Introduction to Data Science With Probability and Statistics

Lecture 22: Introduction to Regression

CSCI 3022 - Summer 2020
Sourav Chakraborty
Dept. of Computer Science
University of Colorado Boulder

What will we learn today?

- ❑ Predictive statistics
- ❑ Simple linear regression (SLR) model
- ❑ Residuals
- ❑ Sum of squared-errors
- ❑ Fitted regression line
- ❑ Least-squares line

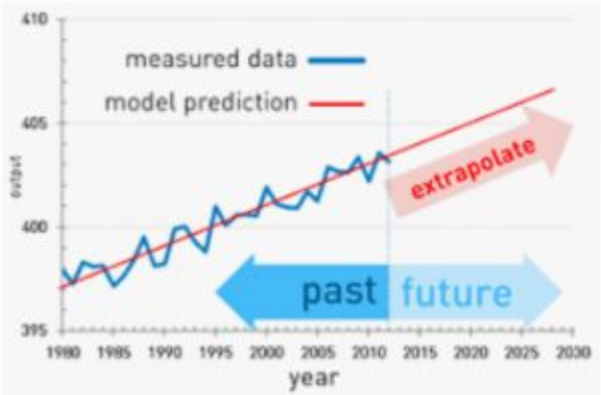
- ❑ *A Modern Introduction to Probability and Statistics, Chapter 22*

Statistical Modeling

Few types of statistics:

- Descriptive Statistics: Summarizing the dataset/sample. “This is the way my sample is”. (done)

- Inferential Statistics: Drawing conclusions from the sample. “This is what I can conclude from my sample with 100(1 - α)% confidence”. (next week)



➢ Today: predictive statistics

Linear regression for prediction

Examples:

- Given a person's age and gender, predict their height
- Given the area of a house, predict its sale price
- Given unemployment, inflation, number of wars and economic growth, predict the president's approval rating
- Given a person's browser history, predict how long they'll stay on a product page
- Given the advertising budget expenditures in various media markets, predict the number of products they'll sell.

Simple Linear Regression

Our goal: figure out the equation of the line through the data.

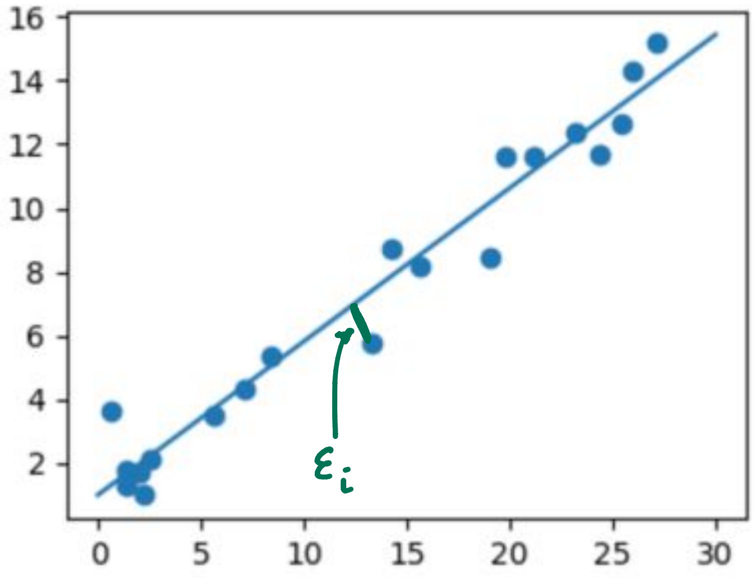
Definitions and Assumptions:

1) $y_i = \alpha + \beta\, x_i + \epsilon_i$

2) Each of the ϵ_i are independent

3) $\epsilon_i \sim N(0, \sigma^2)$

True model.



