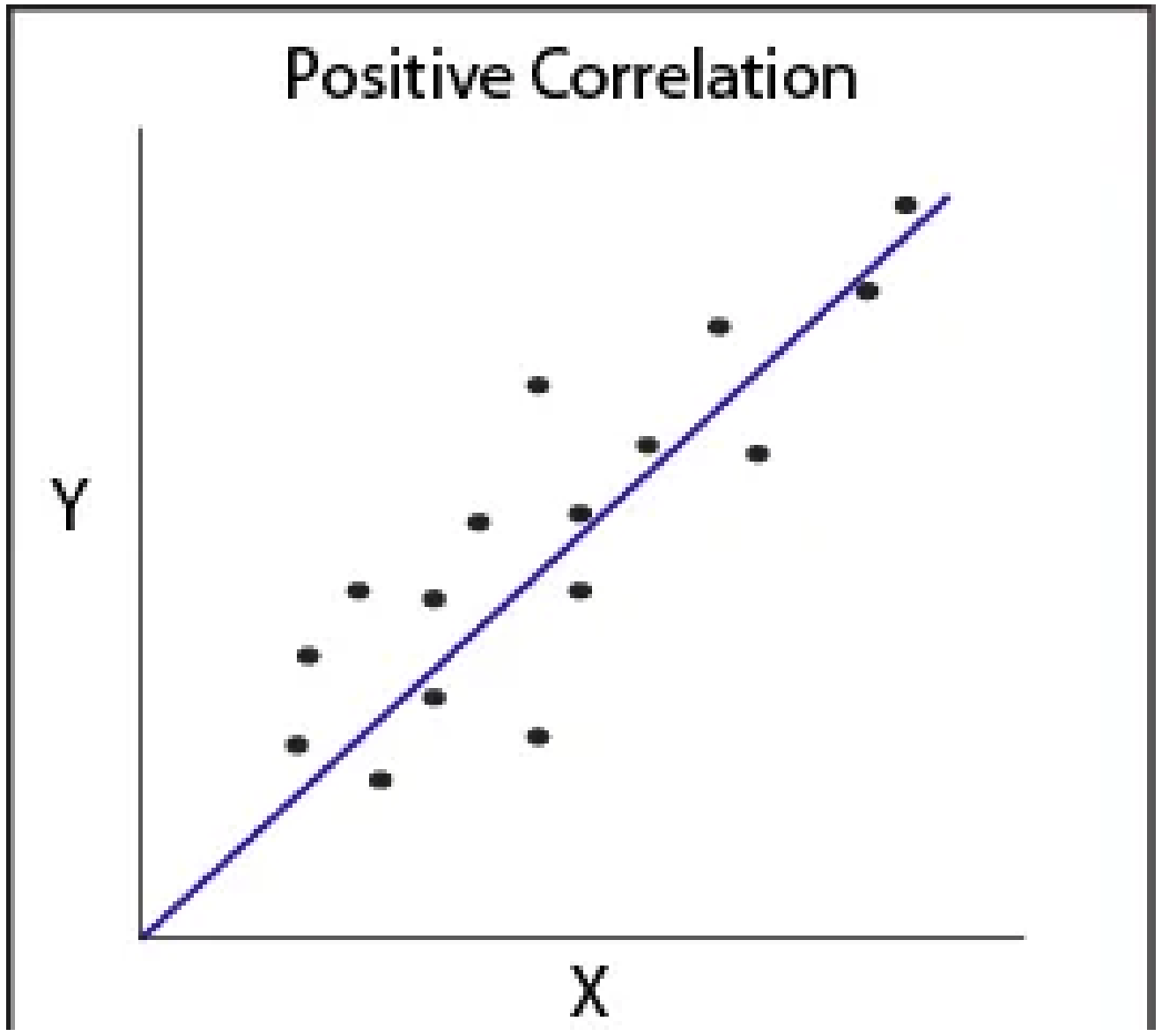
Unit 3

Syllabus

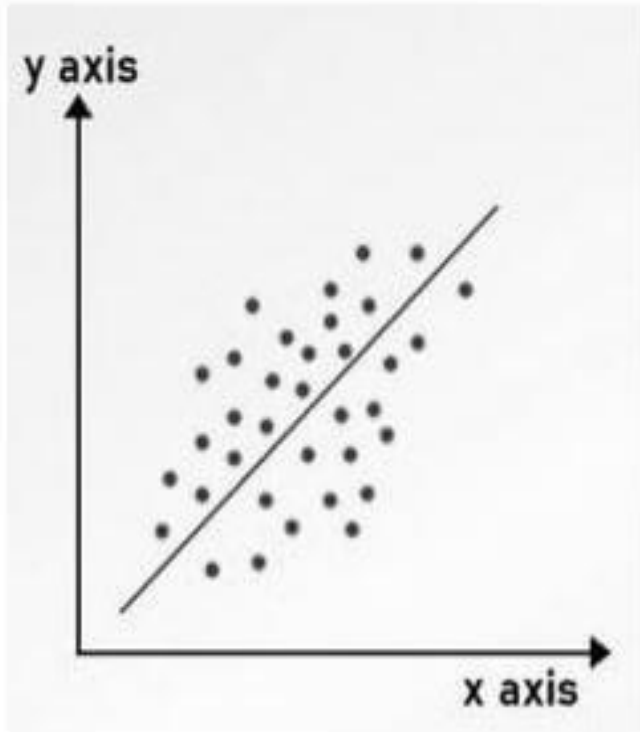
- **Concepts of Correlation**
- **Regression**
- **Linear Square Estimation**
- **Simple Linear Regression**
- **Multiple Regression**

What Is a Correlation?

- Correlation is a statistical term describing the degree to which two variables move in coordination with one another.
