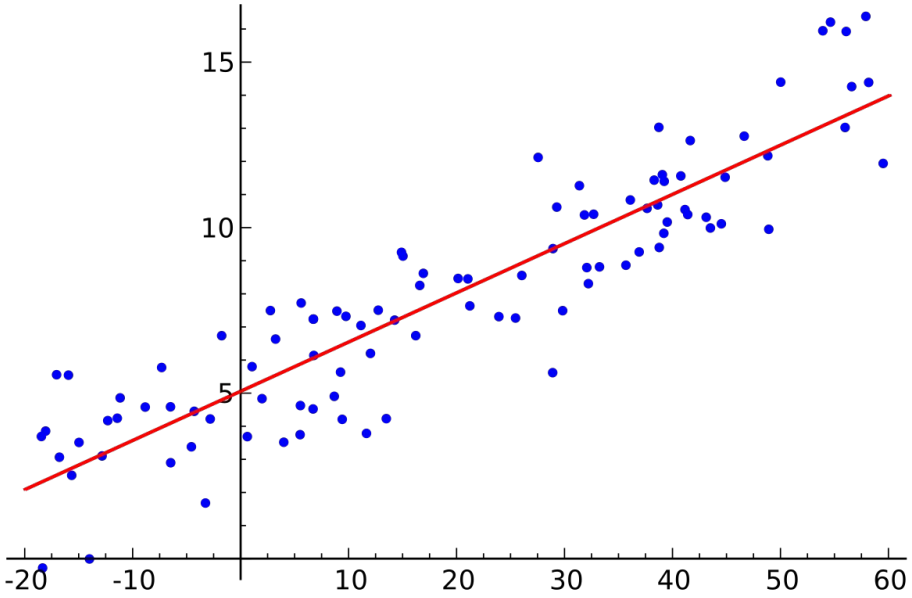
Simple linear regression

Some important aspects

- Independent variables:
  - Data that can be controlled directly.
- Dependent variables:
  - Data that cannot be controlled directly.
- The data that can’t be controlled i.e. dependent variables need to predicted or estimated.
- Model:
  - A model is a transformation engine that helps us to express dependent variables as a function of independent variables.