Simple Linear Regression

$$Y = \alpha + \beta X + \epsilon$$

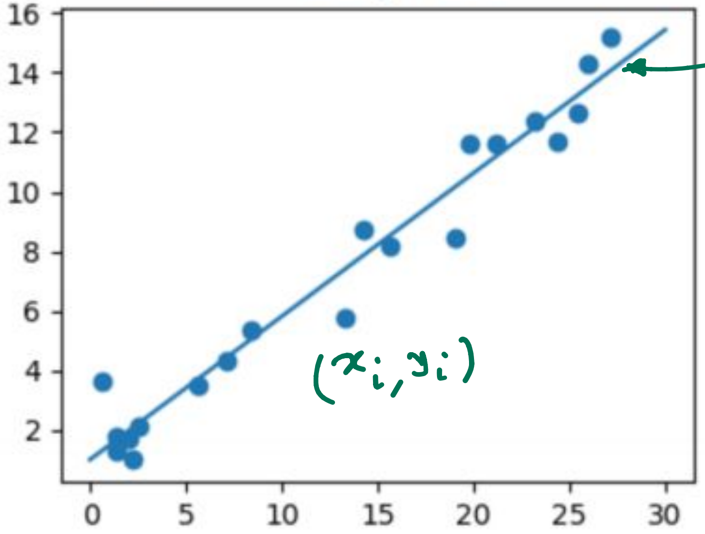
X: the independent variable, the predictor, the explanatory variable, the feature

Y: the dependent variable, the response variable

$\boldsymbol{\epsilon}$: the random deviation, random error – accounts for the fact that the world is uncertain and that there are random deviations around the true process.

Simple Linear Regression

$$Y = \alpha + \beta X + \epsilon$$

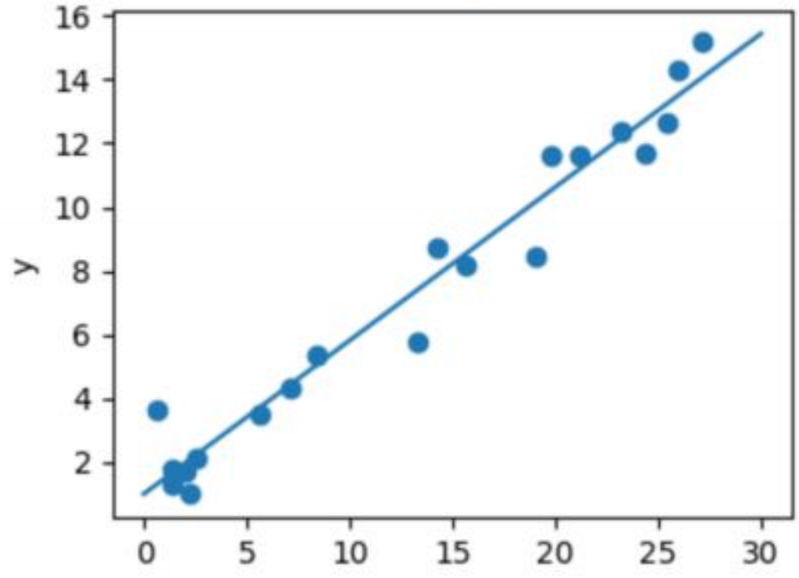


best possible fit. (True line)

The points (x_1, y_1), (x_2, y_2), … , (x_n, y_n) resulting from n independent observations will be scattered about the true regression line.

