

$\hat{y} = 911.5 \text{ bikes} + 181.0 \frac{\text{bikes}}{°C} \times X$

Regression equation describing bike rentals for different temps.

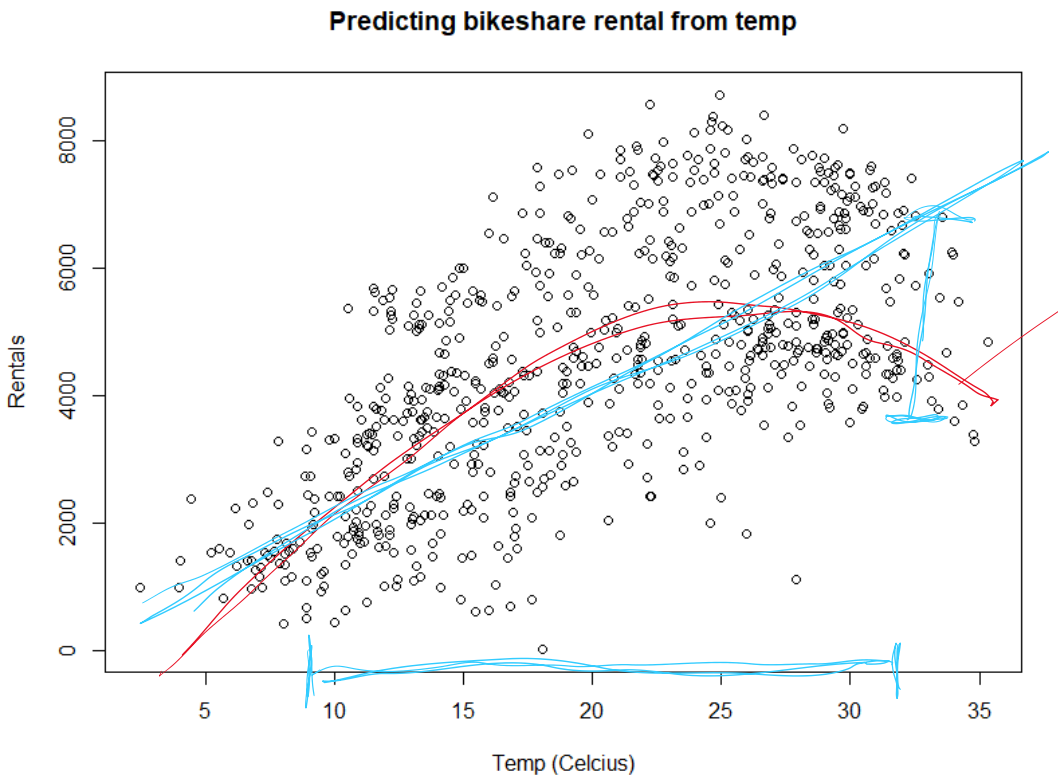
How many bike rentals should we expect if $X = 25°C$

$\hat{y} = 911.5 \text{ bikes} + 181.0 \frac{\text{bikes}}{°C} \times 25°C$

$\hat{y} = 5436.5$ bikes

This is only valid within the range of

within the range of
data used to fit the model.



This is a non-linear relationship

Using a model outside the range of data used to fit it is extrapolation and should be avoided.