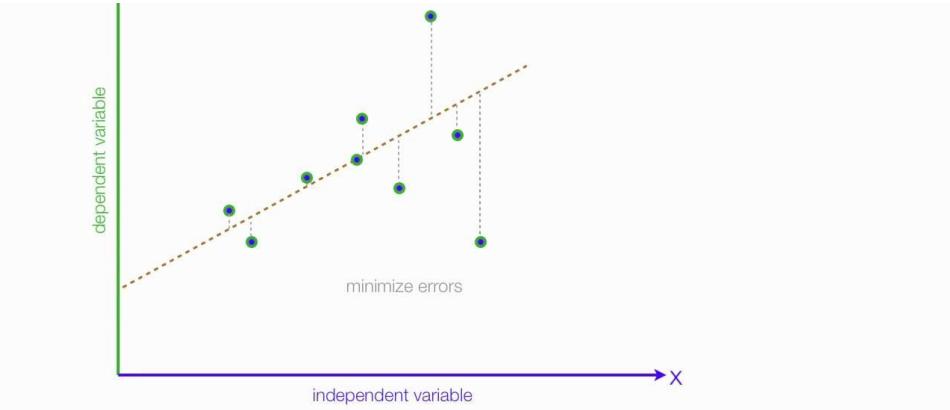
Linear regression

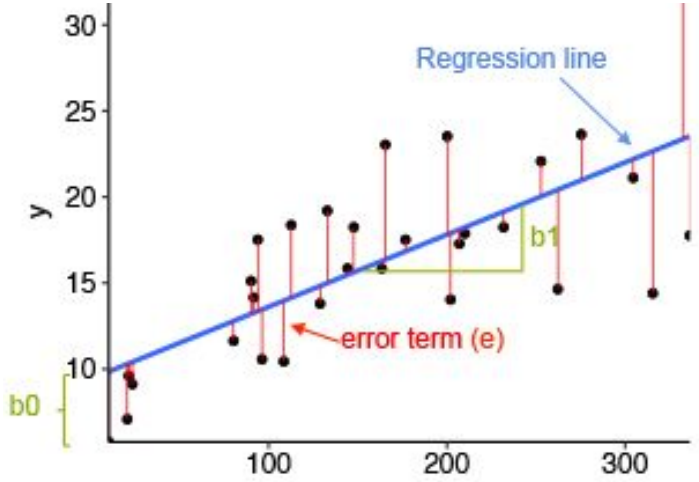
- Provide a simple approach towards supervised learning.
- They are simple yet effective.
- Linear suggests that the relationship between dependent and independent variable can be expressed in a straight line.
- Recall the geometry lesson from high school. What is the equation of a line?
  - y = mx + c