- If the two variables move in the same direction, then those variables are said to have a **positive correlation**.
- If they move in opposite directions, then they have a **negative correlation.**


Positive Correlation
Y
X

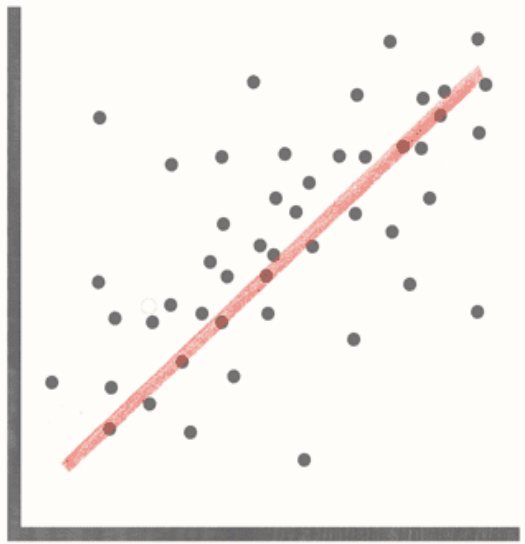
Positive Correlation



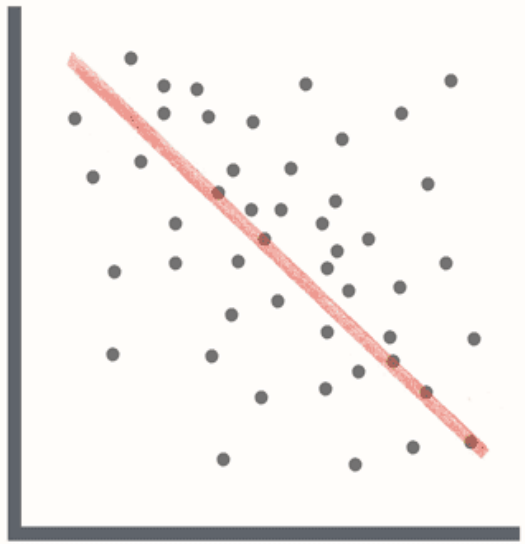
Types

- ❑ Strong Correlation (+1.0)
- ❑ Medium Correlation (+0.5)
- ❑ Low Correlation (+0.2)

