DATA SCIENCE FOR ENGINEERS

Week 6

Session Co-Ordinator : Abhijit Bhakte

Regression

- Regression is a supervised learning that **build functional relationship between dependent and independent variables**

Property Rent Price

Stock Price

Regression Example

➢ House rent price prediction

Model

Rent is 10000 INR

- Location
- Size and area
- Transportation
- Furnished
- Utilities
- Pet allowed/not allowed

Independent Features

Dependent Features

Regression Types

❑Univariate Vs Multivariate

- **Univariate:** One dependent and one independent variable
- ***Multivariate:*** *Multiple independent and multiple dependent variables*

Square Footage (X)	House Price (Y)
1500	$250,000
1800	$280,000
1200	$220,000
2000	$320,000
1350	$240,000

Univariate

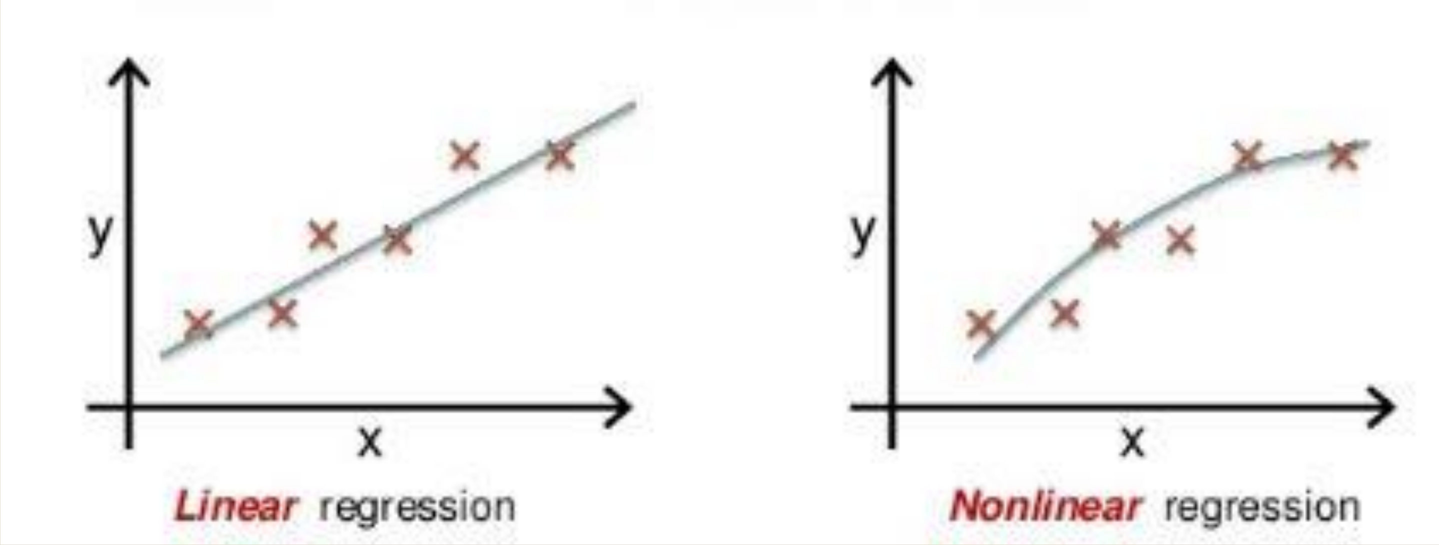
Square Footage (X1)	Bedrooms (X2)	House Price (Y)
1500	3	$250,000
1800	4	$280,000
1200	2	$220,000
2000	4	$320,000
1350	3	$240,000

Multivariate

Regression Types

❑Linear Vs Non-linear

- **Linear:** Relationship is linear between dependent and independent variables
- ***Non-linear:*** *Relationship is nonlinear between dependent and independent variables*