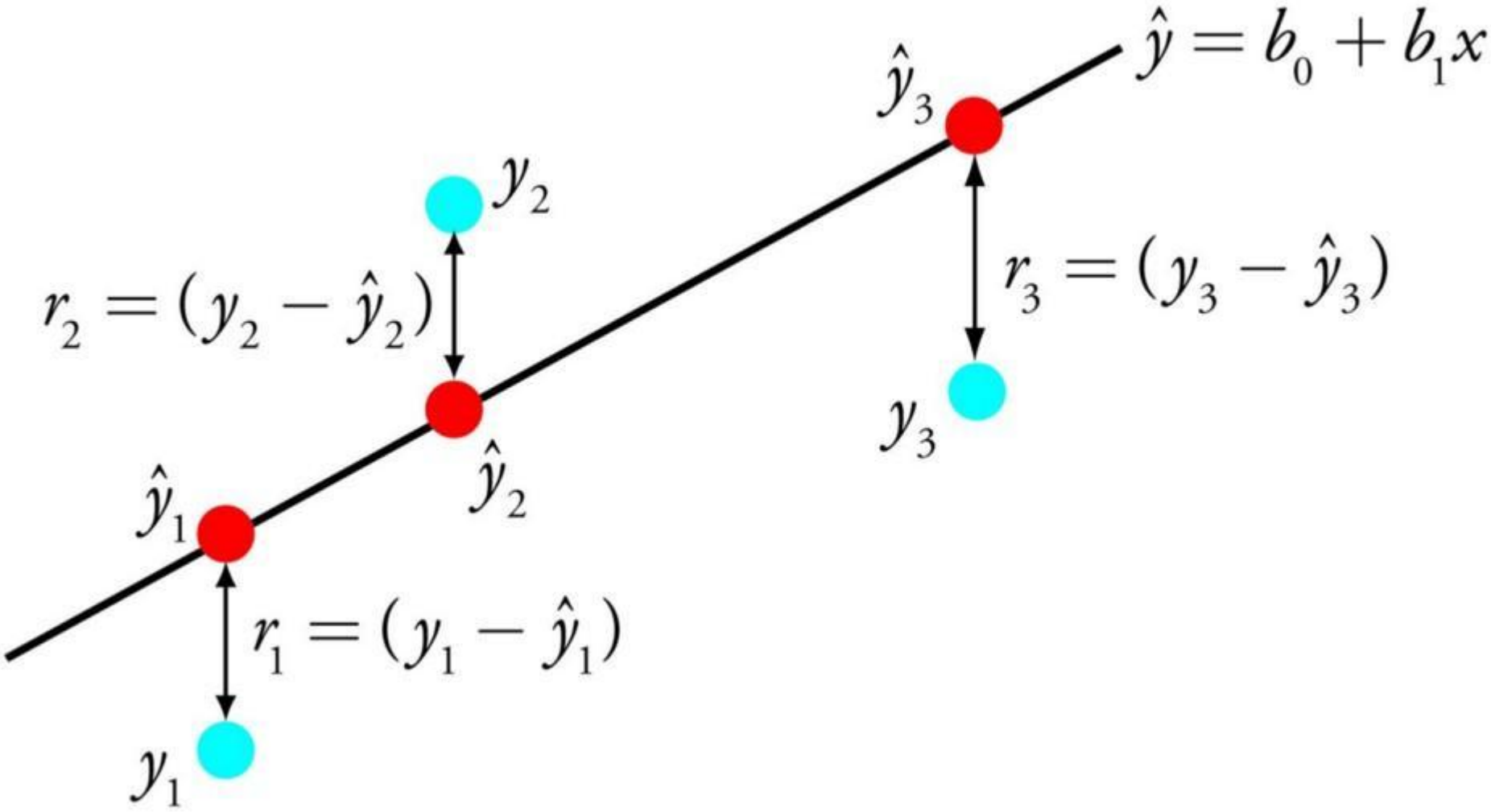
Linear regression

- y is the dependent variable i.e. the variable that needs to be estimated and predicted.
- x is the independent variable i.e. the variable that is controllable. It is the input.
- m is the slope. It determines what will be the angle of the line. It is the parameter denoted as β.
- c is the intercept. A constant that determines the value of y when x is 0.