Simple Linear Regression

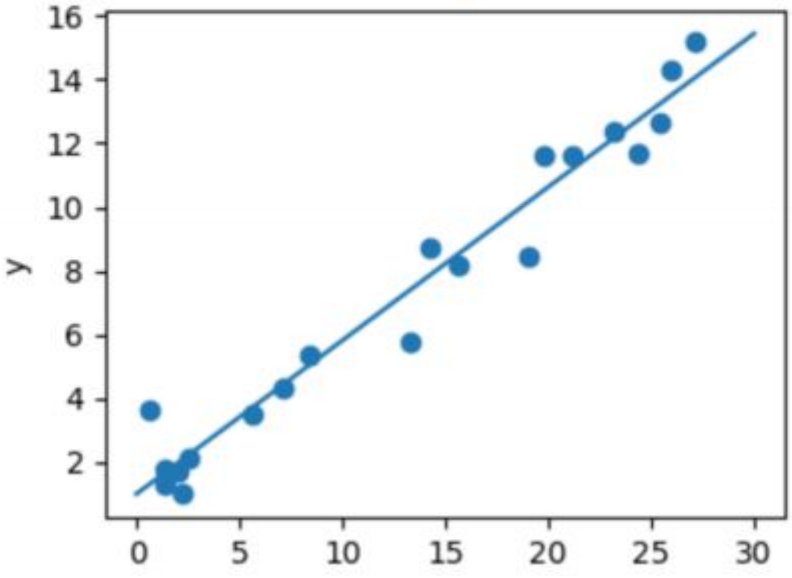
$$Y = \alpha + \beta X + \epsilon$$



Y is a random variable.
What is $E[Y]$?

Simple Linear Regression

$$Y = \alpha + \beta X + \epsilon$$



Y is a random variable.
What is $E[Y]$?

$\varepsilon \sim N(0, \sigma^2)$

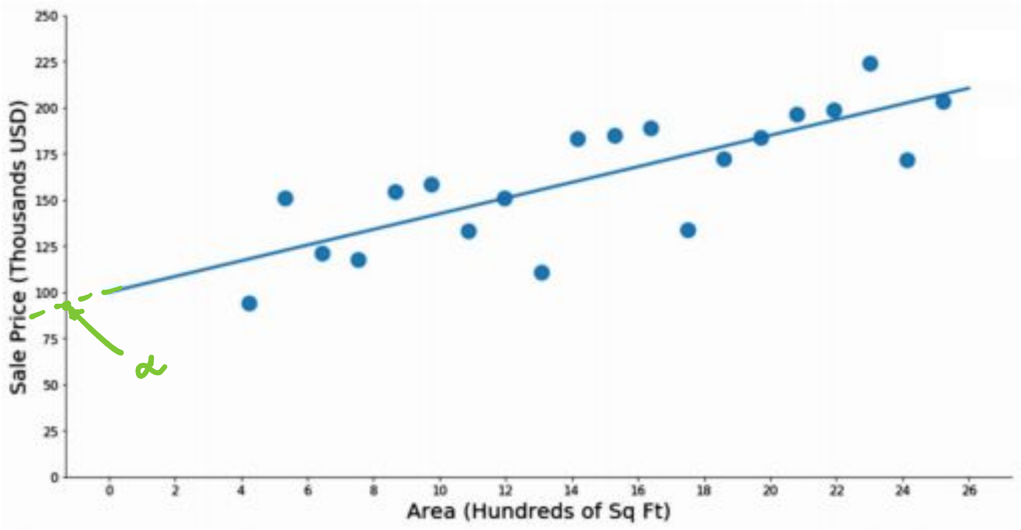
$$E[y] = E[\alpha + \beta X + \varepsilon]$$
$$= E[\alpha] + E[\beta X] + \underbrace{E[\varepsilon]}_{0}$$
$$= \alpha + \beta X$$

X: predictor variable (or the feature)

Simple Linear Regression – Interpreting parameters

α is the intercept of the true regression line (aka the baseline average)

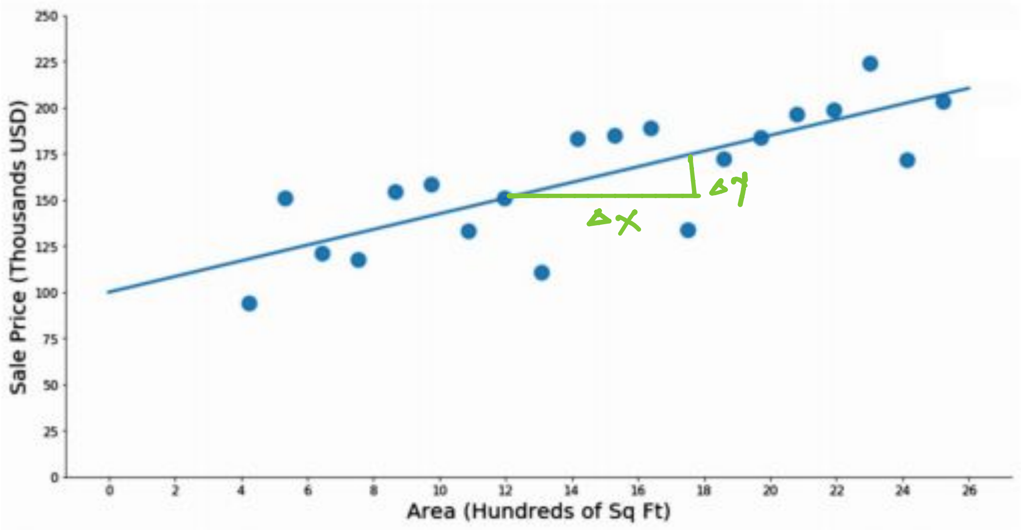
It's called the "baseline" because it's the response when $\underbrace{\text{the feature} = 0}_{x=0}$



Simple Linear Regression – Interpreting parameters

β is the slope of the true regression line

It's the increase in our response from a unit increase in our feature

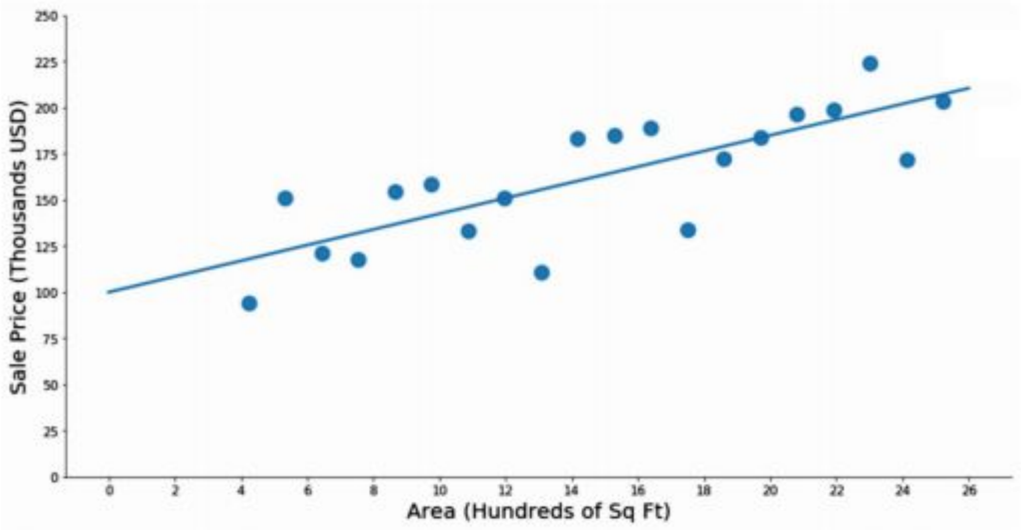


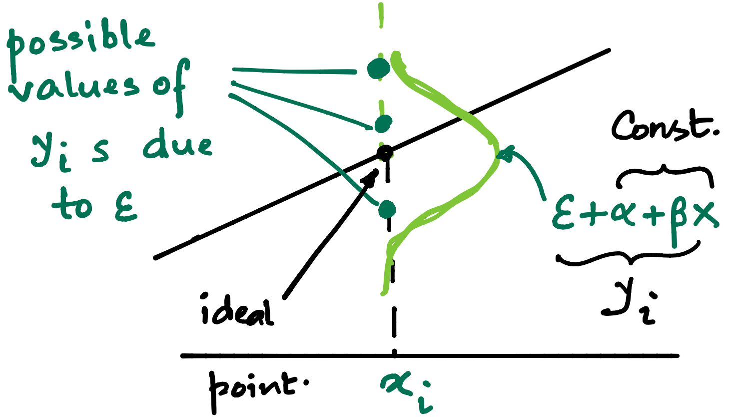
β : Truly determines the relationship between y and x.

if $\beta = 0$ the no relationship bet^n x & y.

Simple Linear Regression – Interpreting parameters

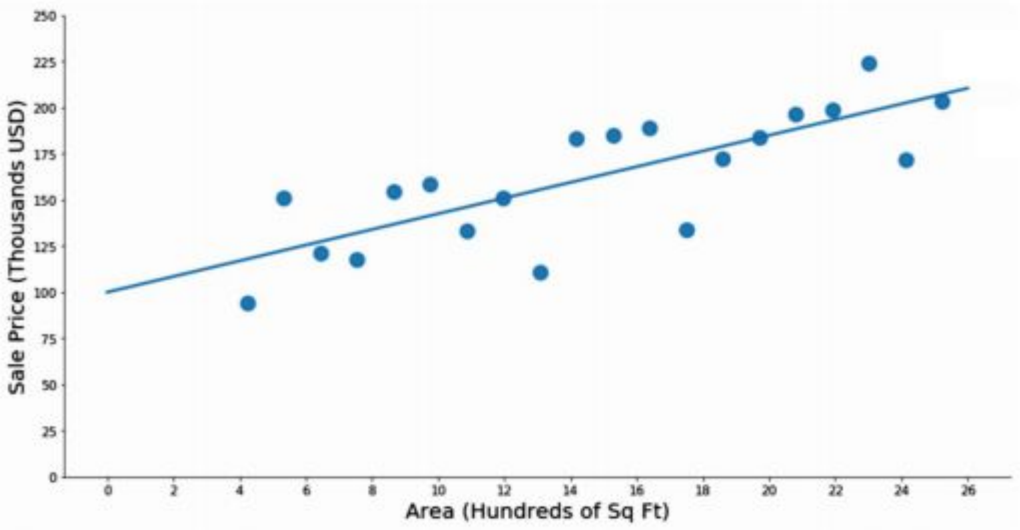
The variance parameter σ^2 determines the extent to which each normal curve spreads about the true regression line.

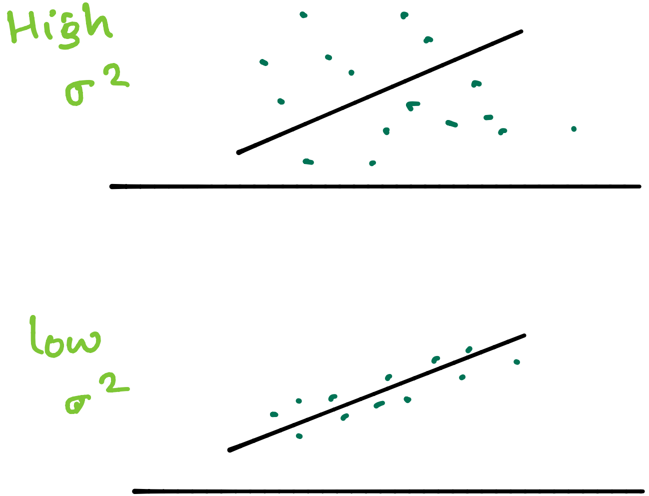




Simple Linear Regression – Interpreting parameters

The variance parameter σ^2 determines the extent to which each normal curve spreads about the true regression line.





Simple Linear Regression – Estimating model parameters

We've got data: $(x_1, y_1), (x_2, y_2), \dots .(x_n, y_n)$

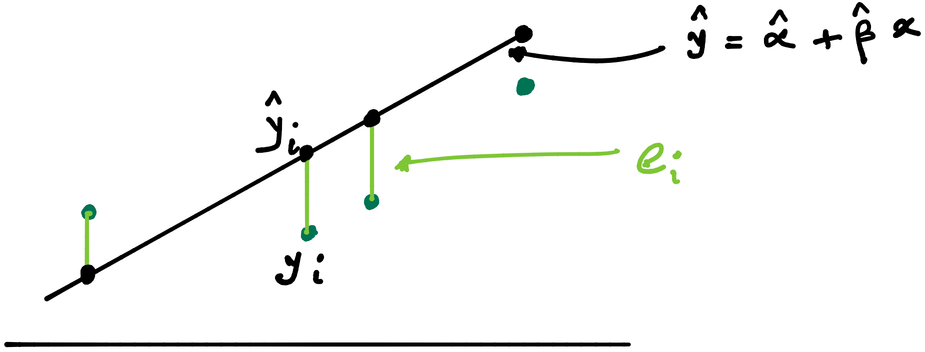
True line

We want to figure out our regression line: $y = \alpha + \beta x$

$\hat{y}_i$

How can we minimize the residuals, $e_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$?

So; we have to estimate $y = \alpha + \beta x$ using $\hat{y} = \hat{\alpha} + \hat{\beta} x$



Simple Linear Regression – Estimating model parameters

The sum of squared-errors for the points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ to the regression line is given by

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The point-estimates (single value estimates from the data) of the slope and intercept parameters are called the least-squares estimates, and are defined to be the values that minimize the SSE.

Goal : $\min(SSE)$

Simple Linear Regression – Estimating model parameters

How do we actually find the parameter estimates?

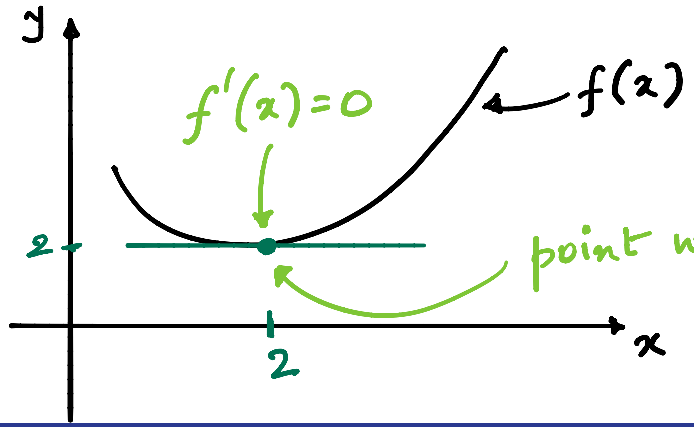
$$\boxed{\frac{\partial SSE}{\partial \hat{\alpha}} = 0 \quad , \quad \frac{\partial SSE}{\partial \hat{\beta}} = 0}$$

Example left: $f(x) = (x-2)^2 + 2$

$$f'(x) = \frac{df}{dx} = 2(x-2) = 0$$

$$\therefore x = 2 \,//$$

$$\boxed{f(2) = 2}$$



point where slope or $\frac{df}{dx} = 0$

OR $\underline{\min f(x)}$

Simple Linear Regression – Estimating model parameters

How do we actually find the parameter estimates?

ONE WAY

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

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

ANOTHER WAY

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{Cov(x,y)}{Var(x,y)}$$

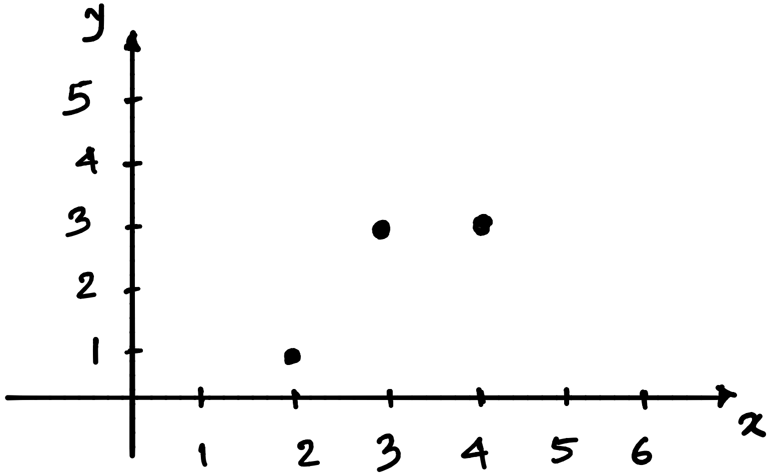
where : $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}$

$$\overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n}$$

$\hat{\alpha}$: same in all

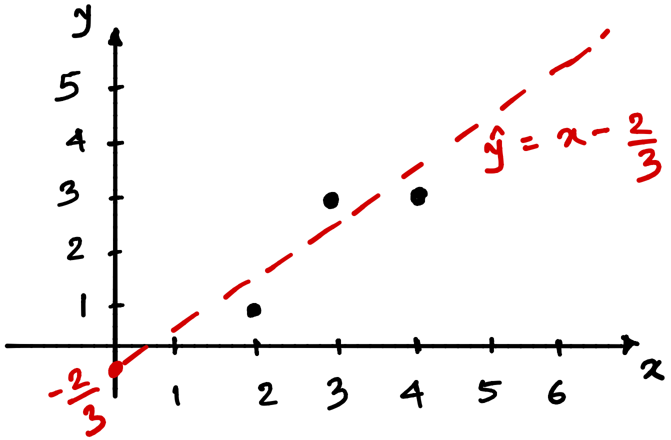
Simple Linear Regression – Estimating model parameters

Example: Find the regression line for the following data: $(2,1), (3,3), (4,3)$



Simple Linear Regression – Estimating model parameters

Example: Find the regression line for the following data: (2,1), (3,3), (4,3)



$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$\therefore \hat{\beta} = \frac{\frac{23}{3} - 3 \cdot \frac{7}{3}}{\frac{29}{3} - (3)^2} = 1.$$

$$\hat{\alpha} = \frac{7}{3} - 1 \cdot 3 = \frac{-2}{3}$$

$$\bar{x} = (2+3+4)/3 = 3$$

$$\bar{y} = (1+3+3)/3 = 7/3$$

$$\overline{xy} = \frac{2 \cdot 1 + 3 \cdot 3 + 4 \cdot 3}{3} = \frac{23}{3}$$

$$\overline{x^2} = \frac{2^2 + 3^2 + 4^2}{3} = \frac{29}{3}$$

$$\therefore \hat{y} = -\frac{2}{3} + x$$

Residuals

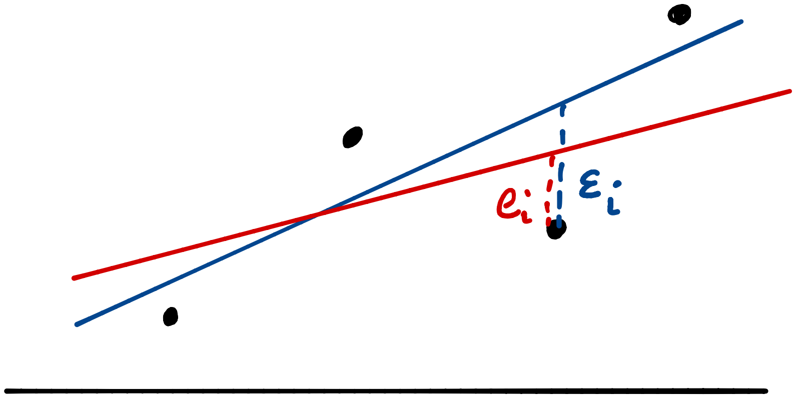
The residuals are the difference between the observed and the predicted responses:

$$r_i = y_i - \hat{y}_i$$

OR

e_i

The residuals r_i are estimates of the unknown true error ϵ_i



$SSE \equiv RSS$