Welcome to

# Linear Regression

**Day 2**

# ABOUT INSTRUCTOR - Abhinav Singh

| | | |
|---|---|---|
| 2016 | **CloudxLab** | Building platform for practicing Big Data Technologies |
| 2015 | **Byjus** | #1 Edtech application in India on Play Store |
| 2015 | | |
| 2013 | **Specadel** | Developed platform for disrupting indian education |
| 2012 | | |
| 2009 | **HashCube** | Developed #1 Sudoku Game on Facebook |

# Linear Regression

When you think of Regression, think prediction. A regression uses the historical relationship between an independent and a dependent variable to predict the future values of the dependent variable.

Two types

1. Simple linear Regression
2. Multiple Linear Regression

# Application of Linear Regression

- Predicting future sales,
- Predicting Stock prices
- Predicting currency exchange rates and productivity gains resulting from a training program

**Sales of a company = function(Marketing budget, Pricing, Season, Discount)**

**Independent variables** are variable which are used to explain certain phenomena like marketing budget, pricing, season, discount etc

**Dependent variables** are driven by independent variables like sales of a company

# Simple Linear Regression

**Problem**

Is there any relationship between wages and number of years spent in education?

**Dependent Variable -** wages
**Independent Variable -** number of years spent in education

Wages(dependent variable) = (Y-Intercept) + β1Education(Independent Variable)

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

# Multiple Linear Regression

**Problem**

Is there any relationship between wages and number of years spent in education, age, gender, worx experience?

**Dependent Variable -** wages

**Independent Variable -** number of years spent in education, age, gender, work experience etc

$$\text{Wages(dependent variable)} = (\text{Y-Intercept}) + \beta 2(\text{Education}) + \beta 3(\text{age}) + \beta 4(\text{Gender}) + \beta 5(\text{Work Experience}) + \beta 6(\text{Sector})$$