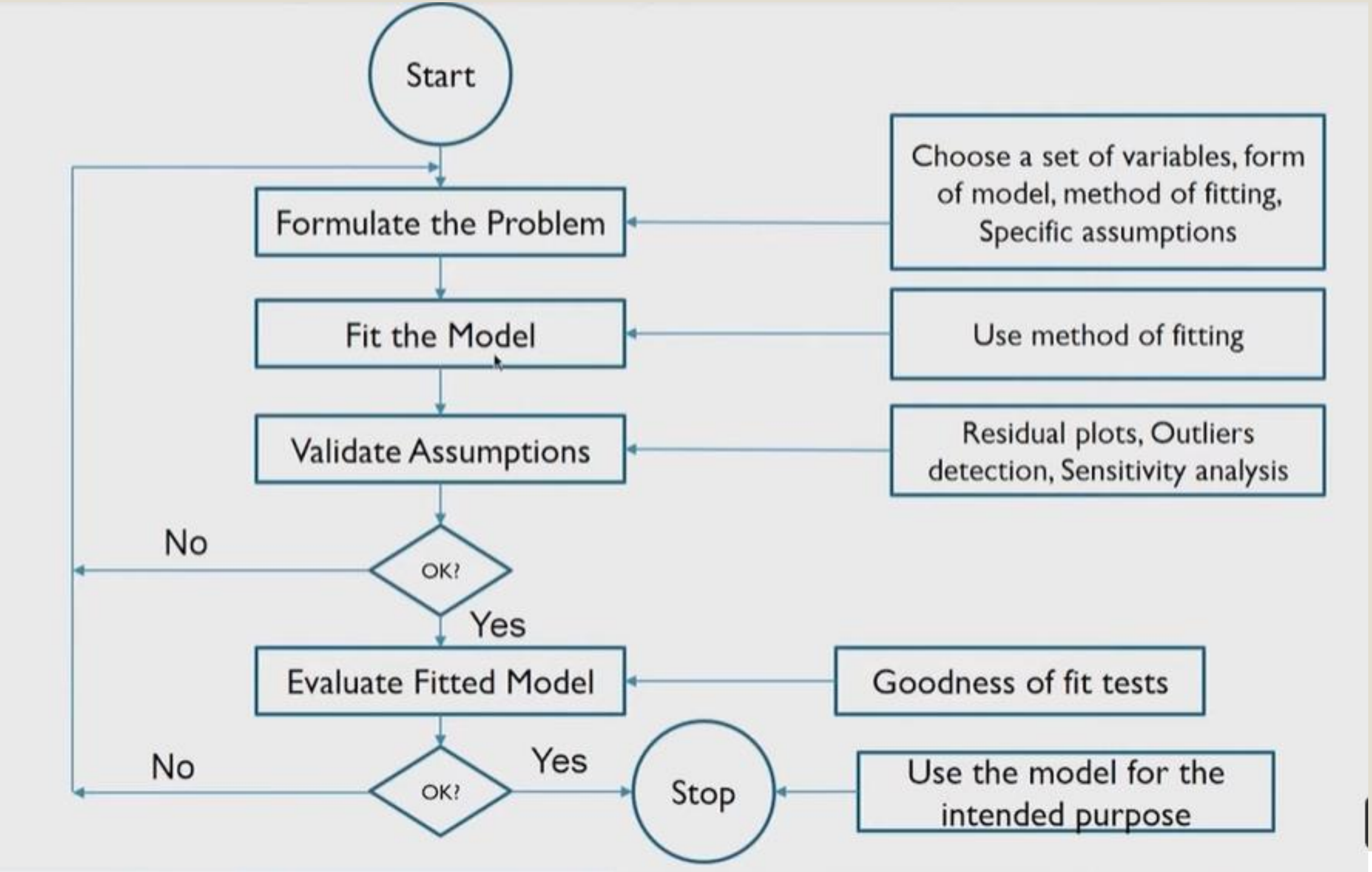


Regression Methods

- **Linear**
  - Ordinary Least Squares (OLS) Regression
  - Ridge Regression (L2 Regularization)
  - Lasso Regression (L1 Regularization)
  - Partial Least Square (PLS) Regression
  - Principle Component Analysis (PCA)