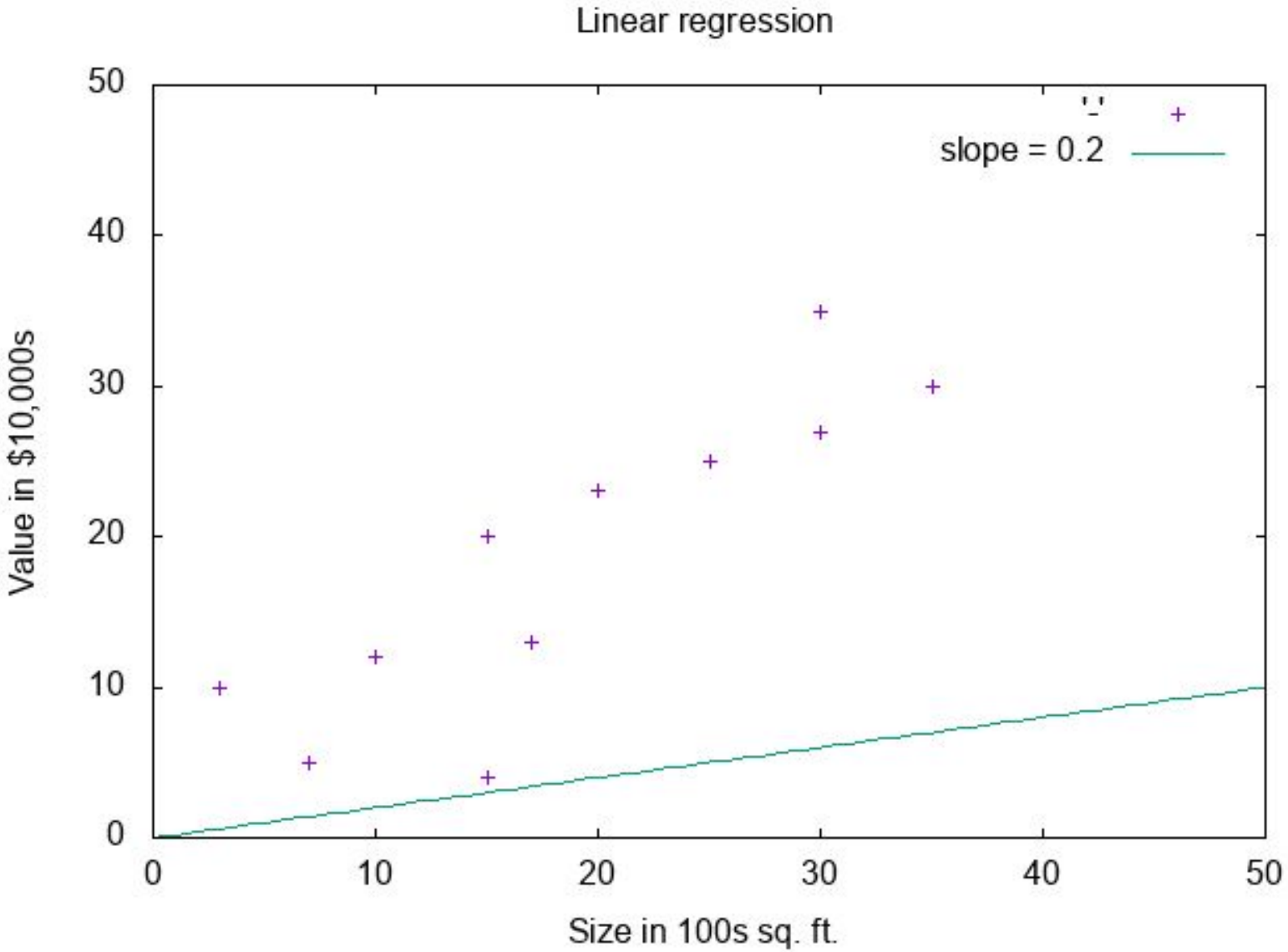




Linear regression



Linear regression



Linear regression Models

- Linear regression models are not perfect.
- It tries to approximate the relationship between dependent and independent variables in a straight line.
- Approximation leads to errors:
  - Some errors can be reduced.
  - Some errors are inherent in the nature of the problem. These errors cannot be eliminated. They are called as an irreducible error.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

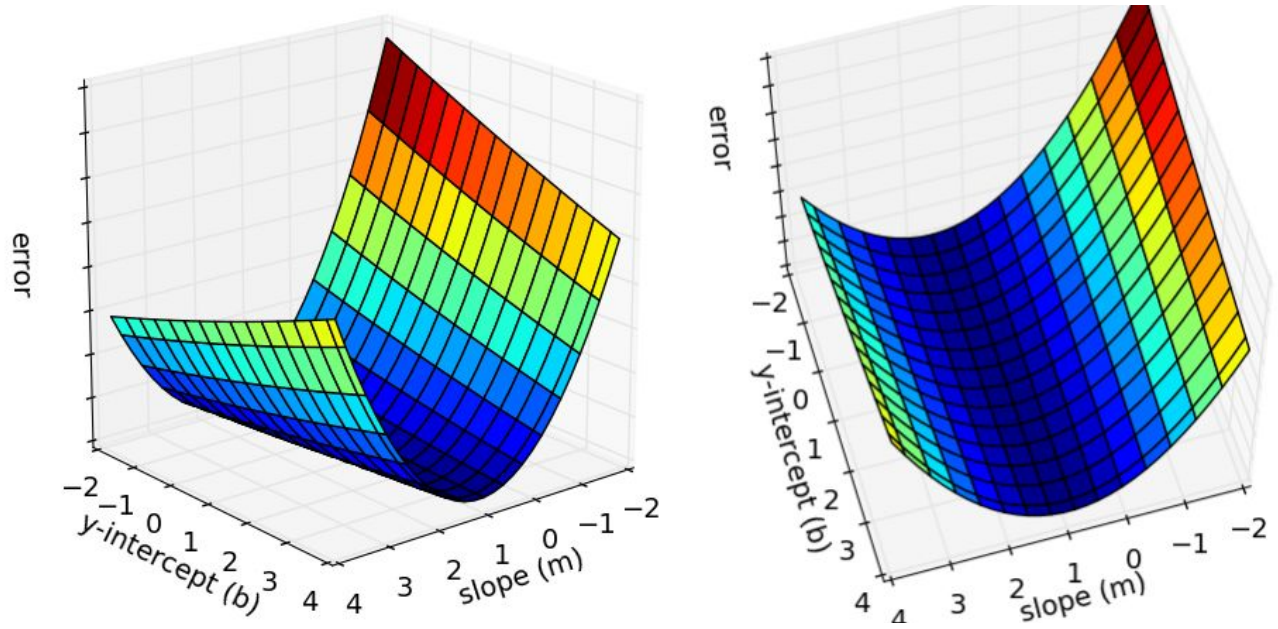
- β0 and β1 are two unknown constants that represent the intercept and slope. They are the parameters.
- ε is the error term.