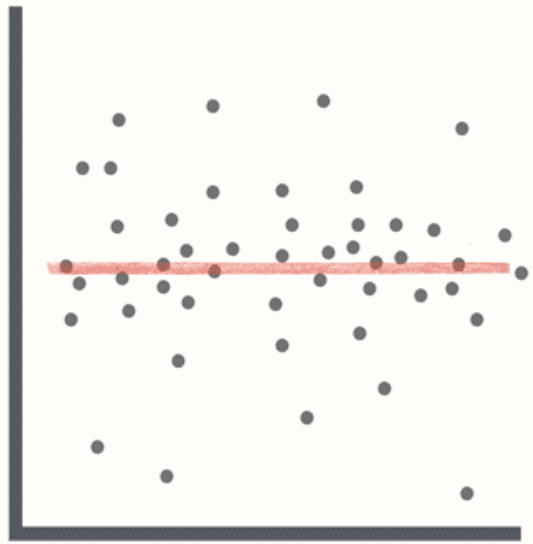
Correlation Coefficient



Positive Correlation



Negative Correlation



No Correlation

Introduction to Linear Regression

- Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables.
- Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation −

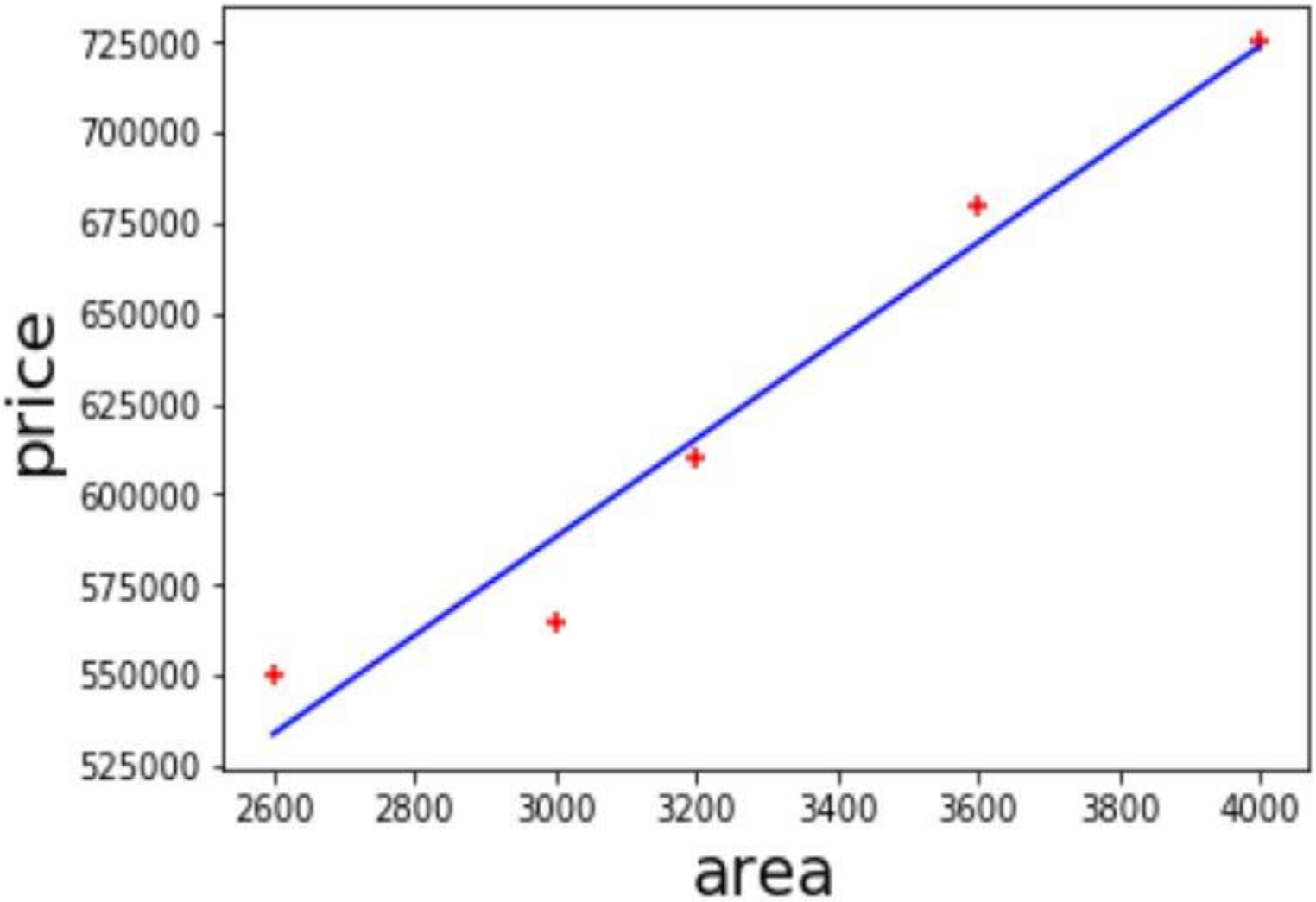
$$Y=mX+b$$

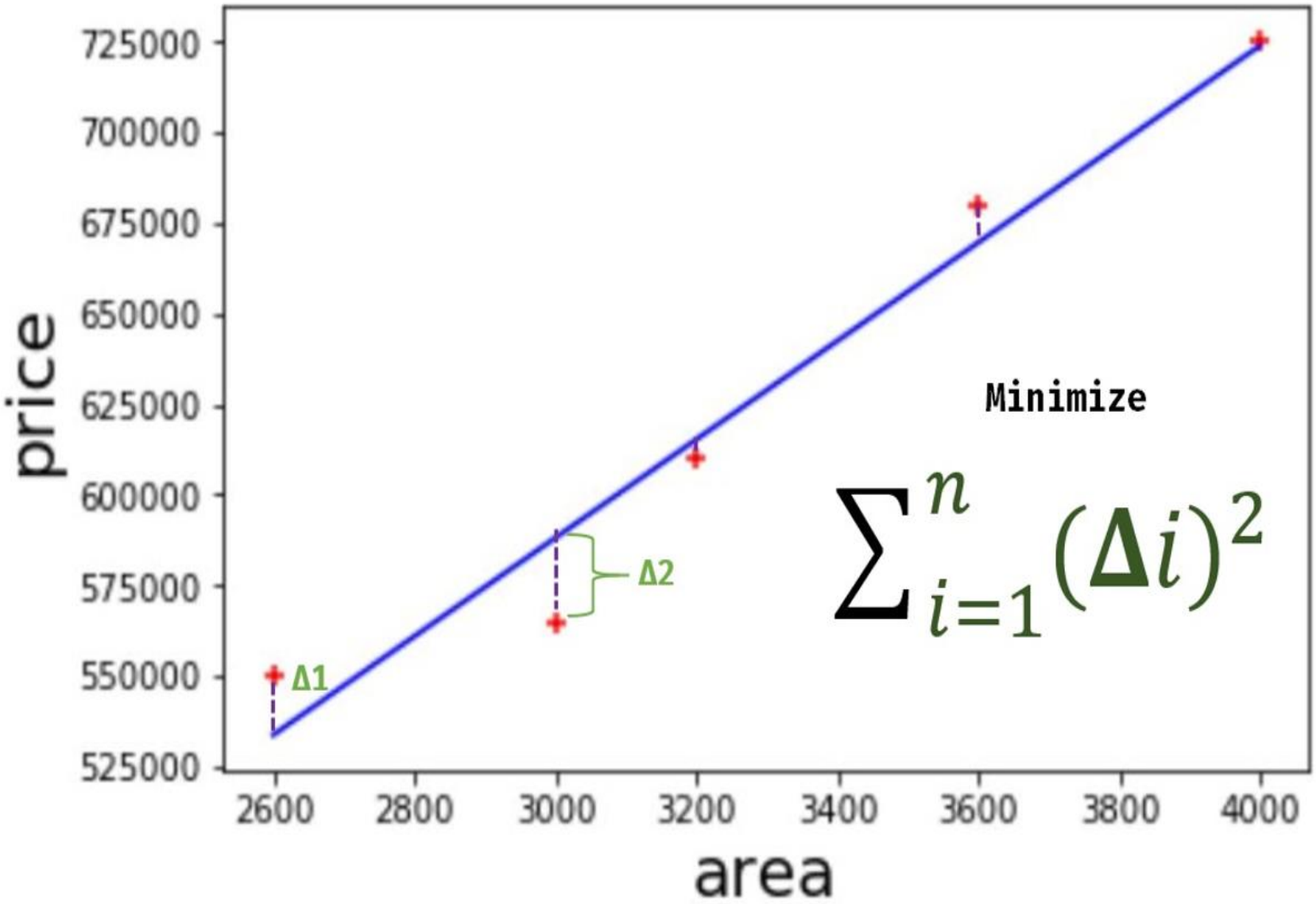
Here,

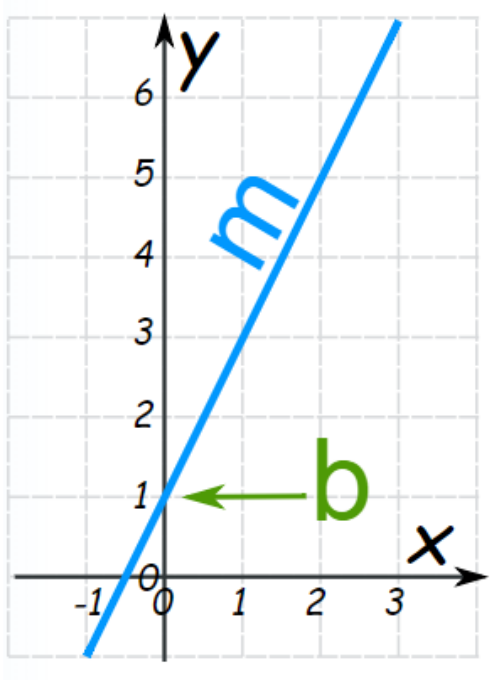
- Y is the dependent variable we are trying to predict.
- X is the independent variable we are using to make predictions.
- m is the slop of the regression line which represents the effect X has on Y