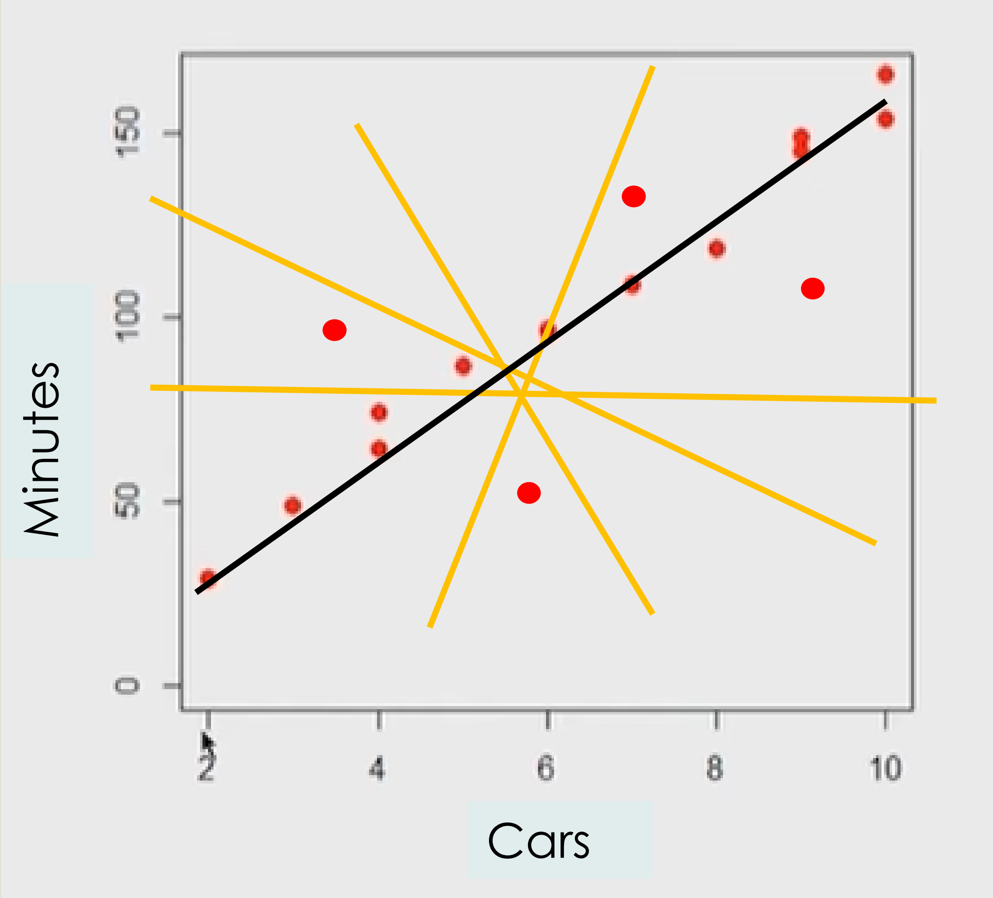
- **Non-linear**
  - Polynomial Regression
  - Neural Network
  - Spline Regression

Regression Process



Regression Illustration

- We have the dataset of car service center
- It contains number of cars (independent variable) and Minute for service (dependent variable)
- *We want to find the best functional relationship between both variables which can be given by linear line*



Ordinary Least square (OLS)

- Linear model between $\boldsymbol{y_i}$ and $\boldsymbol{x_i}$, $\boldsymbol{i = 1, \dots, n}$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Error in only dependent variable and no error in independent variable

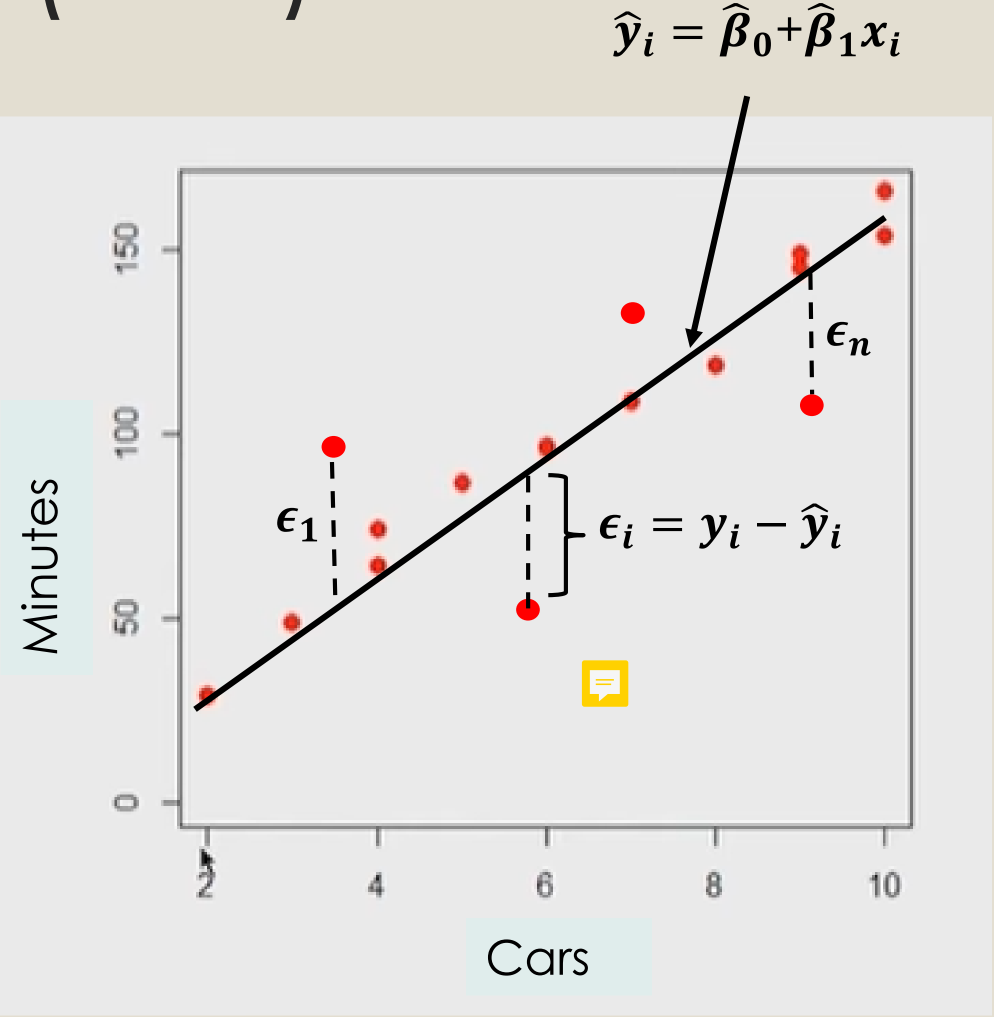
$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