Cost function

- A cost function is something you want to minimize. For example, your cost function might be the sum of squared errors over your training set.
- Cost function is also called squared error function.

$$error = (guess - actual) = (mx + b) - y$$

$$cost function : J(m, b) = \sum_{i=0}^{n} error^2$$



Formulation

- You need to predict the price of a new car. You have the following data:

make	fuelType	nDoors	engineSize	price
alfa-romero	gas	two	130	13495
alfa-romero	gas	two	130	16500
alfa-romero	gas	two	152	16500
audi	gas	four	109	13950
audi	gas	four	136	17450
audi	gas	two	136	15250
audi	gas	four	136	17710
audi	gas	four	136	18920
audi	gas	four	131	23875

Formulation

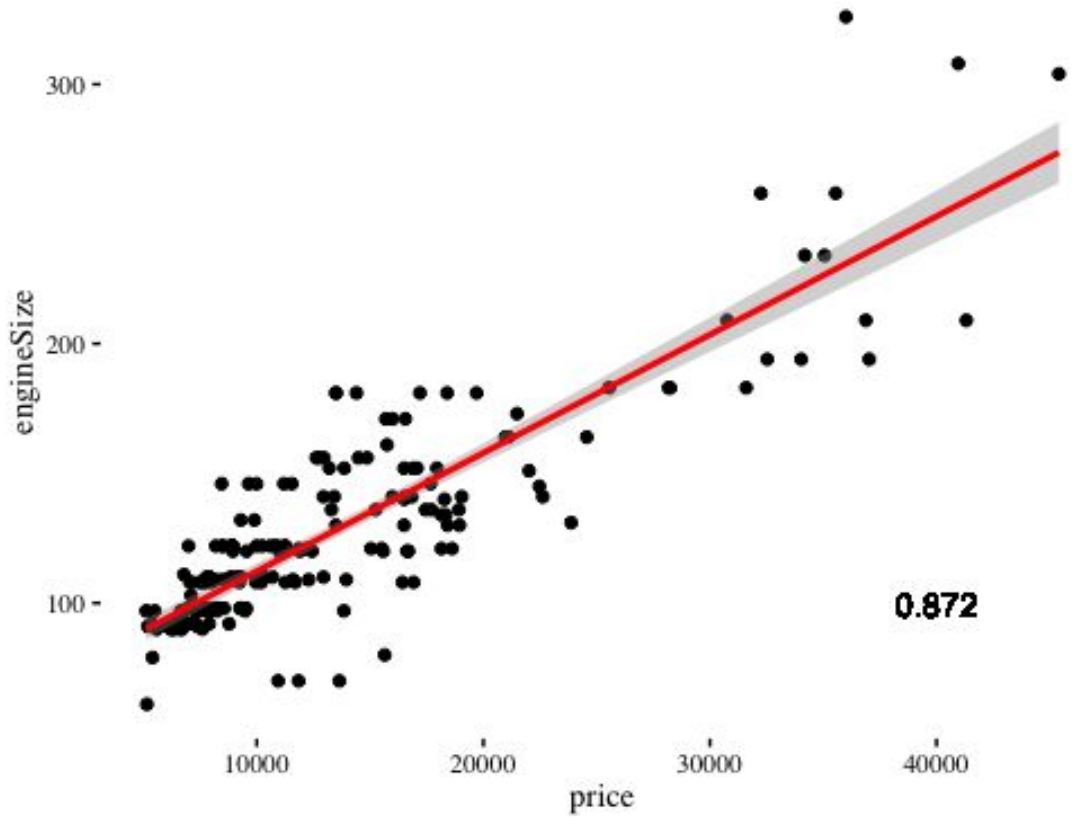
- The following are the data provided to him:
  - make: make of the car.
  - fuelType: type of fuel used by the car.
  - nDoor: number of doors.
  - engineSize: size of the engine of the car.
  - price: the price of the car.

make	fuelType	nDoors	engineSize	price
alfa-romero	gas	two	130	13495
alfa-romero	gas	two	130	16500
alfa-romero	gas	two	152	16500
audi	gas	four	109	13950
audi	gas	four	136	17450
audi	gas	two	136	15250
audi	gas	four	136	17710
audi	gas	four	136	18920
audi	gas	four	131	23875

- You want to evaluate if indeed he can predict car price based on engine size. The first set of analysis seeks the answers to the following questions:
  - Is price of car price related with engine size?
  - How strong is the relationship?
  - Is the relationship linear?
  - Can we predict/estimate car price based on engine size?

Formulation

- Let’s do a correlation analysis.
- Correlation is a measure of how much the two variables are related.