- b is a constant, known as the XY-intercept.
- If X = 0,Y would be equal to b.

Sample problem of predicting home price in monroe, new jersey (USA)

area	price
2600	550000
3000	565000
3200	610000
3600	680000
4000	725000


725000
700000
675000
650000
625000
600000
575000
550000
525000
price
2600
2800
3000
3200
3400
3600
3800
4000
area


725000
700000
675000
650000
625000
600000
575000
550000
525000
price
2600
2800
3000
3200
3400
3600
3800
4000
area
Δ1
Δ2
Minimize
$\sum_{i=1}^{n} (\Delta i)^2$



price = m * area + b

$$y = mx + b$$

Slope (or Gradient) Y Intercept

Reference: **https://www.mathsisfun.com/algebra/linear-equations.html**

Multiple Linear Regression

A linear regression model that contains more than one predictor variable is called a *multiple linear regression model*.

$$\mathbf{Y=m_1X_1+m_2X_2+b}$$

Where

Y is a dependent variable

X_1 and X_2 are independent variable.

m_1 and m_2 are coefficient.

b is intercept or constant.

Examples:

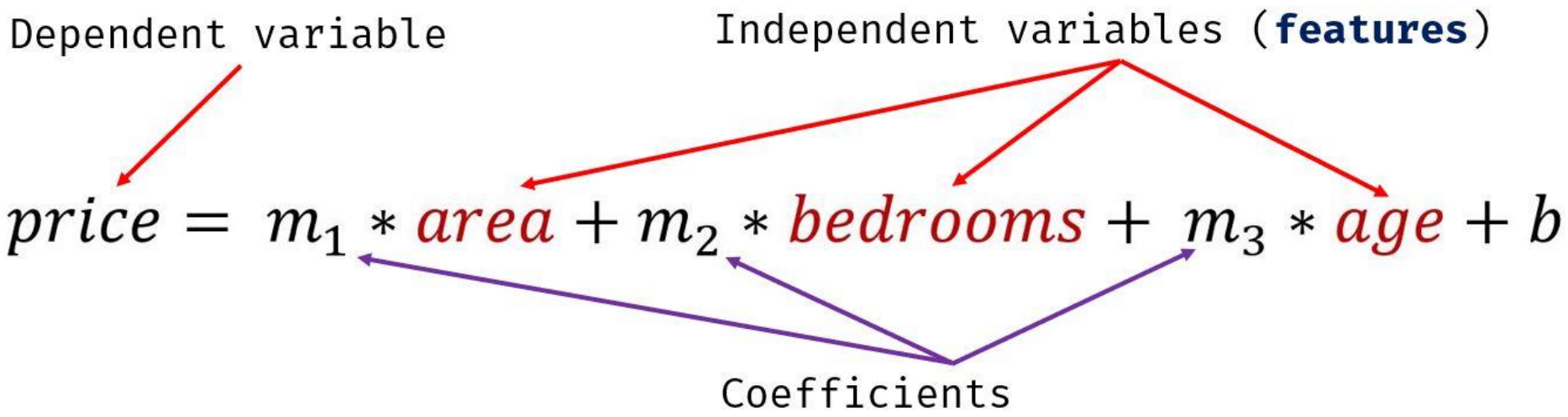
- The selling price of a house can depend on the desirability of the location, the number of bedrooms, the year the house was built, the square area and a number of other factors.
- The height of a child can depend on the height of the mother, the height of the father, nutrition, and environmental factors.