- The sum of square of errors (SSE)

$$\sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

- *The minimization of SSE gives estimate of B0 and B1*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Testing goodness of fit

- R^2 is one of the measure use to test determine goodness of fit
- R^2 calculates the variability in output variable calculated by input variable

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Variability explained by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Total variability in y

- The value of R^2 lie between **1(good fit) and 0 (bad fit)**
- Adjusted R^2 is the modification of R^2 metric to **take into account the number of independent variables**

$$\bar{R}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2 / (n-p-1)}{\sum (y_i - \bar{y})^2 / (n-1)}$$

Q) In a linear regression equation, what does the slope (coefficient) represent?

a) The intercept of the regression line
b) The change in the dependent variable for a unit change in the independent variable
c) The average of the dependent variable
d) The variance of the dependent variable

Solution

The slope coefficient in a linear regression equation indicates how much the dependent variable is expected to change for a unit change in the corresponding independent variable, while holding other variables constant.

Q) What does the coefficient of determination (R-squared) measure in a regression model?

a) The accuracy of the model's predictions
b) The proportion of variance explained by the model
c) The bias of the model
d) The standard error of the model

Solution

R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model.

We have the following data for which we want to calculate the best fit

X	**1**	**2**	**3**	**4**	**5**	**6**
Y	**3**	**5**	**7**	**8**	**10**	**12**

Step 1) Calculate the Means:
Mean of X: (1+2+3+4+5+6)/6 = *3.5*
Mean of Y: (3+5+7+8+10+12)/6 = *7.5*

Step 2) Calculate the Deviations from mean:
$(x - \bar{x})$: *[(1-3.5),(2-3.5),(3-3.5),(4-3.5),(5-3.5),(6-3.5)] = [-2.5, -1.5, -0.5, 0.5, 1.5, 2.5]*
$(\mathrm{y} - \bar{y})$: *: [(3-7.5),(5-7.5),(7-7.5),(8-7.5),(10-7.5),(12-7.5)] = [-4.5, -2.5, -0.5, 0.5, 2.5, 4.5]*

Step 3) Calculate the covariance between x and y and variance of x:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\mathrm{y_i} - \bar{y})}{n} = \frac{(-2.5 * -4.5) + (-1.5 * -2.5) + (-0.5 * -0.5) + (0.5 * 0.5) + (1.5 * 2.5) + (2.5 * 4.5)}{6} = 5.083$$

$$S_{xx} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{(-2.5)^2 + (-1.5)^2 + (-0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2}{6} = 2.916$$