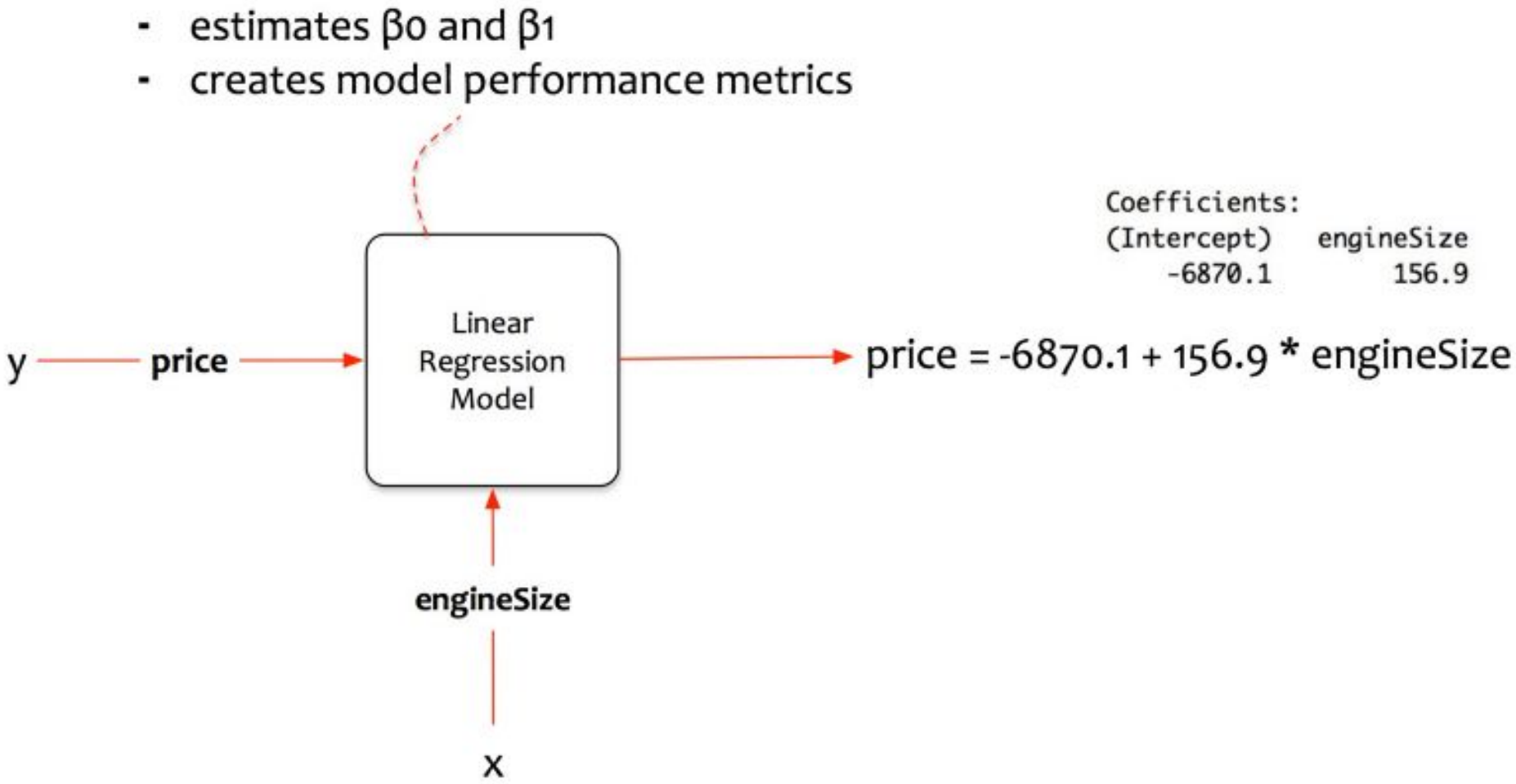


Formulation

- Following are the answers to the questions:
  - Is price of car price related with engine size?
    - Yes, there is a relationship.
  - How strong is the relationship?
    - There is a strong relationship.
  - Is the relationship linear?
    - Yes.
  - Can we predict/estimate the car price based on engine size?
    - Yes, car price can be estimated based on engine size.