Sample problem of predicting home price

area	bedrooms	age	price
2600	3	20	550000
3000	4	15	565000
3200		18	610000
3600	3	30	595000
4000	5	8	760000
4100	6	8	810000

Here price depends on area (square feet), bed rooms and age of the home (in years).

Price can be calculated using following equation,



$$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + b$$

Find out price of a home that has,

3000 sqr ft area, 3 bedrooms, 40 year old

2500 sqr ft area, 4 bedrooms, 5 year old

$$price = m_1 * area + m_2 * bedrooms + m_3 * age + b$$

Regression Model

```
df = pd.read_csv('homeprices.csv')
df
```

area	bedrooms	age	price
2600	3	20	550000
3000	4	15	565000
3200		18	610000
3600	3	30	595000
4000	5	8	760000
4100	6	8	810000

```
reg = linear_model.LinearRegression()

reg.fit(df.drop('price',axis='columns'),df.price)

reg.coef_

array([112.06244194, 23388.88007794, -3231.71790863])
```

m1= 112.06244194

m2= 23388.88007794

m3=-3231.71790863

reg.intercept_

221323.00186540408

b=221323.00186540408

$$price = m_1 * area + m_2 * bedrooms + m_3 * age + b$$

Price=112.06244194*area+23388.88007794*bedrooms+(3231.71790863)*age+221323.00186540408

Example:

Experience	Test_score(out of 10)	Interview_score(out of 10)	Salary($)
3	8	9	50000
2	8	6	45000
5	6	7	60000
2	10	10	65000
7	9	6	70000
3	7	10	62000
10	6	7	72000
11	7	8	80000

Question ?

- **Predict salary whose experience is 2, test score is 9 and interview score is 6.**
- **Predict salary whose experience is 12, test score is 10 and interview score is 10.**

$$Y=m_1X_1+m_2X_2+b$$

salary=m1*experience+m2*test_score+m3*interview_score+b

```
d = pd.read_csv("hiring-1.csv")
d
```

```
reg = linear_model.LinearRegression()
reg.fit(d[['experience','test_score(out of 10)','interview_score(out of 10)']],d['salary($)'])
```

reg.coef_

array([3489.78042144, 2361.95659052, 2515.2819085])

reg.intercept_

6424.666202619774

Predict salary whose experience is 2, test score is 9 and interview score is 6.

reg.predict([[2,9,6]])

Manual Calculation is:

salary=m1*experience+m2*test_score+m3*interview_score+b

3489.78042144*2+2361.95659052*9+2515.2819085*6+6424.666202619774

Advantages Regression Models

1. Regression models are easy to understand as they are built upon basic statistical principles such as correlation and least square error.

2. Regression models provide simple algebraic equations that are easy to understand and use.

3. The strength (or the goodness of fit) of the regression model is measured in terms of the correlation coefficients, and other related statistical parameters that are well understood.

4. Regression models can match and beat the predictive power of other modeling techniques.

5. Regression models can include all the variables that one wants to include in the model.

6. Regression modeling tools are pervasive. They are found in statistical packages as well as data mining packages. MS Excel spreadsheets can also provide simple regression modeling capabilities.

Thank You