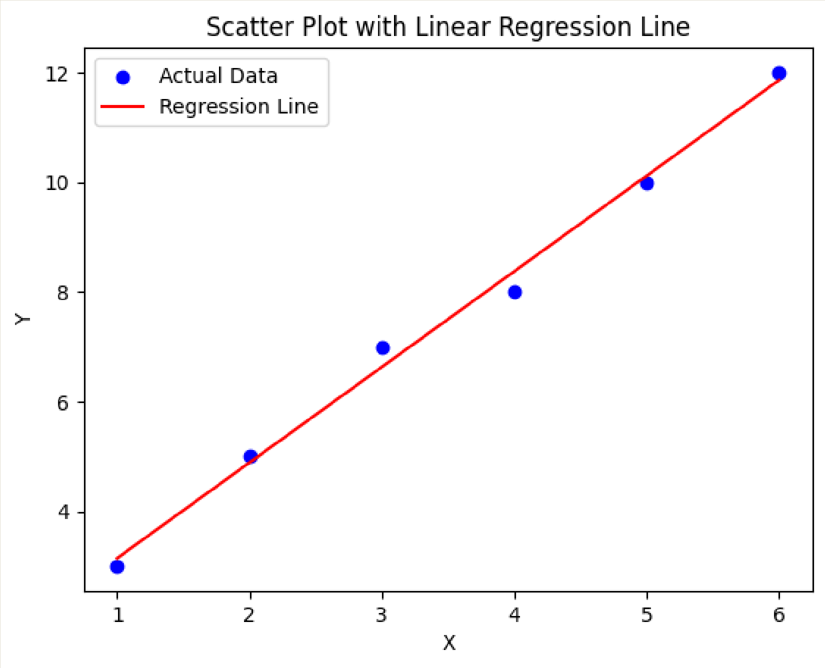
Step 4) Use formulae to calculate coefficient

$$\hat{\boldsymbol{\beta}}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5.083}{2.916} = 1.743$$

Step 5) Use formulae to calculate $\hat{\boldsymbol{\beta}}_0$

$$\hat{\boldsymbol{\beta}}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.5 - 1.743 * 3.5 = 1.4$$

So the final equation is $\boldsymbol{\hat{y} = 1.4 + 1.743x}$

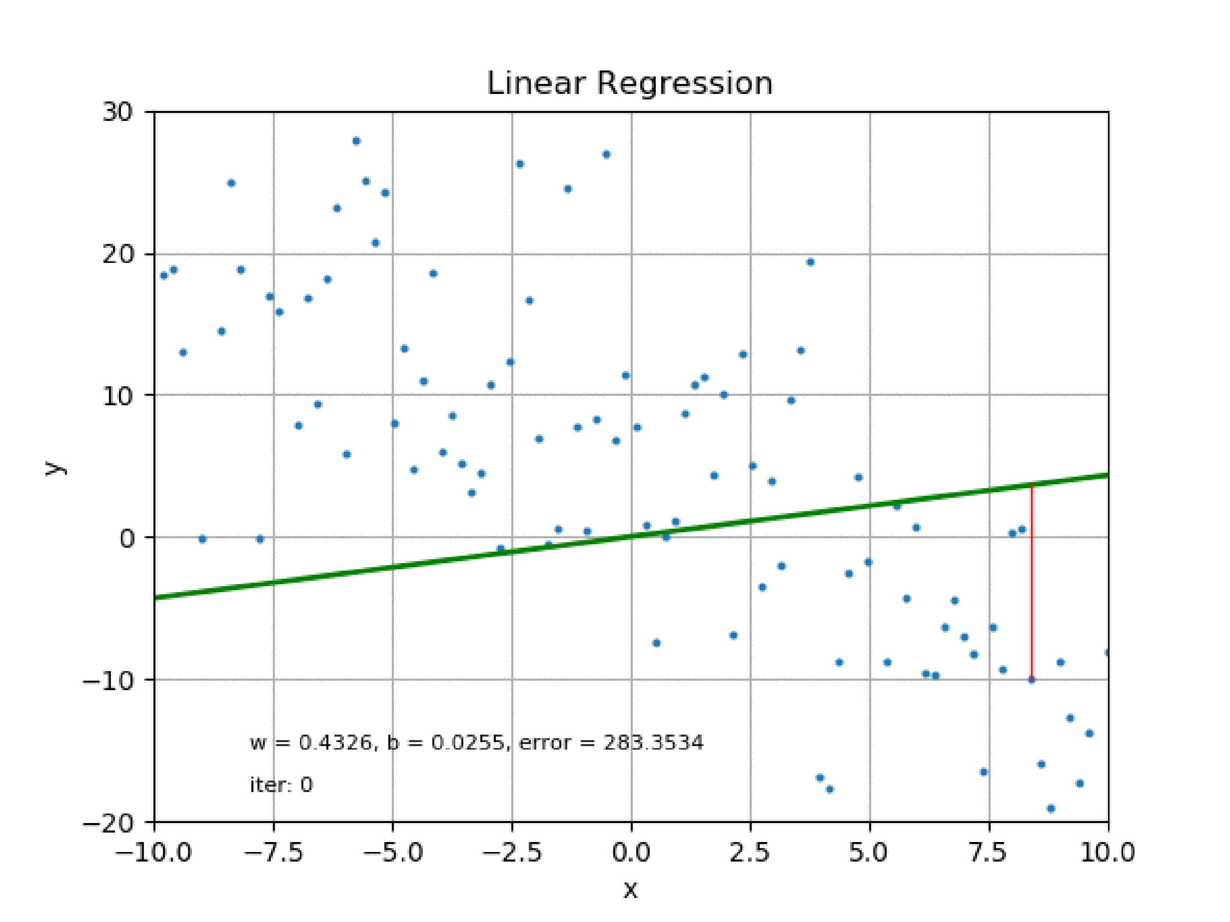


Step 6) Calculate the R^2 value to evaluate model

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \mathbf{0.9936;}$$

$$\textit{Similarly calculating Adjusted } R^2 = \mathbf{0.992}$$

R studio



Linear Regression
30
20
10
0
−10
−20
y
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
x
w = 0.4326, b = 0.0255, error = 283.3534
iter: 0

We have the following data for which we want to calculate the best fit

X	**1**	**2**	**3**	**4**	**5**
Y	**3**	**8**	**7**	**5**	**11**

Q1) What is the slope (coefficient) of the best-fitting linear regression line for this dataset?

a) 2.1
b) 1.3
c) 1.7
d) 2.3

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Q2) What is the intercept of the best-fitting linear regression line for this dataset?

a) 2.8
b) 1.8
c) 1.9
d) 2.9

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Q3) What is the predicted value of Y when X = 6 using the linear regression model?

a) 10.1
b) 8.6
c) 11.3
d) 10.7

Solution

Y = 1.3x+2.9 = 1.3(6)+2.9 = 10.7

Q4) What is the (R-squared) for the linear regression model fitted to this dataset?

a) 0.55
b) 0.45
c) 0.35
d) 0.60

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Variability explained by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Total variability in y

Q5) What is the mean squared error (MSE) for the linear regression model fitted to this dataset?

a)3.49
b) 2.93
c) 3.98
d) 4.21

Solution
MSE=(residuals)^2/n=3.98

Q) If the slope of the linear regression line is 3 and the intercept is 2, what would be the predicted Y value when X = 8?

a) 24
b) 26
c) 28
d) 30

Solution

Y = 3x+2 = 3(8)+2 = 26

Sum Square Quantity Definitions

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

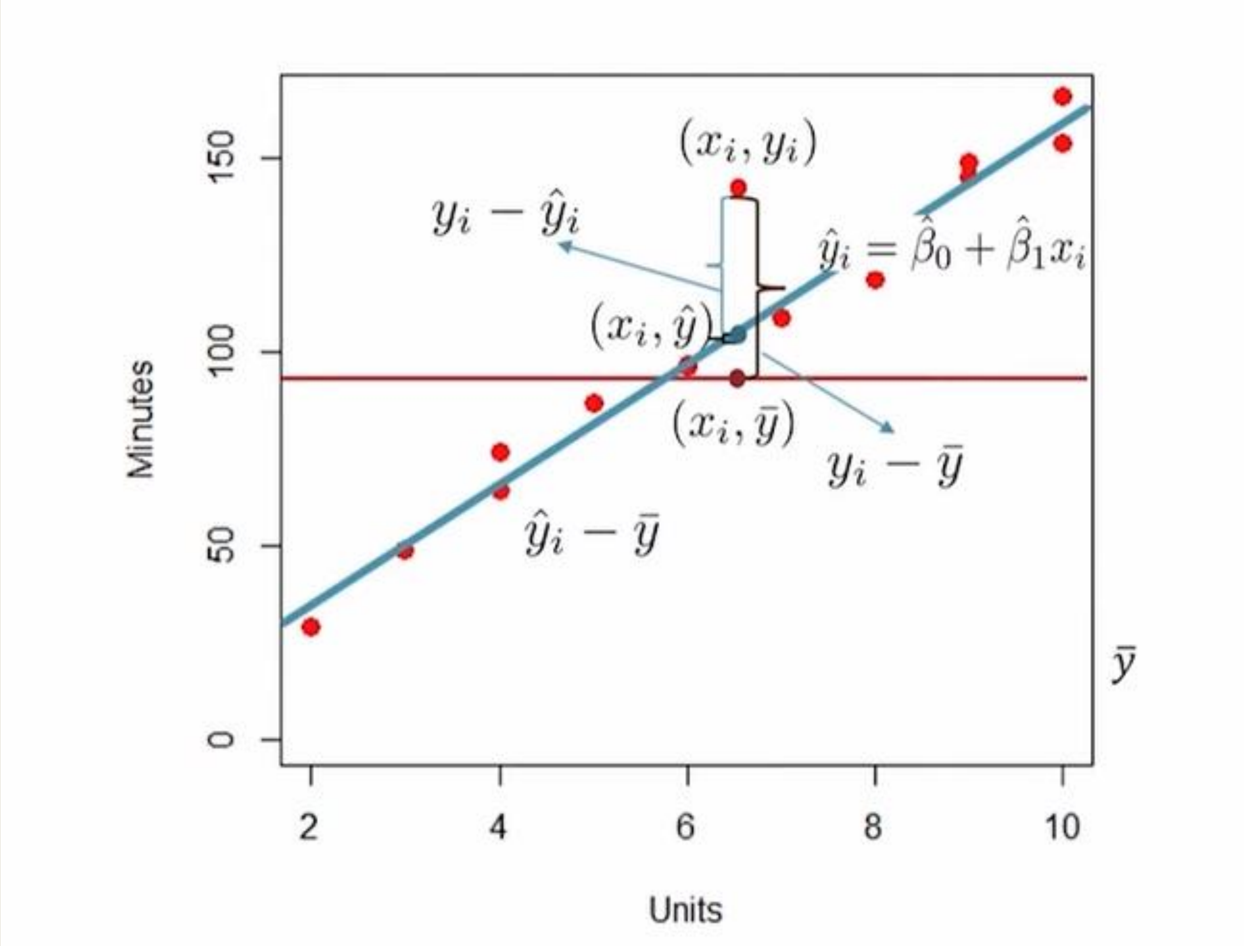
$$\mathrm{SSR} = \sum (\hat{y}_i - \bar{y})^2$$

$$\mathrm{SST} = \sum (y_i - \bar{y})^2$$

- SSR (residual sum-of-squares)
- SST (total sum-of-squares)
- SSE (sum-squared error)

- **SST = SSE+SSR**

- **R^2 = 1-SSE / SST**



Which of the following formulas correctly calculates SSE (Sum of Squares Error)?

A) $SSE = \Sigma(y_i - \bar{y})^2$
B) $SSE = \Sigma(y_i - \hat{y}_i)^2$
C) $SSE = \Sigma(\hat{y}_i - \bar{y})^2$

Q) Which equation relates SST, SSE, and SSR?

A) SST = SSE + SSR
B) SST = SSE - SSR
C) SST = SSE * SSR
D) SSE = SST - SSR

Solution

The total variability (SST) is the sum of the explained variability (SSR) and the unexplained variability (SSE).

What is the possible range of values for R-squared (R^2)?

A) -∞ to +∞
B) 0 to 1
C) -1 to 1
D) 0 to ∞

Solution

R^2 values range from 0 to 1, where 0 indicates no variability explained by the model, and 1 indicates perfect fit.

Q) If SSE = 200 and SST = 800, what is the value of R^2?

A) 0.25
B) 0.5
C) 0.75
D) 0.8

Solution
R^2=1-SSE/SST=1-200/800=0.75

If the linear regression model perfectly fits the data, what would be the value of SSE?

A) 0
B) Equal to SST
C) Equal to SSR
D) Indeterminate

Solution

If the model perfectly fits the data, all predicted values ($\hat{y}_i$) will be equal to the actual values (y_i), resulting in no residual differences and SSE being equal to 0.

Q) What happens to R^2 when the regression model's fit improves?

A) R^2 decreases
B) R^2 increases
C) R^2 remains unchanged
D) R^2 becomes negative

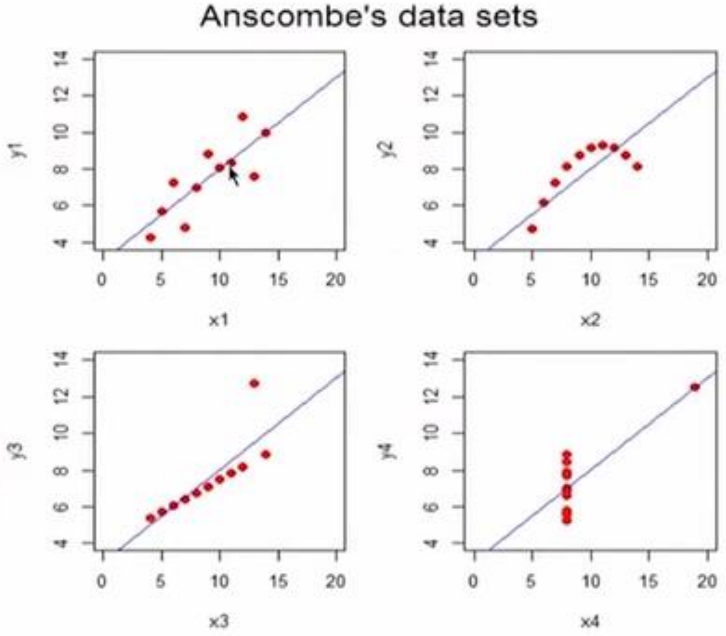
R studio Example

Hypothesis test on regression coefficient

- In order to check if linear model fit is good or not we can test whether estimate $\hat{\beta}_1$ is significant (different from zero) or not
- Null hypothesis $H_0 : \beta_1 = 0$
- Alternative hypothesis $H_1 : \beta_1 \neq 0$
- Null hypothesis implies $\hat{y}_i = \hat{\beta}_0 + \epsilon_i$ ⟵ Reduced Model
- Alternative hypothesis implies $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i$ ⟵ Full Model

Next step to check the linear fit

- Linear regression of Anscombe data sets




```
$lm1
              Estimate Std. Error
(Intercept) 3.0000909  1.1247468
x1          0.5000909  0.1179055
```

```
$lm2
              Estimate Std. Error
(Intercept) 3.000909  1.1253024
x2          0.500000  0.1179637
```

```
$lm3
              Estimate Std. Error
(Intercept) 3.0024545  1.1244812
x3          0.4997273  0.1178777
```

```
$lm4
              Estimate Std. Error
(Intercept) 3.0017273  1.1239211
x4          0.4999091  0.1178189
```

- R^2, CI for regression coefficients, hypotheses tests all give identical results for all four data sets!

Residual plots

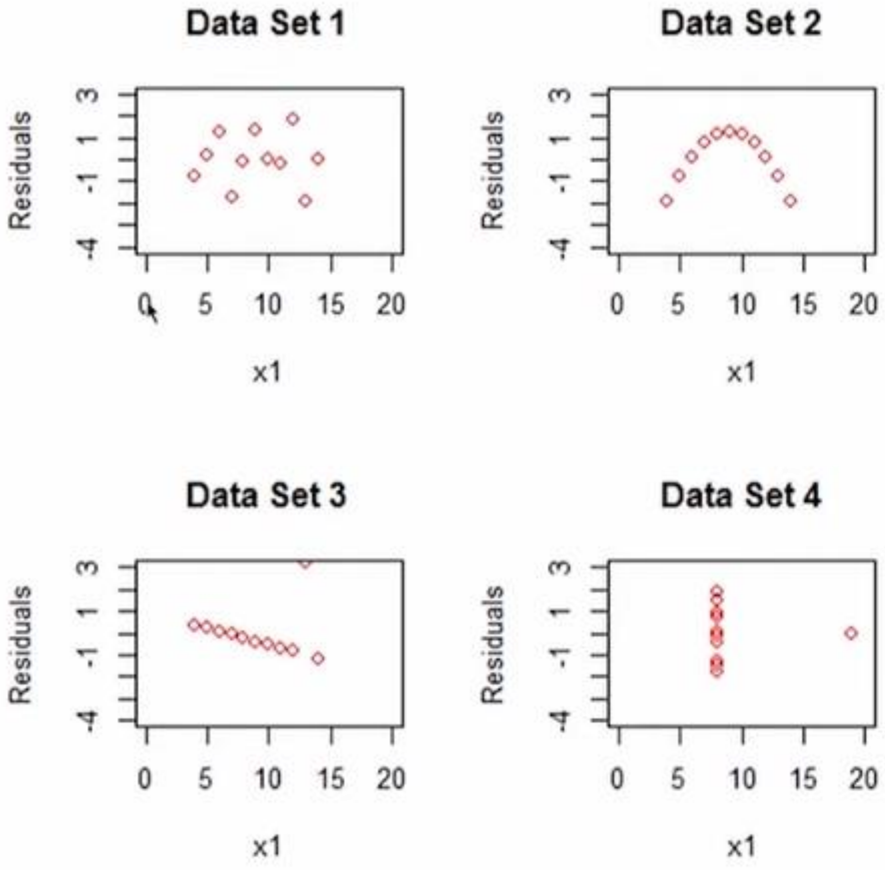
- Residual plots for Anscombe data

No pattern

Pattern

Pattern

Pattern



- Look for patterns
  - Random
    A valid model
  - Pattern
    Not a valid model
- Shape of Pattern
  Information on the function of x

Thank you