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 4X4 + \beta 5X5$$

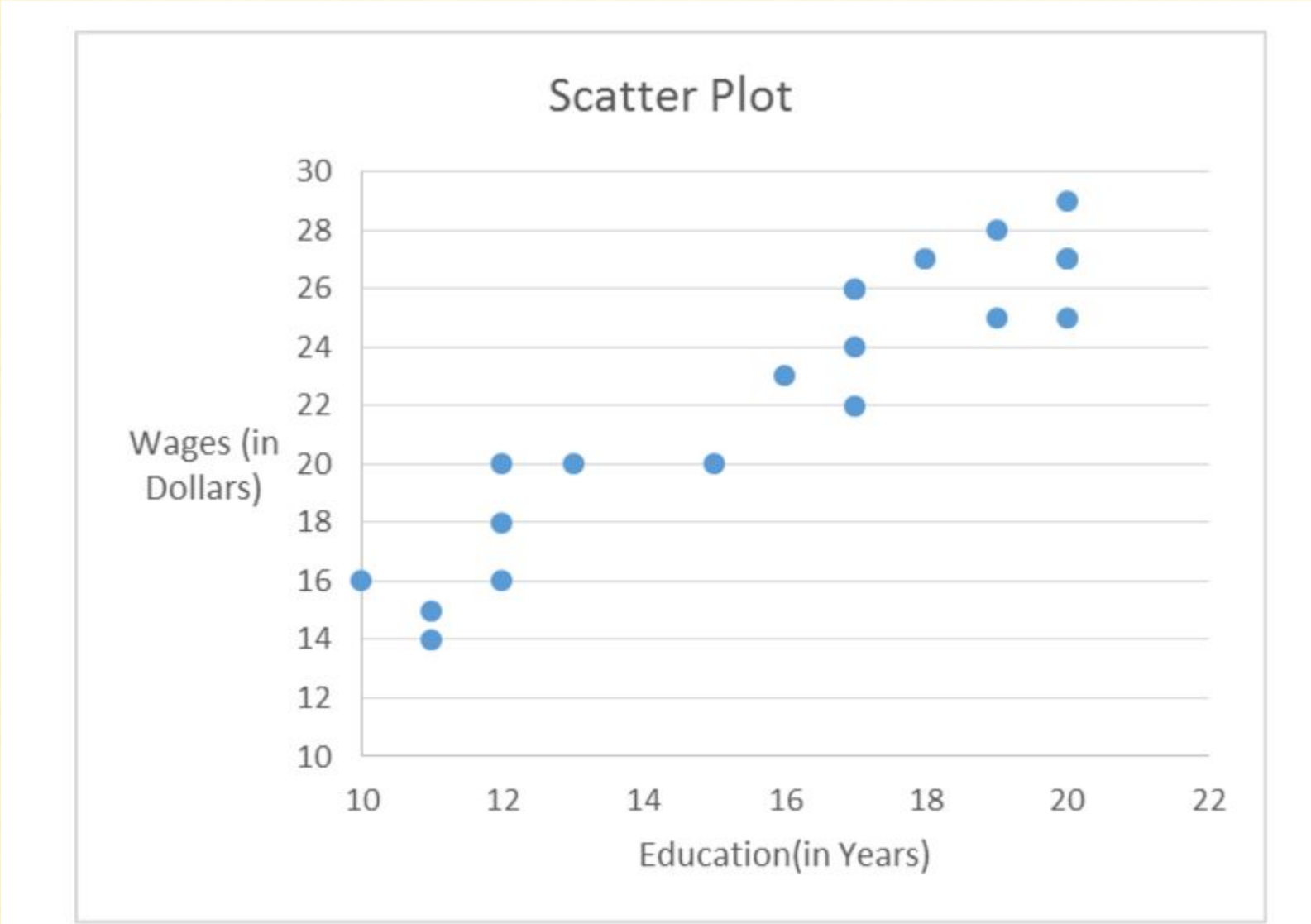
# Linear Regression - Problem

Given a data how do we find out if we can apply Linear Regression?

