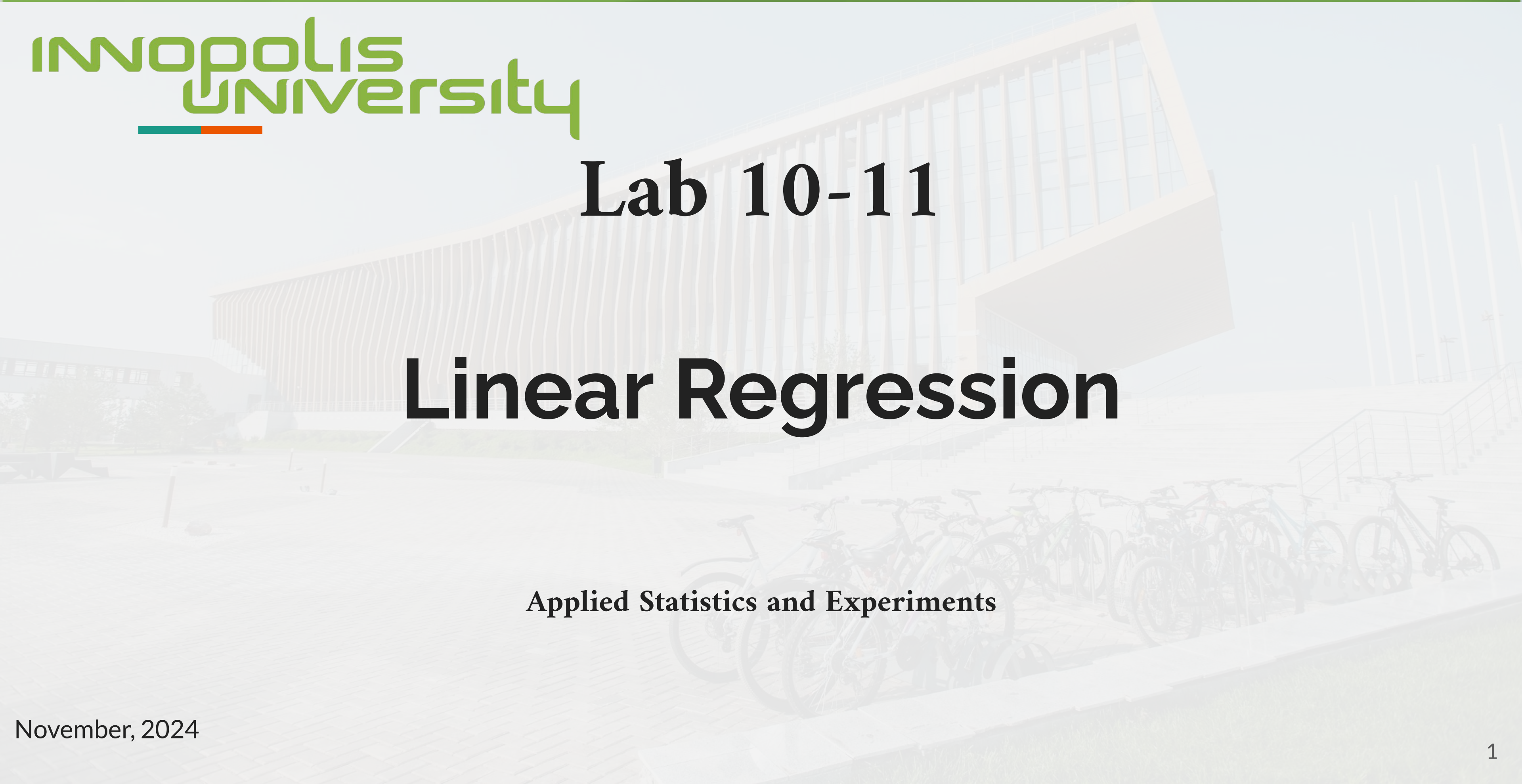


Lab 10-11

Linear Regression

Applied Statistics and Experiments

November, 2024

Agenda

1. Fitting curves to bivariate data
2. Linear regression
3. Simple linear regression
4. What is linear about linear regression?
5. Big assumptions in linear regression
6. Formulas of simple linear regression
7. Measuring goodness of fit
8. Coefficient of determination

Lecture Recap

https://quizizz.com/join

Join and enter game code

Regression model

- Answers “What is the relationship between the variables?”
- Equation used
  - One numerical dependent (response) variable (y)
    - What is to be predicted
  - One or more numerical independent (explanatory) variables (X)
- Used mainly for prediction and forecasting