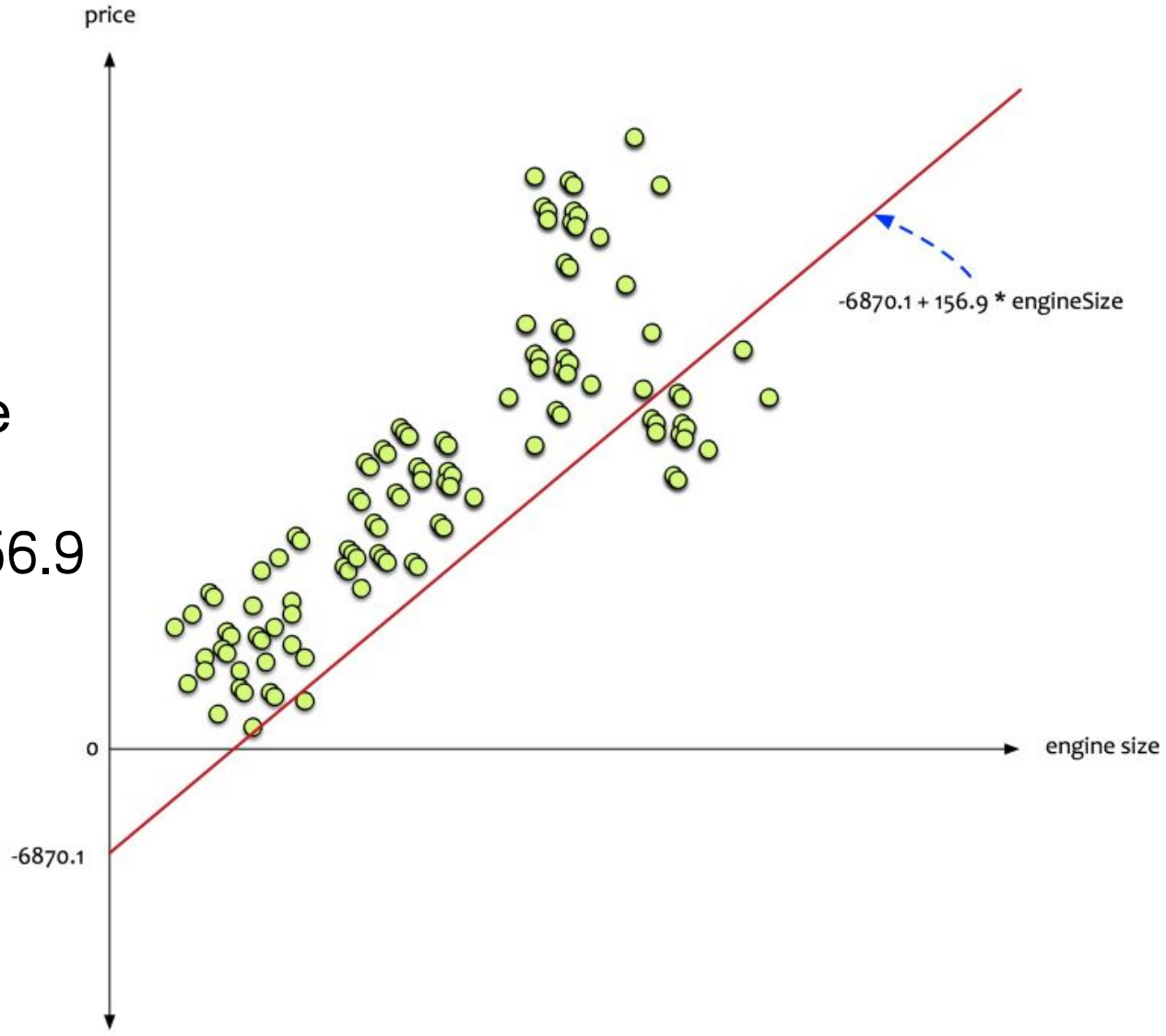
Equation for price prediction

- price = β0 + β1 x engine size



Interpretation

One unit increase in engine size will increase the average price of the car by 156.9 units.




Thank
you !!!