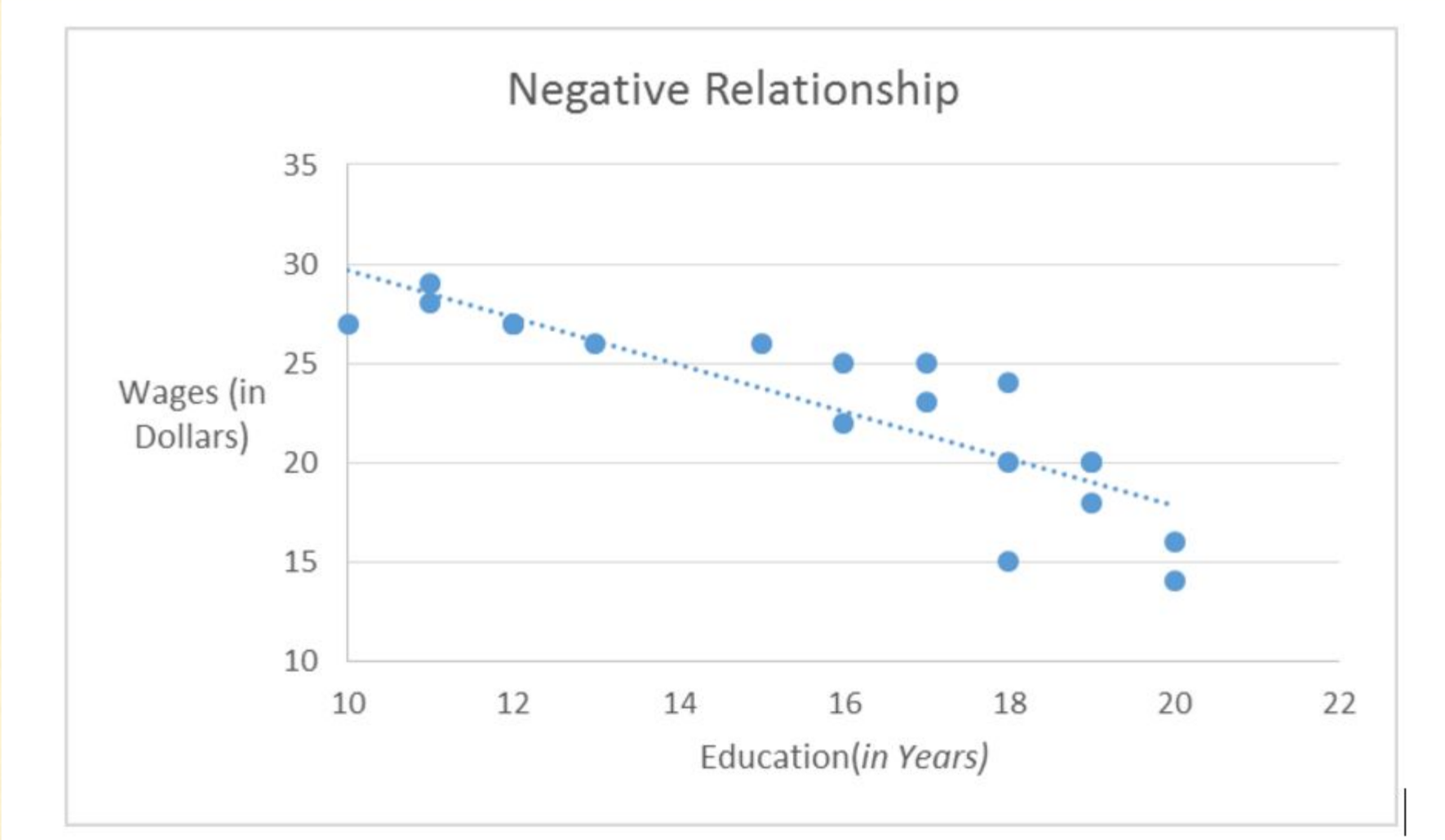
# Linear Regression - Solution

1) Find the Correlation between two variables and if correlation is closer to one it means that we can model the linear regression
2) Plot the Scatter plot between two variables and see if there is a line

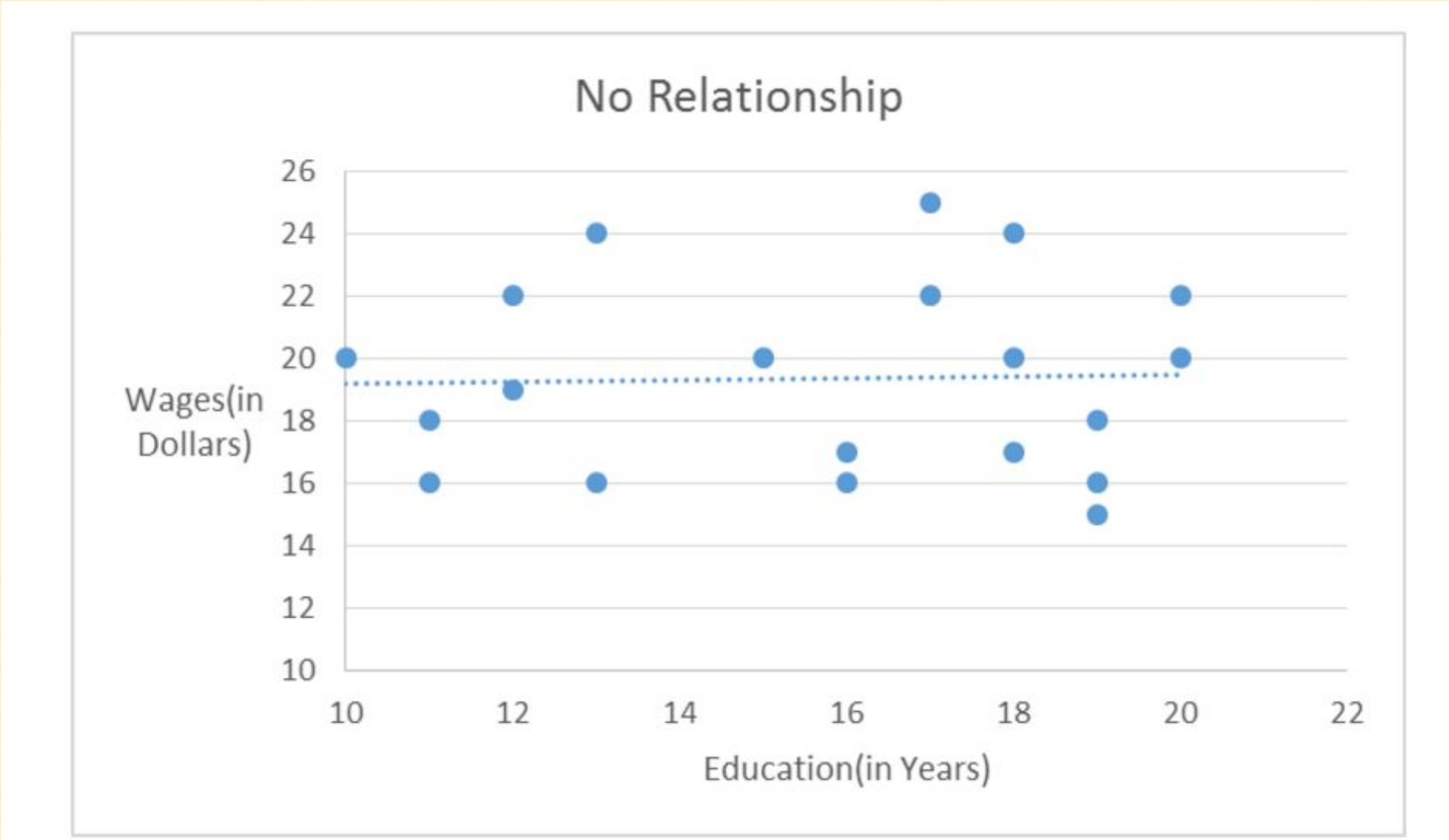
# Linear Regression - Positive Relationship