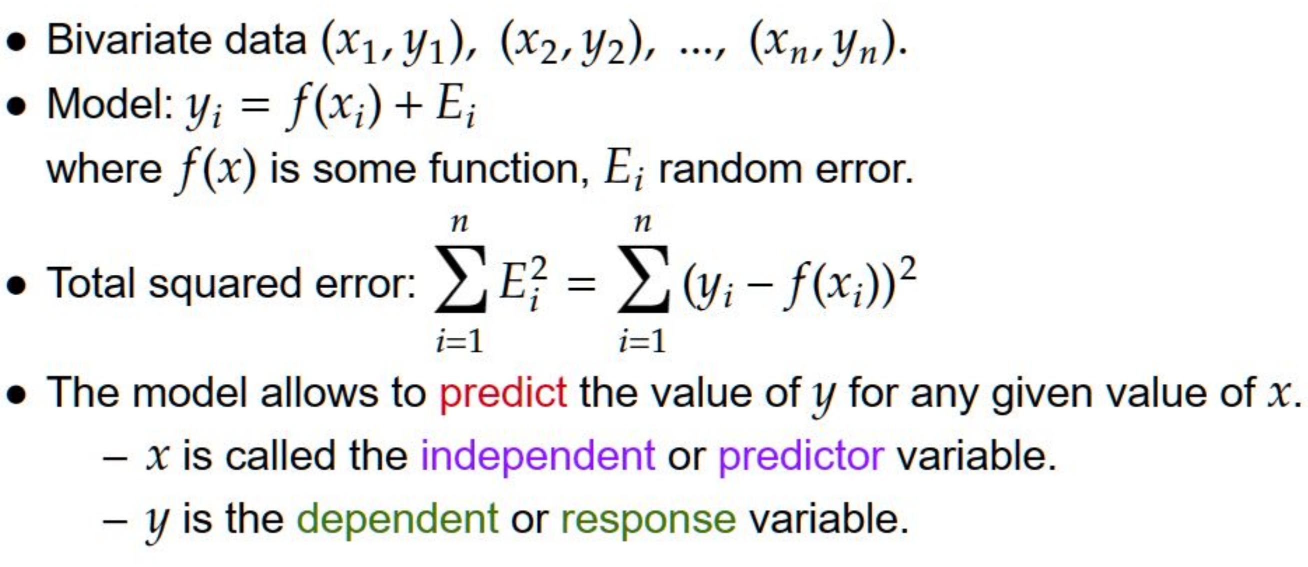
Regression model

- You are a marketing analyst in an e-commerce company. You gather the following data:
- The independent variable x
  - Ad expenditure (predictor)
- The dependent variable y
  - Sales revenue (response)

Ad expenditure ($100)	Sales revenue ($1000)
1	1
2	1
3	2
4	2
5	4

Regression model

Source



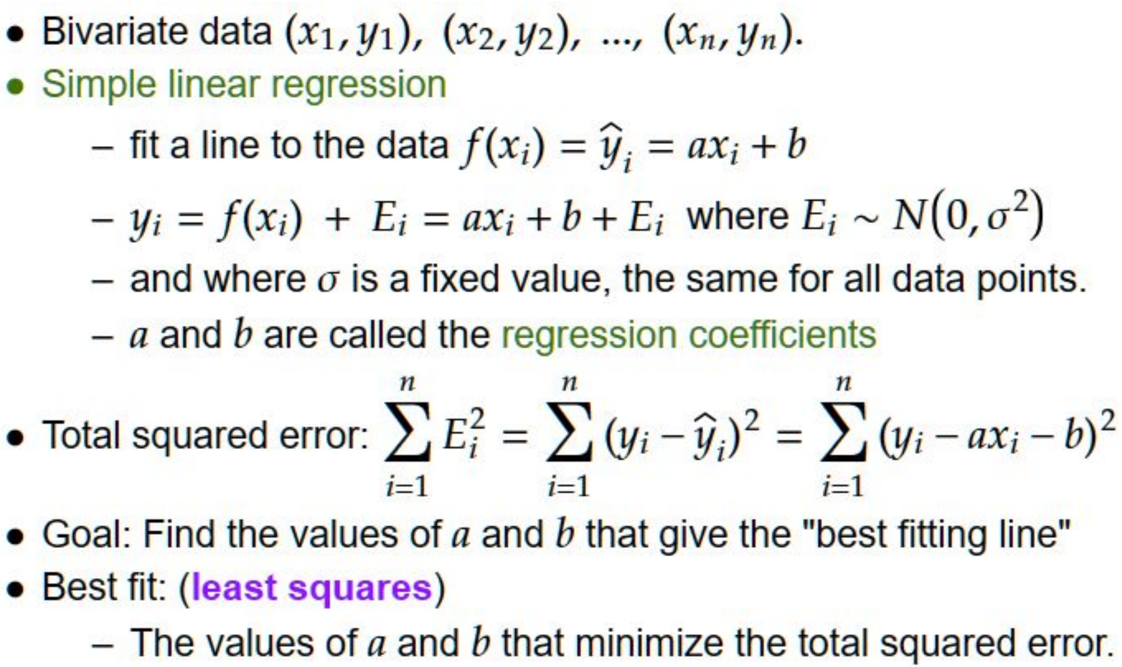
- Bivariate data $(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$.
- Model: $y_i = f(x_i) + E_i$
  where $f(x)$ is some function, E_i random error.
- Total squared error: $\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$
- The model allows to predict the value of y for any given value of x.
  - x is called the independent or predictor variable.
  - y is the dependent or response variable.

Examples of f(x)

Mean function

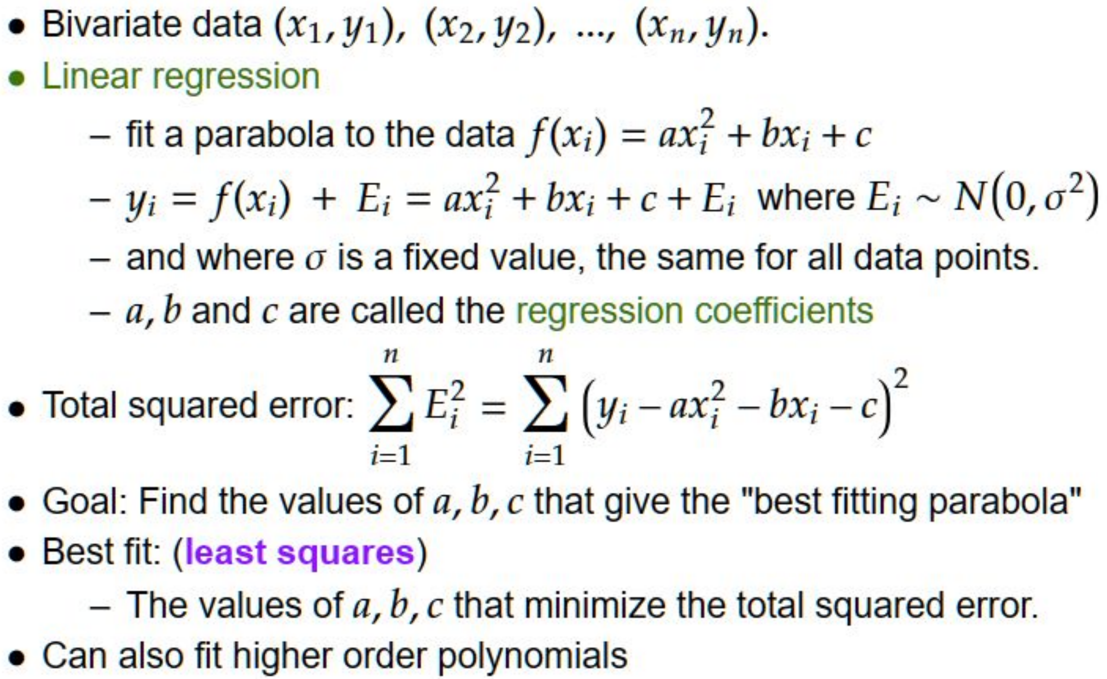
- lines: $y = ax + b + E$
- polynomials: $y = ax^2 + bx + c + E$
- other: $y = \frac{a}{x} + b + E$
- other: $y = a\,sin(x)\ + b\ + E$

Simple linear regression: finding the best fitting line



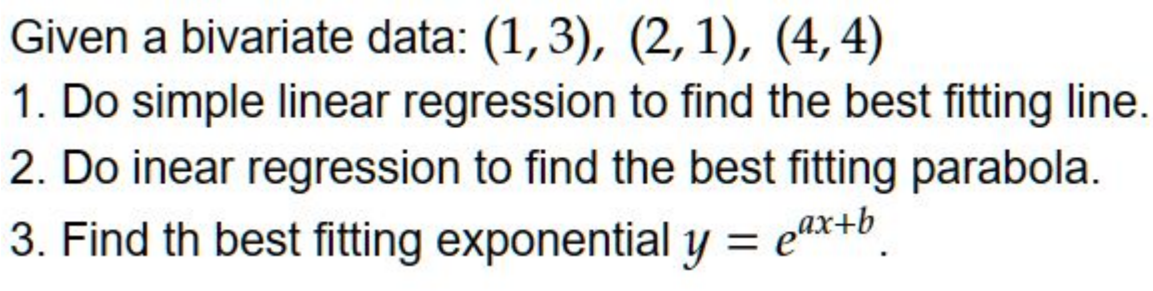
- Bivariate data $(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$.
- Simple linear regression
  - fit a line to the data $f(x_i) = \hat{y}_i = ax_i + b$
  - $y_i = f(x_i) + E_i = ax_i + b + E_i$ where $E_i \sim N(0, \sigma^2)$
  - and where σ is a fixed value, the same for all data points.
  - a and b are called the regression coefficients
- Total squared error: $\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2$
- