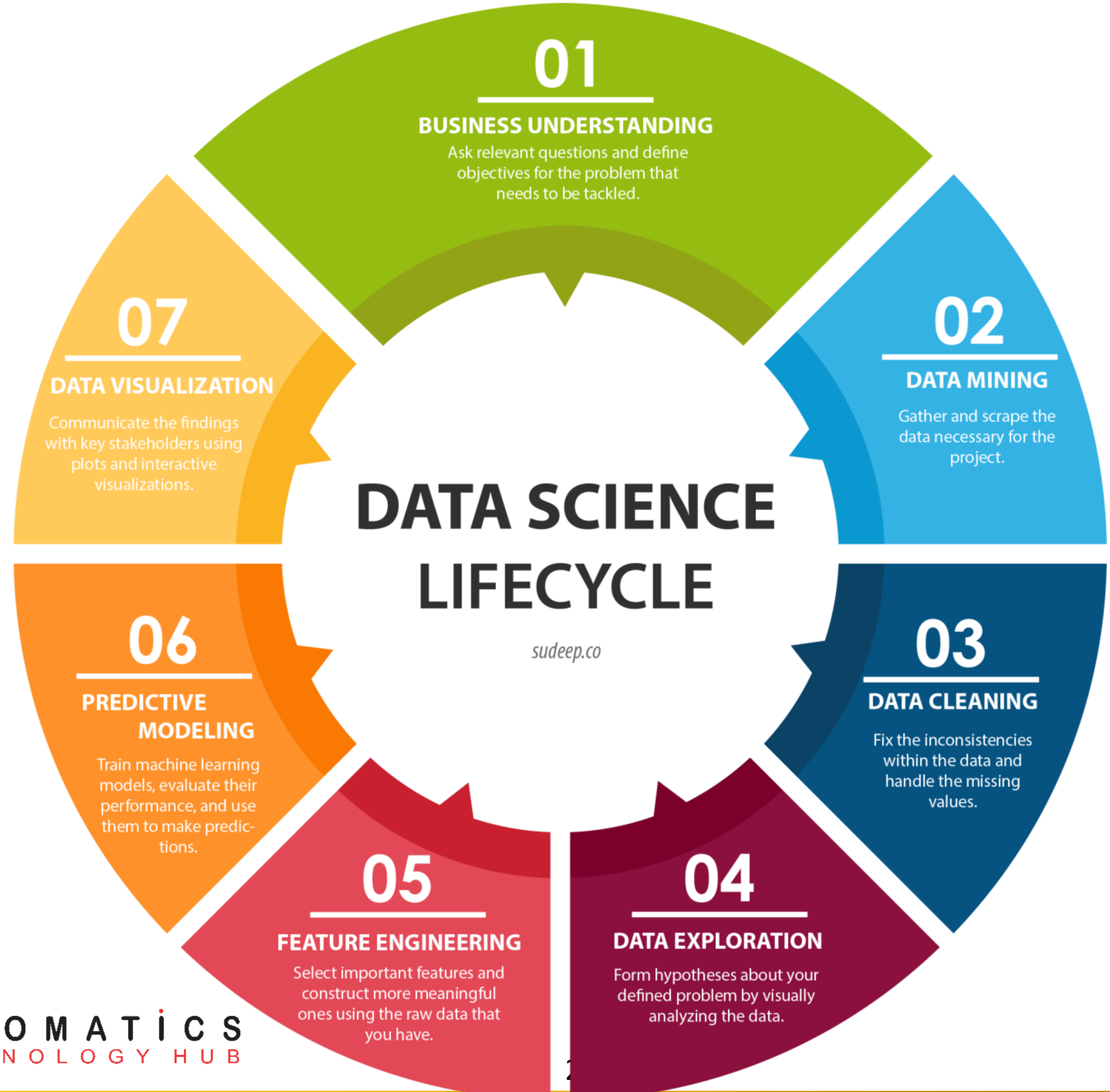
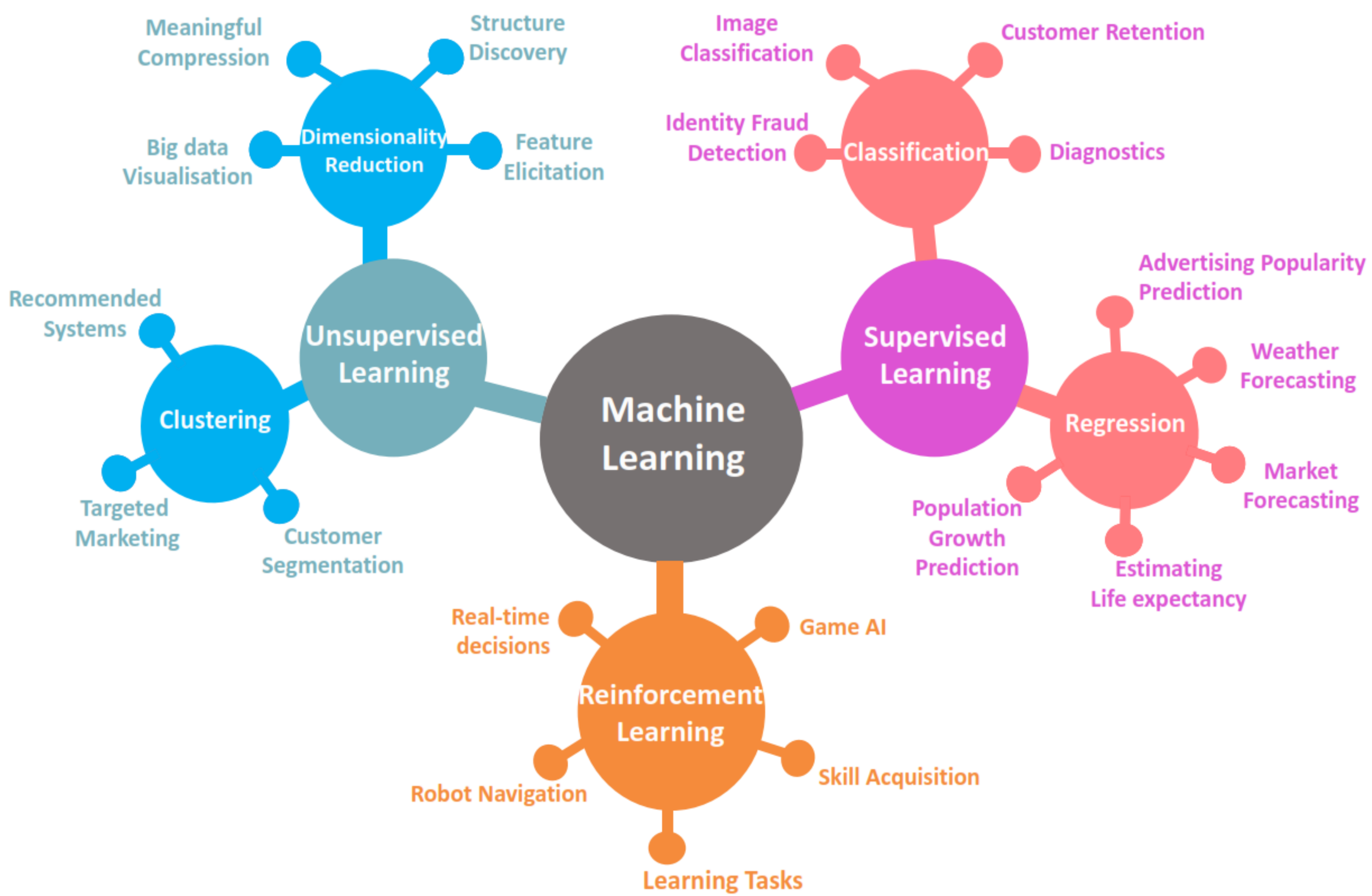
Machine Learning



INNOMATICS TECHNOLOGY HUB


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualisation
Feature Elicitation
Clustering
Recommended Systems
Targeted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Identity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Population Growth Prediction
Estimating Life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Statistics for Decision Modelling

(Machine Learning)

Supervised Models

Linear Regression

Why Model Building ?

In any business, there are some easy-to-measure metrics

- *Age, Gender, Income, Education level etc.*

and a difficult-to-measure metric

- *Amount of loan to give; Will she buy or not; How many days a patient will stay in the hospital etc.,*

- Regression enables you to compute the latter form the former

Simplest Learning Models

- **Linear Regression:** Measuring the relation between two or more analog variables (class variable is numeric)

- **Logistics Regression:** A classification model (class variable is categorical)

Simple Linear Regression

Speed vs Stopping distance

I13

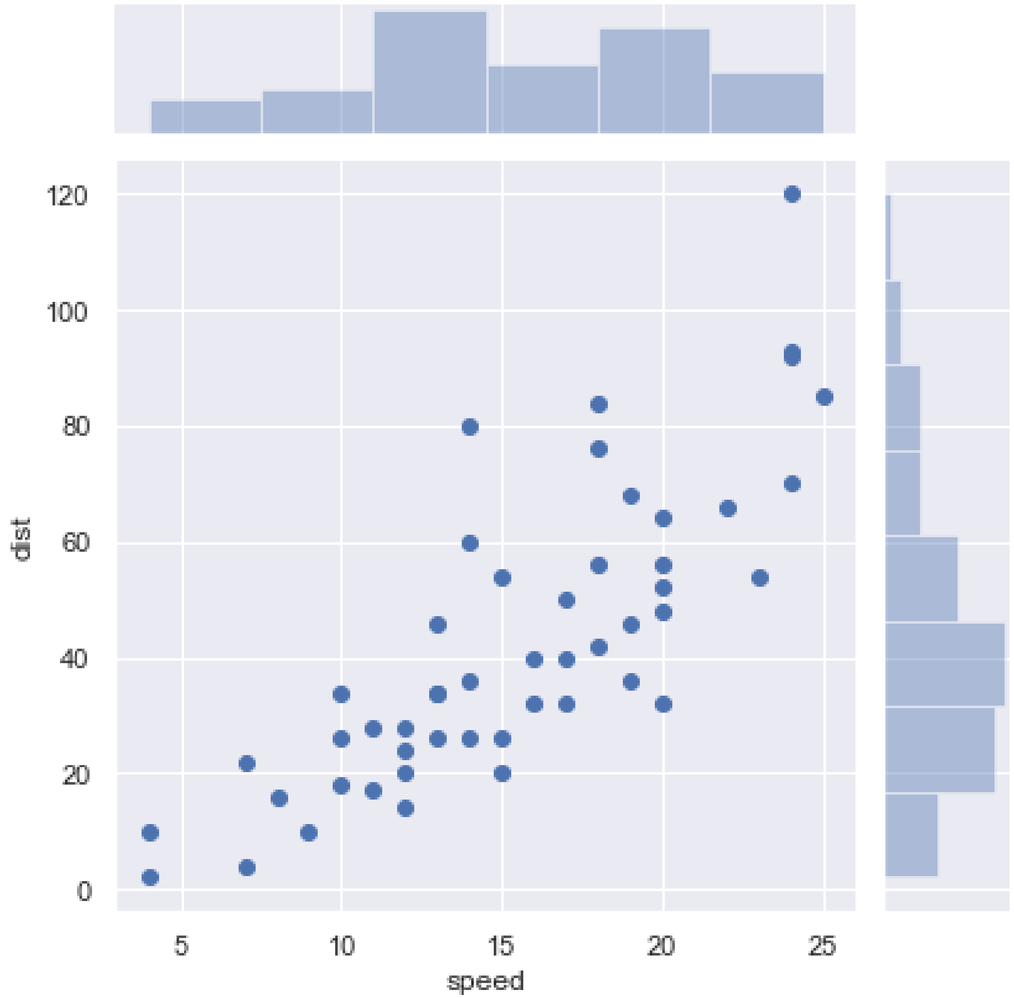
	A	B	C	D
1		speed	dist	
2	1	4	2	
3	2	4	10	
4	3	7	4	
5	4	7	22	
6	5	8	16	
7	6	9	10	
8	7	10	18	
9	8	10	26	
10	9	10	34	
11	10	11	17	
12	11	11	28	
13	12	12	14	
14	13	12	20	
15	14	12	24	
16	15	12	28	
17	16	13	26	
18	17	13	34	
19	18	13	34	
20	19	13	46	
21	20	14	26	
22	21	14	36	
23	22	14	60	
24	23	14	80	

cars

The “cars” dataset contains 50 pairs of data points of Speed(mph) vs stopping distance(ft). That were collected in 1920

Speed vs Stopping distance

- Independent variable (explanatory) –Speed(mph) – Plotted on x-axis
- Dependent variable (response) – Stopping distance(ft) – Plotted on Y-axis

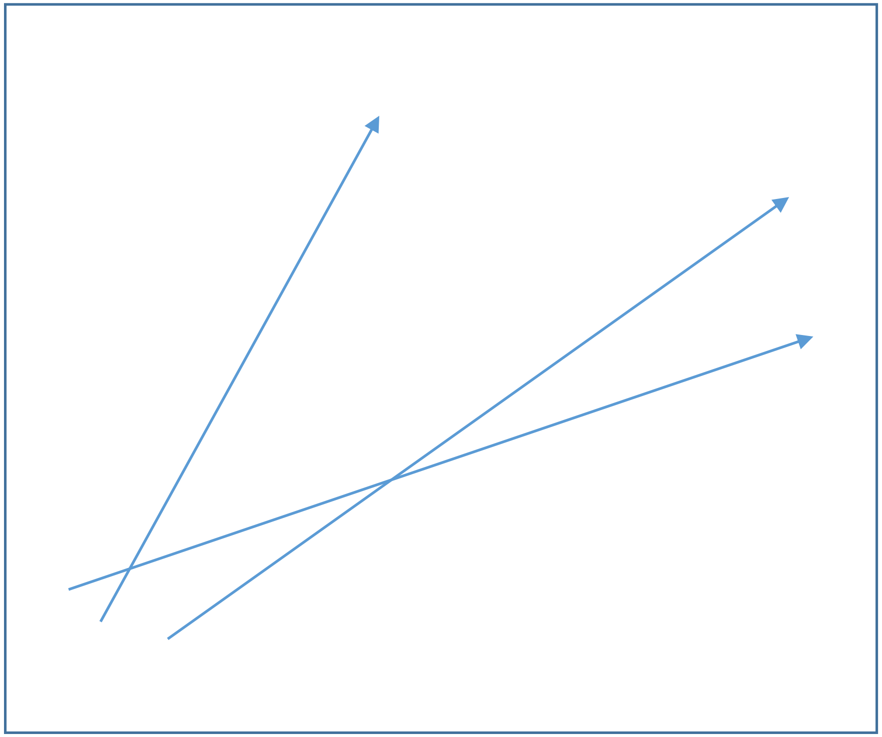


Speed vs Stopping distance

- Another car with the same speed, might not have the same stopping distance
- X is known (No uncertainty)
- Y has uncertainty (It's a sample picked from some unknown distribution)

Start with a Function/Hypothesis with some parameters

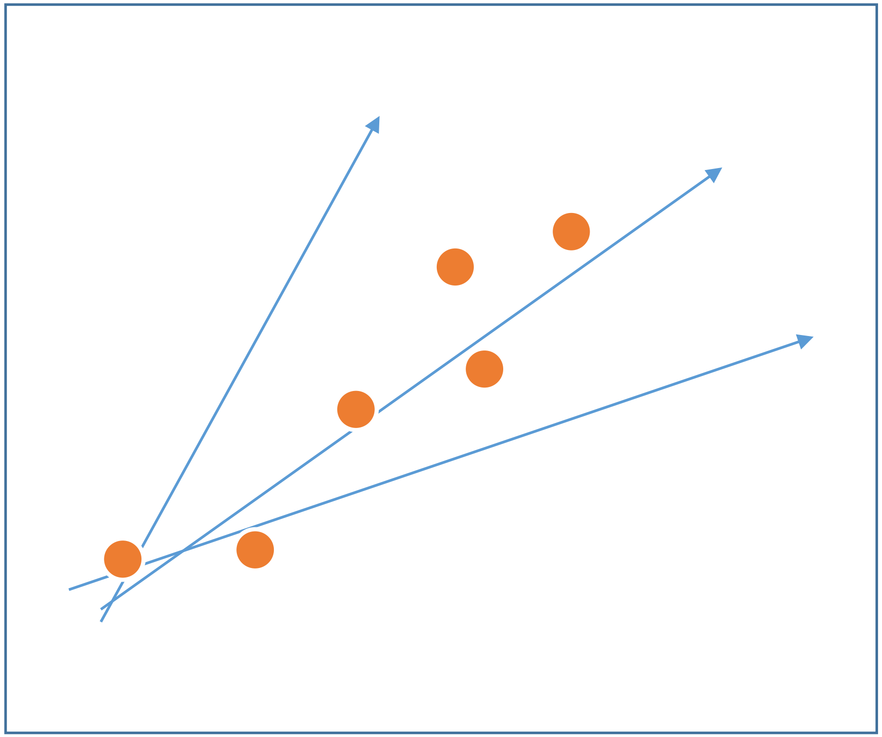
$y = \beta_0 + \beta_1 x$ (Deterministic model)



How to pick the Best Model

$y = \beta_0 + \beta_1 x + \varepsilon$ (Probabilistic model)

$y = E(Y|X = x) + \varepsilon$

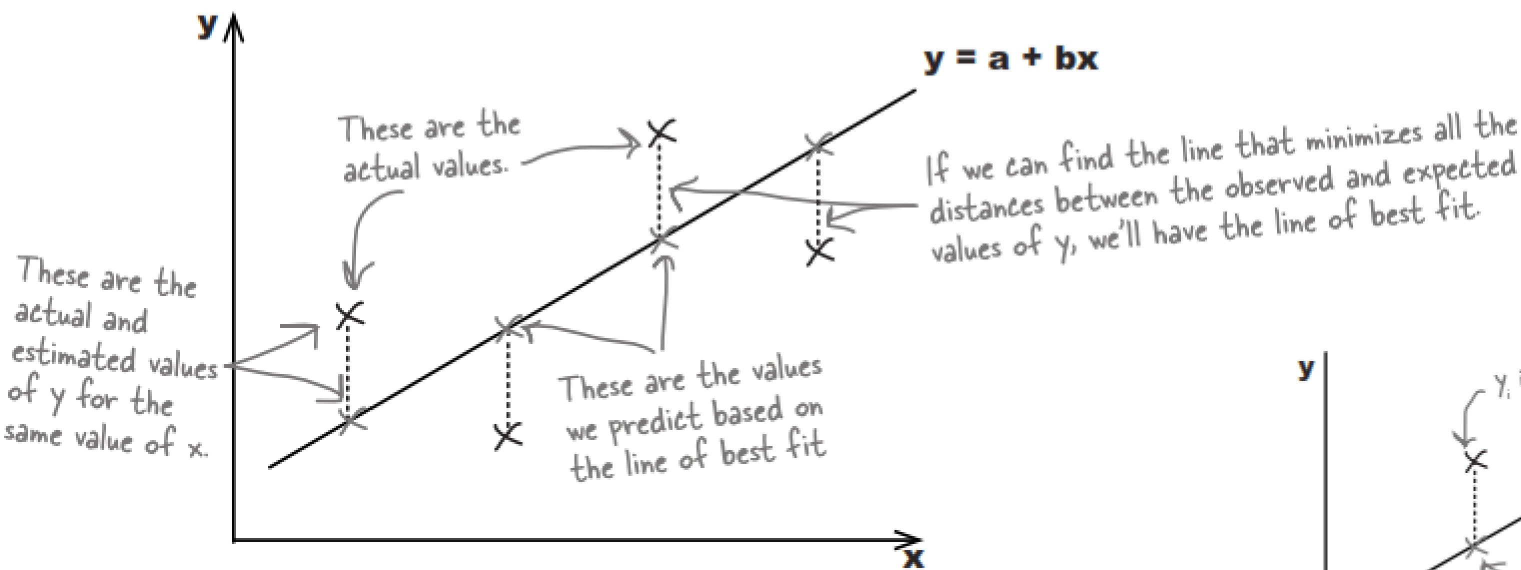


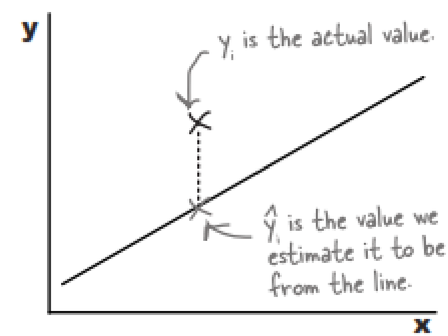
The lines whose residual error on all points is the least is the best line.

To ensure residual errors don't cancel. We take squares of residual errors35

We need to minimize errors.

- We can do that by minimizing $\sum(y_i - \hat{y}_i)$, where y_i is the actual value and $\hat{y}$ its estimate. $(y_i - \hat{y}_i)$ is also known as the **residual**.





We need to minimize errors.

Just as we did when finding variance, we find the sum of squared errors or SSE. *Note in variance calculations, we subtract mean, $\overline{y_i}$, not $\widehat{y_i}$.*

$$SSE = \sum (y_i - \widehat{y_i})^2$$

The value of b, the slope, that minimizes the SSE is given by

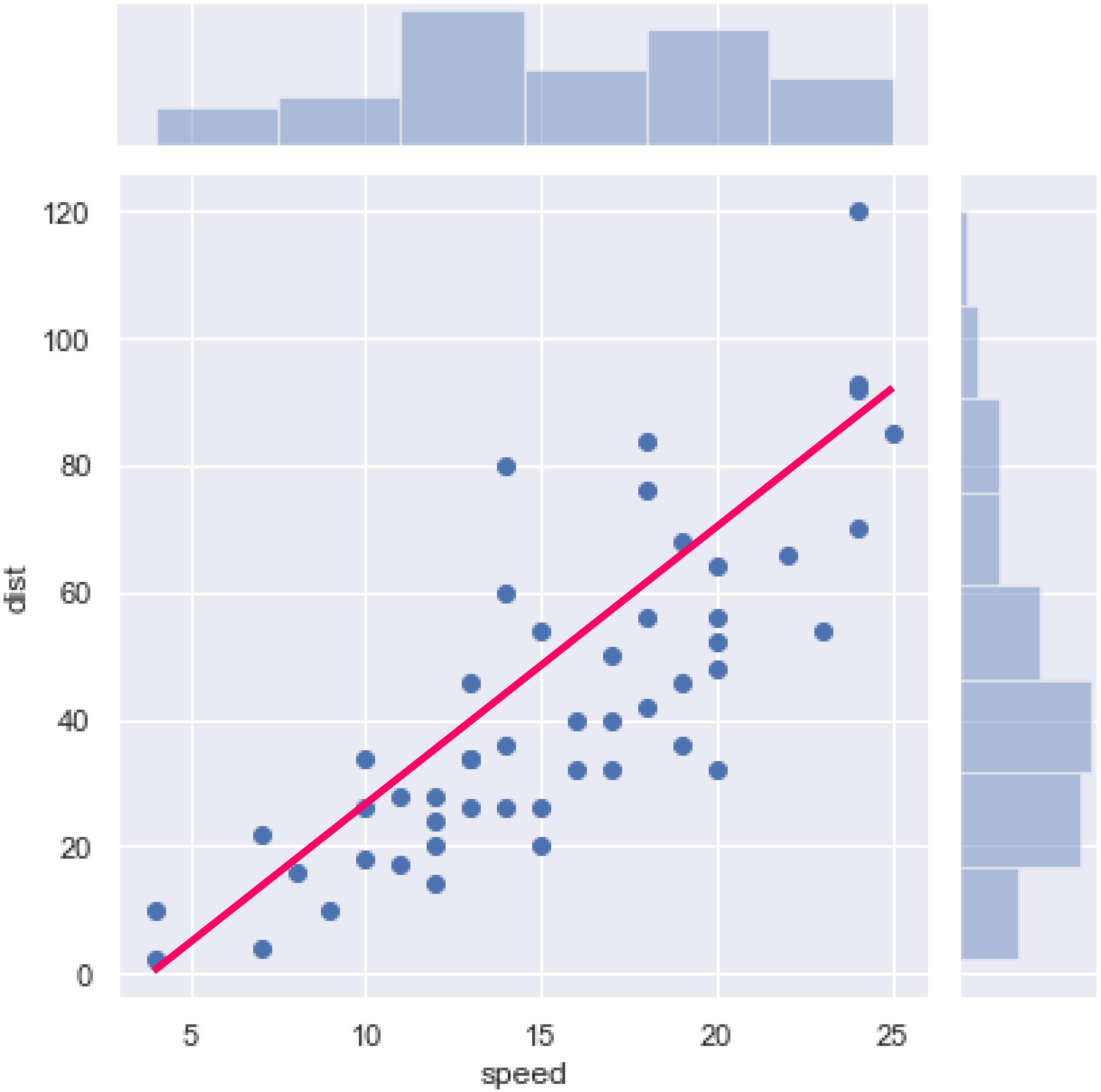
$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

The value of *b,* the slope, that minimizes the SSE is given by

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

How do you calculate a ? The line of best fit must pass through $(\bar{x}, \bar{y})$. Substituting in the equation y = a+ b x, we can find a.

This method of fitting the line of best fit is called **least squares regression.**



$$y = 3.93\,X\; - 17.58$$

Covariance

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

Covariance $s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$

Slope coefficient can be expressed in terms of covariance & Standard deviation s_x

$$b = \frac{s_{xy}}{s_x^2}$$

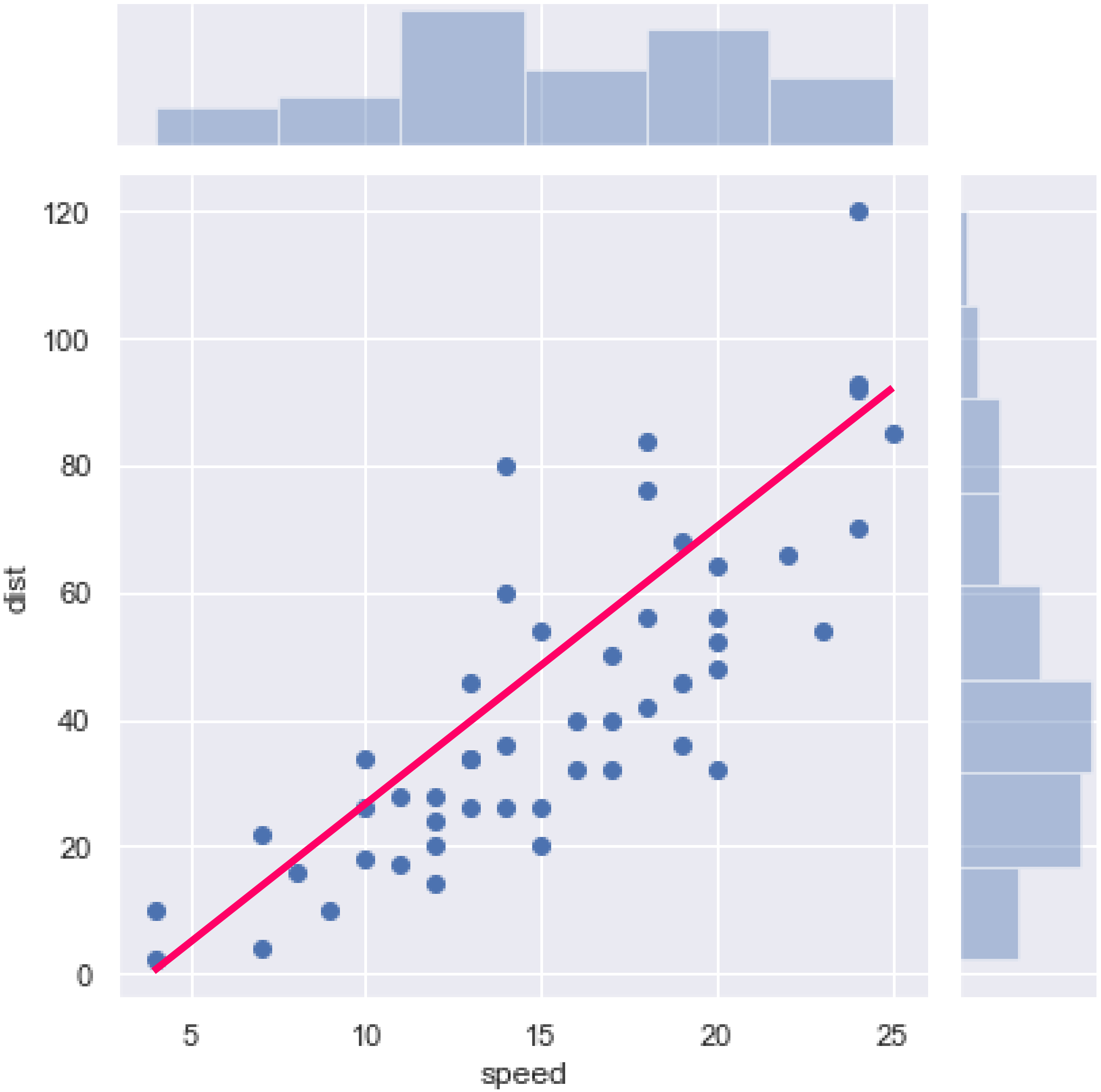
Covariance

$$\text{Covariance } s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

- If both x and y are large distance away from their respective means, the resulting covariance will be even larger
  - The value will be positive if both are below the mean or both are above.
  - If one is above and the other below, the covariance will be negative

- If even one of them is very close to the mean, the covariance will be small.

Cov(x,x) = Var(x)



$covariance:$

$dataframe.cov()$

	speed	dist
speed	27.95	109.94
dist	109.94	664.06

Covariance and correlation

$$\text{Covariance } s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

- The value of covariance itself doesn't say much. It only show whether the variable are moving together (positive value) or opposite to each other (negative value).
- To find the strength of how the variable move together, covariance is standardized to the dimensionless quantity, called correlation (r).

$$r = \frac{s_{xy}}{s_x s_y}$$

Covariance and Correlation

$$s_{xy} = \frac{\sum(x-\bar{x})\,(y-\bar{y})}{(n-1)},\ r = \frac{s_{xy}}{s_x s_y}$$

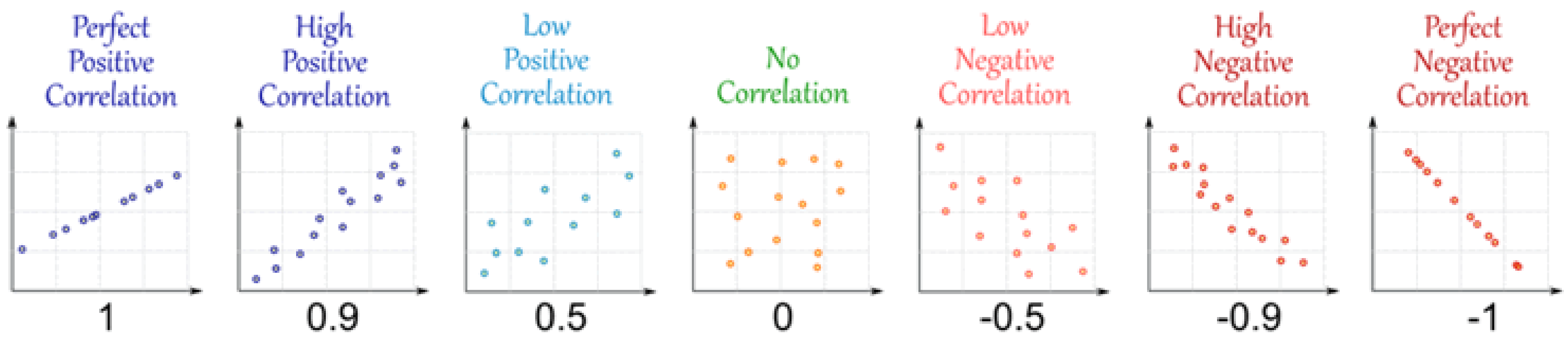
- The value of covariance itself doesn't say much. It only shows whether the variables are moving together (positive value) or opposite to each other (negative value).

- To know the strength of how the variables move together, covariance is standardized to the dimensionless quantity, correlation.

Correlation Coefficient

Correlation coefficient, *r, is a* number between *-1* and *1* and tells us how well a regression line fits the data.

$$r = \frac{bs_x}{s_y}$$



It gives the strength and direction of the relationship between two variables.

Correlation Coefficient

$$r = \frac{bs_x}{s_y}$$

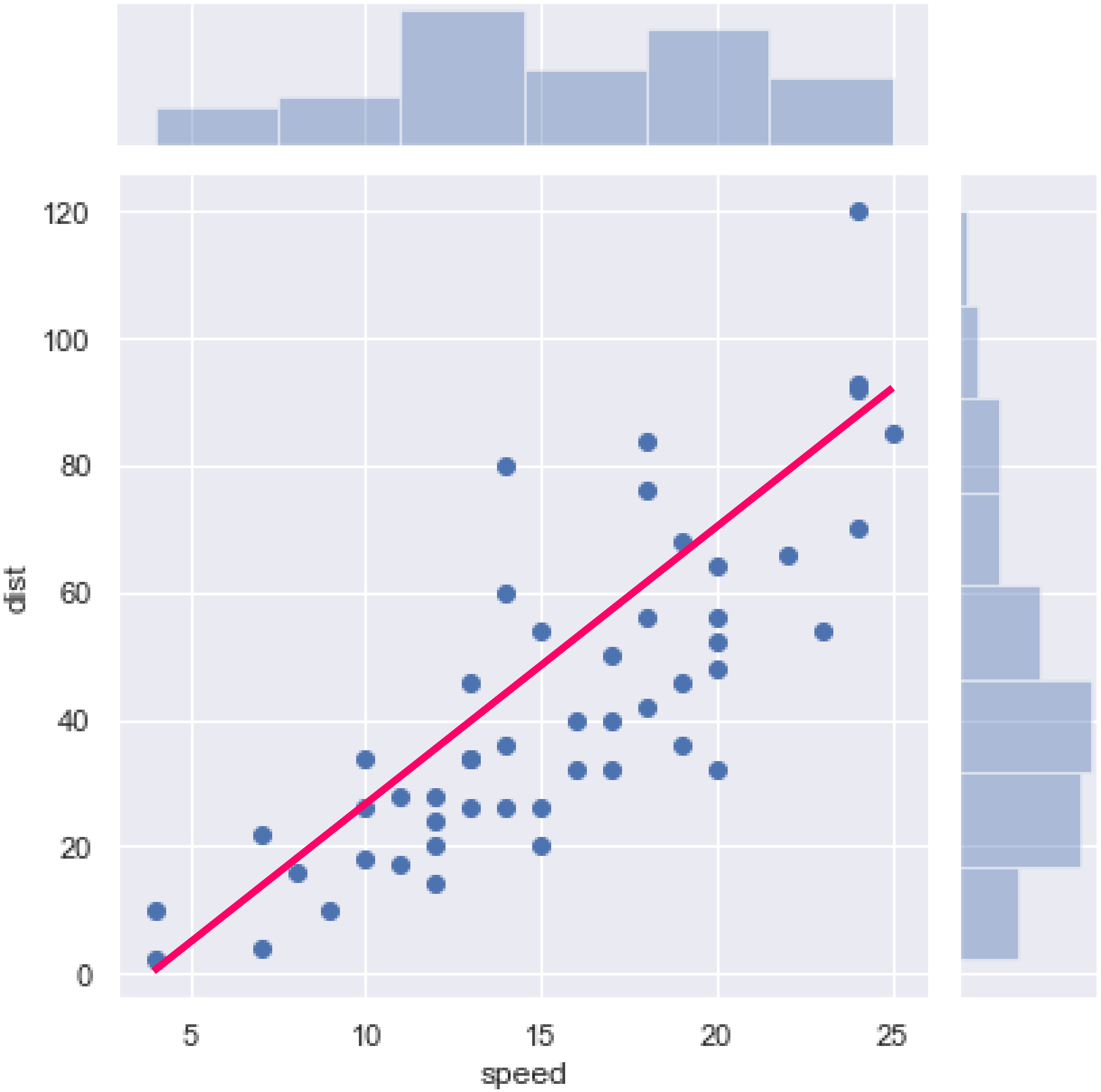
where *b* is the slope of the line of best fit, s_x is the standard deviation of the *x* values in the sample, and s_y is the standard deviation of the y values in the sample.

$$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \text{ and } s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}$$

Correlation Coefficient and Covariance

$s_x^2 = \frac{\sum(x-\bar{x})^2}{(n-1)}$, $s_y^2 = \frac{\sum(y-\bar{y})^2}{(n-1)}$, $s_{xy} = \frac{\sum(x-\bar{x})\,(y-\bar{y})}{(n-1)}$, where s_x^2 is the sample variance of the x values, s_y^2 is the sample variance of the y values and s_{xy} is the covariance.

$b = \frac{s_{xy}}{s_x^2}$ and so, $r = \frac{s_{xy}}{s_x s_y}$

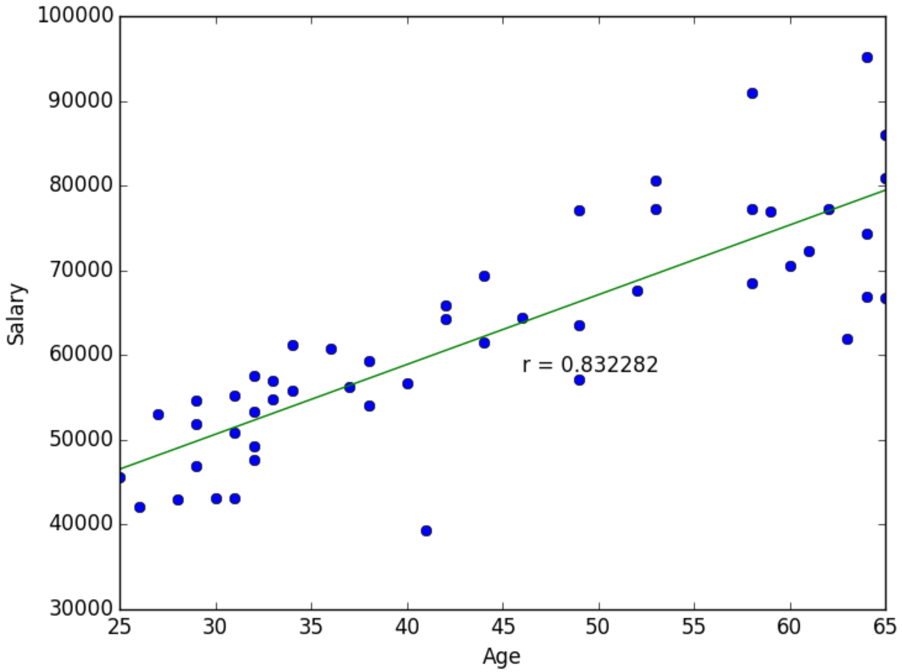


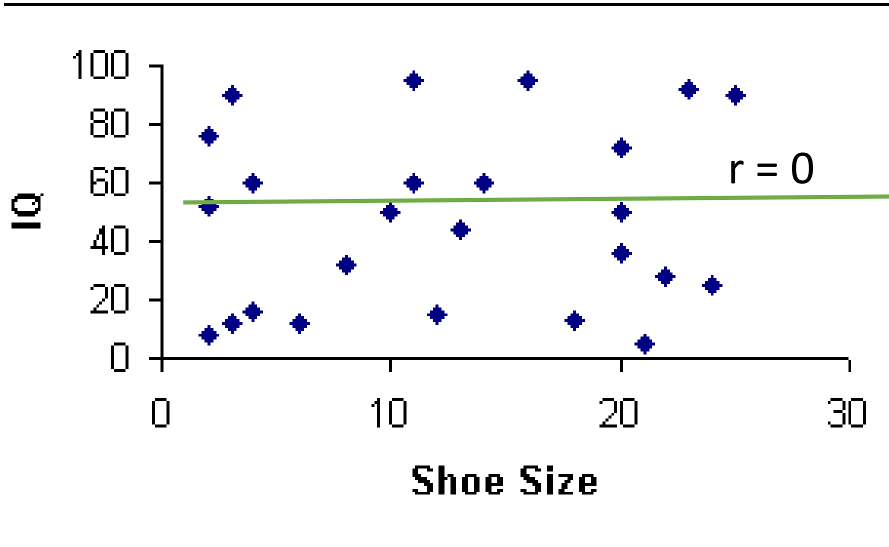
$covariance$:

$dataframe.corr()$

	speed	dist
speed	1.00	0.806
dist	0.806	1.00

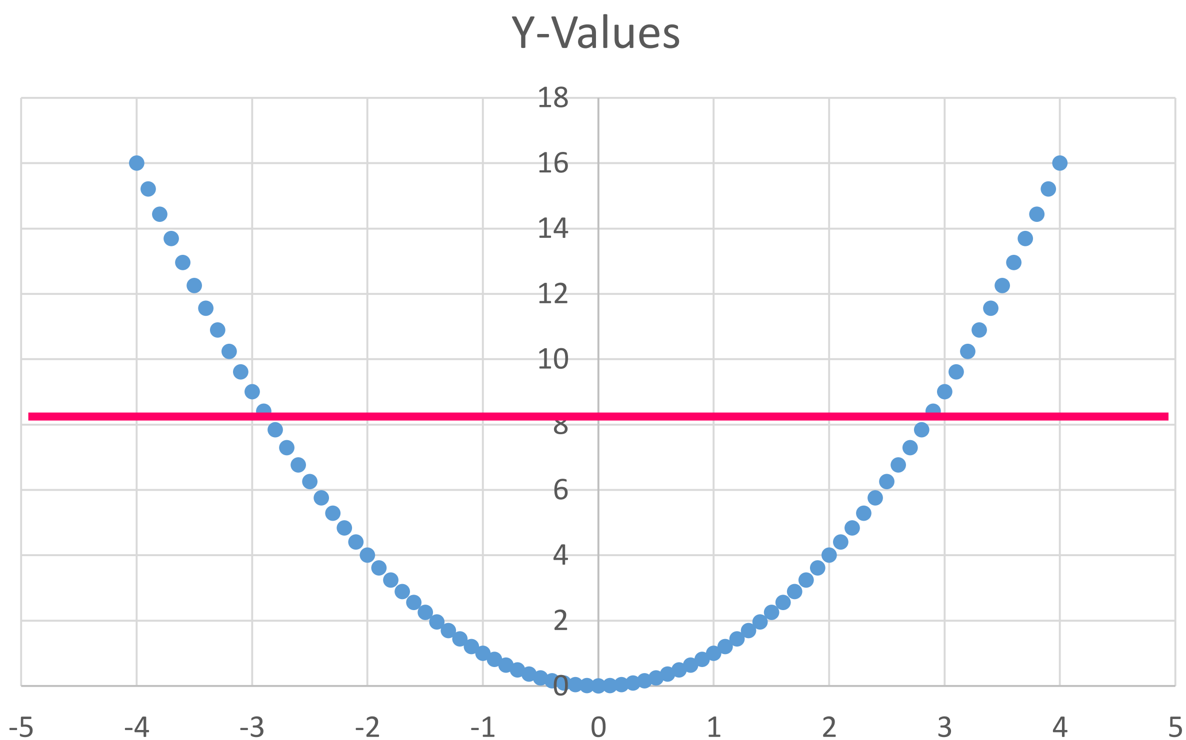
Correlation





Correlation

	A	B
1	-4	16
2	-3.9	15.21
3	-3.8	14.44
4	-3.7	13.69
5	-3.6	12.96
6	-3.5	12.25
7	-3.4	11.56
8	-3.3	10.89
9	-3.2	10.24
10	-3.1	9.61
11	-3	9
12	-2.9	8.41
13	-2.8	7.84
14	-2.7	7.29
15	-2.6	6.76
16	-2.5	6.25
17	-2.4	5.76
18	-2.3	5.29
19	-2.2	4.84
20	-2.1	4.41
21	-2	4
22	-1.9	3.61
23	-1.8	3.24



$$r = 0$$

Correlation coefficient (r) is ***0*** doesn't imply there is no relation
⇒ It implies there is no linear relationship

Coefficient of Determination

The coefficient of determination is given by r^2 or R^2. It is the percentage of variation in the *y* variable that is explainable by the x variable. For example, what percentage of the variation in **distance** is explainable by the **speed of car**.

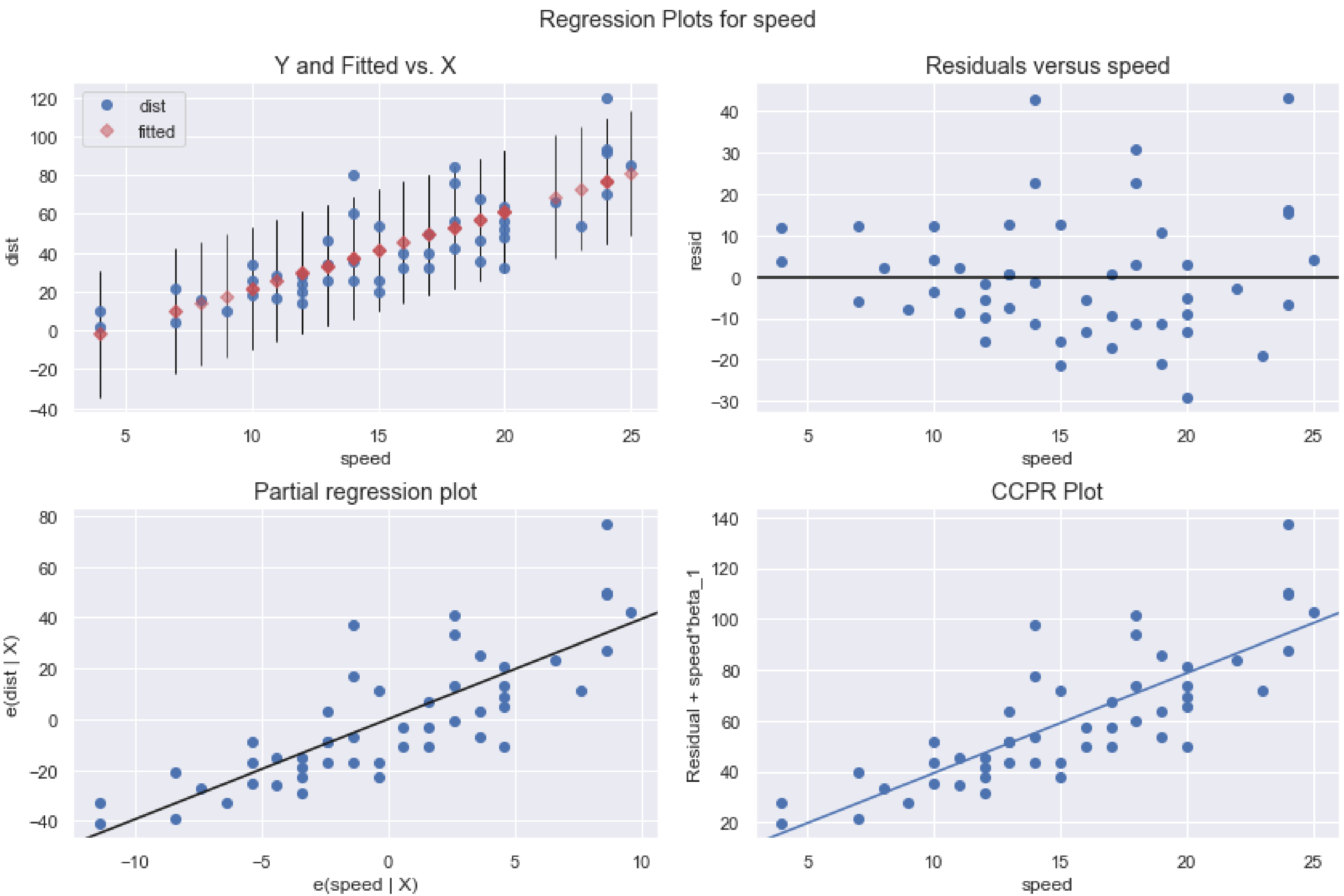
If $r^2 = 0$, it means you can't predict the *y* value from the x value.

If $r^2 = 1$, it means you can predict the *y* value from the x value without any errors.

Usually, r^2 is between these two extremes.

```
                       OLS Regression Results
==============================================================================
Dep. Variable:                   dist   R-squared:                       0.651
Model:                            OLS   Adj. R-squared:                  0.644
Method:                 Least Squares   F-statistic:                     89.57
Date:                Sun, 09 Dec 2018   Prob (F-statistic):           1.49e-12
Time:                        21:34:40   Log-Likelihood:                -206.58
No. Observations:                  50   AIC:                             417.2
Df Residuals:                      48   BIC:                             421.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -17.5791      6.758     -2.601      0.012     -31.168      -3.990
speed          3.9324      0.416      9.464      0.000       3.097       4.768
==============================================================================
Omnibus:                        8.975   Durbin-Watson:                   1.676
Prob(Omnibus):                  0.011   Jarque-Bera (JB):                8.189
Skew:                           0.885   Prob(JB):                       0.0167
Kurtosis:                       3.893   Cond. No.                         50.7
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```


Regression Plots for speed
Y and Fitted vs. X
dist
fitted
speed
Residuals versus speed
resid
speed
Partial regression plot
e(dist | X)
e(speed | X)
CCPR Plot
Residual + speed*beta_1
speed

Reference

- Head First Statistics
- Business-Statistics for contemporary decision making bu Ken Black