# Linear Regression - Negative Relationship



# Linear Regression - No Relationship



# Linear Regression - Hand-on

1) Go to equiskill lab http://lab1.equiskill.com
2) Open test_for_linear_regression.R
3) Load data wages_education.csv
4) Plot the graph
5) Find correlation

# Linear Regression - First model

Let's build our first regression model

1) Go to equiskill lab http://lab1.equiskill.com
2) Open wages_education.R
3) Load data wages_education.csv
4) Fit the model
5) Find Coefficients and Y-intercept
6) Make Prediction

# Linear Regression - First model

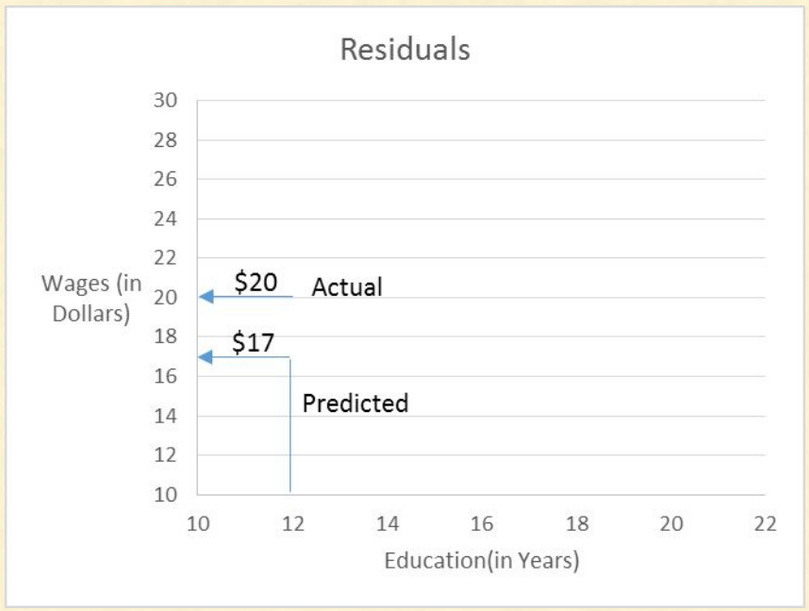
Our wages equation will be

Wages = 2.372 (Y-intercept) + 1.267 Education

# Linear Regression - Residuals

**Residuals are the difference between the actual value and the predicted value.**

Suppose as per our predictions, the wage for a professional who has a 12 years of education(Let say #11 from table) which is $17 per hour. Actually the wage of that professional is $20 per hour. So Residual will be $3 which is represented as μ



# Linear Regression - Residuals

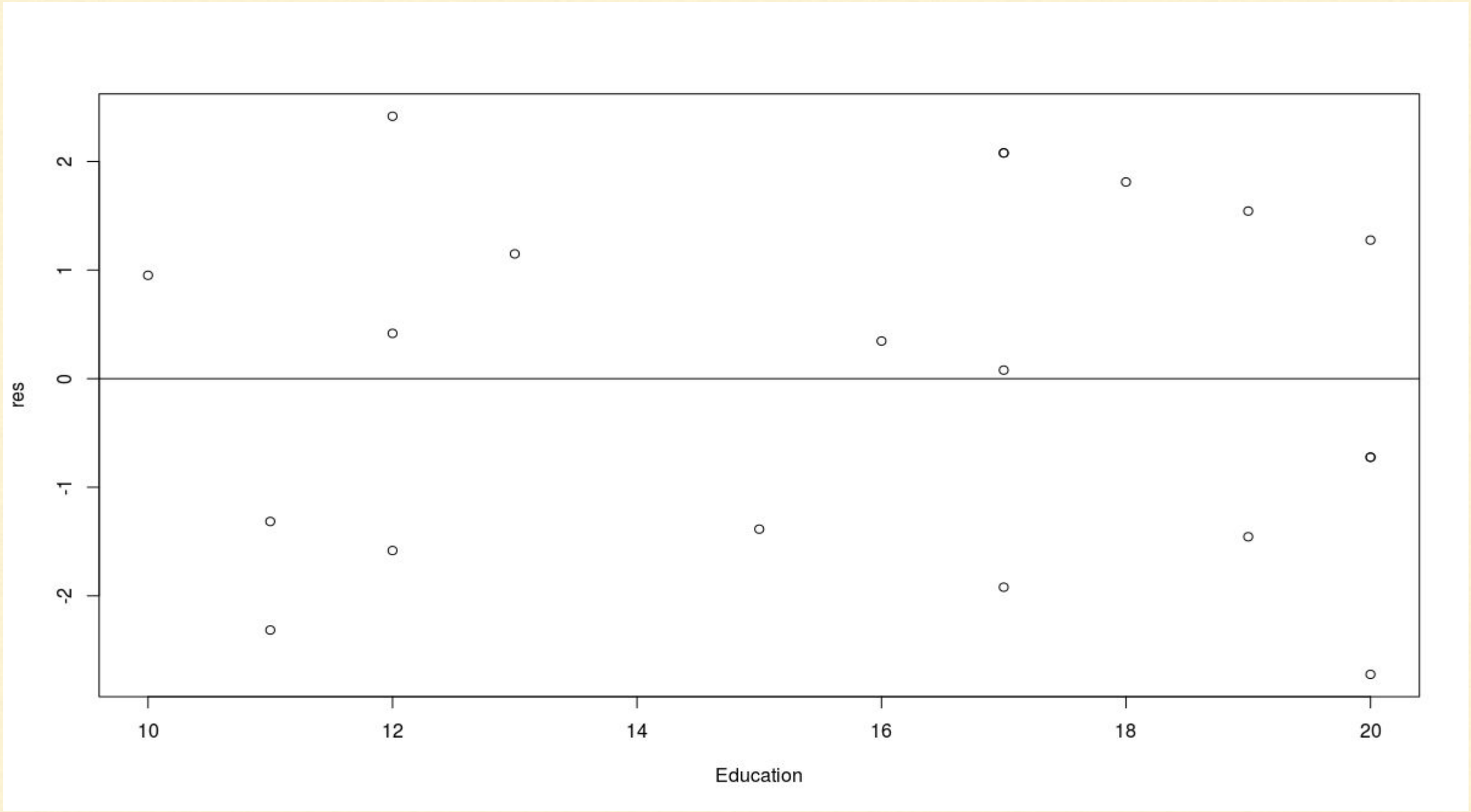
Points to note

- We can apply linear regression only when dependent variable is continuous
- Independent variables can be numerical and categorical

# Linear Regression - Good fit

- Residuals are randomly distributed
- There is no relationship between residuals

# Linear Regression - Residual plot



# Linear Regression - Perfect fit

- When coefficient of determination $r^2 = 1$

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

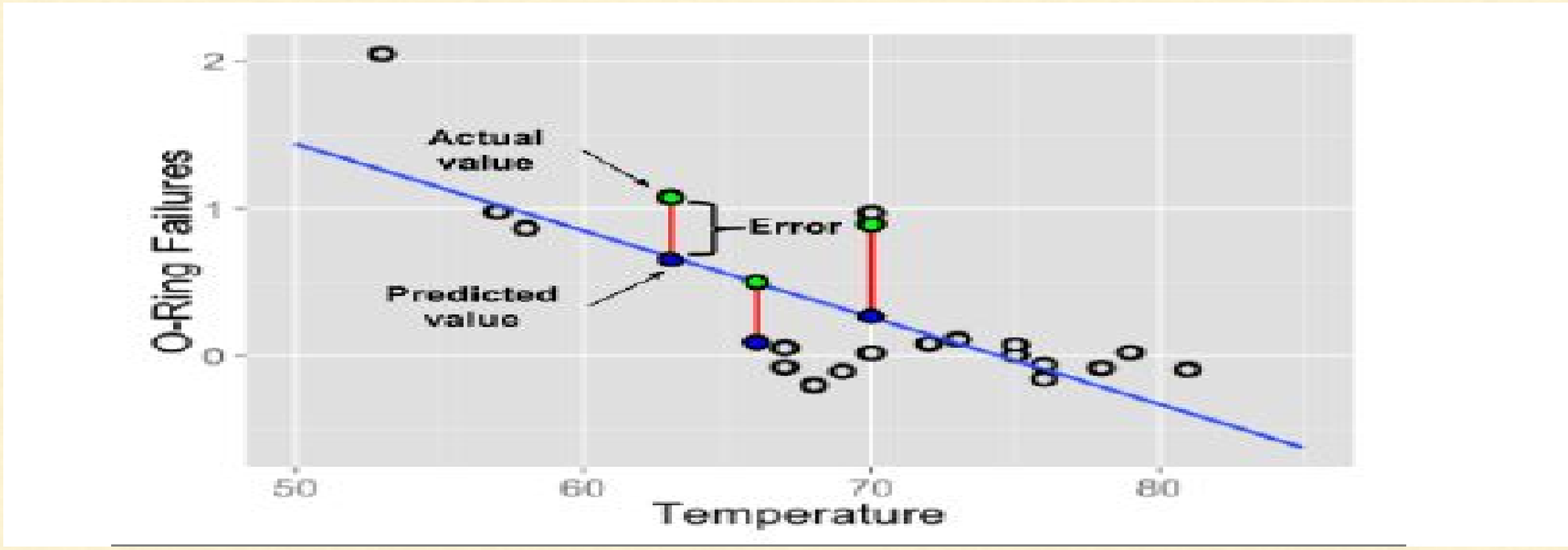
Where $\hat{Y}_i$ = Predicted values

$\bar{Y}$ = Mean of actual values

$Y_i$ = Actual values

# ORDINARY LEAST SQUARES ESTIMATION

In OLS regression, the slope and intercept are chosen such that they minimize the sum of the squared errors




$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

# Simple Linear Regression - Hands-on

- We will explore faithful dataset provided by R
- Go to Equiskill lab http://lab1.equiskill.com
- Check faithful dataset, we will use this dataset for simple linear regression
- ?faithful
- Open file simple_linear_regression.R

# Simple Linear Regression - Hands-on

**faithful dataset**

Waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA.

# Simple Linear Regression - Hands-on

**faithful dataset**

[,1] eruptions numeric Eruption time in mins
[,2] waiting numeric Waiting time to next eruption (in mins)

| | eruptions | waiting |
|---|---|---|
| 1 | 3.600 | 79 |
| 2 | 1.800 | 54 |
| 3 | 3.333 | 74 |
| 4 | 2.283 | 62 |
| 5 | 4.533 | 85 |
| 6 | 2.883 | 55 |
| 7 | 4.700 | 88 |
| 8 | 3.600 | 85 |
| 9 | 1.950 | 51 |
| 10 | 4.350 | 85 |
| 11 | 1.833 | 54 |
| 12 | 3.917 | 84 |
| 13 | 4.200 | 78 |

# Simple Linear Regression - Hands-on

**Problem 1-**

Apply the simple linear regression model for the data set faithful, and estimate the next eruption duration if the waiting time since the last eruption has been 80 minutes.

Steps -

1. Fit the model
2. Predict next eruption duration manually using equation
3. Predict next eruption duration using predict function

# Simple Linear Regression - Hands-on

**Problem 2-**

Find the coefficient of determination for the simple linear regression model of the data set faithful.

Steps -

1. Fit the model
2. Find coefficient of determination

# Simple Linear Regression - Hands-on

**Problem 3-**

Plot the residual of the simple linear regression model of the data set faithful against the independent variable waiting.

Steps -

1. Fit the model
2. Find residual
3. Plot residual

# Multiple Linear Regression - Hands-on

- We will explore stackloss dataset provided by R
- Go to Equiskill lab http://lab1.equiskill.com
- Check stackloss dataset, we will use this dataset for multiple linear regression
- ?stackloss
- Open file multiple_linear_regression.R

# Multiple Linear Regression - Hands-on

**stackloss dataset**

Operational data of a plant for the oxidation of ammonia to nitric acid.

[,1] 'Air Flow' Flow of cooling air
[,2] 'Water Temp' Cooling Water Inlet Temperature
[,3] 'Acid Conc.' Concentration of acid [per 1000, minus 500]
[,4] 'stack.loss' Stack loss

# Multiple Linear Regression - Hands-on

**Problem 1-**

Apply the multiple linear regression model for the data set stackloss, and predict the stack loss if the air flow is 72, water temperature is 20 and acid concentration is 85.

Steps -

1. Fit the model
2. Predict stack loss manually using equation
3. Predict stack loss using predict function

# Multiple Linear Regression - Hands-on

**Problem 2-**

Find the coefficient of determination for the multiple linear regression model of the data set stackloss.

Steps -

1. Fit the model
2. Find the coefficient of determination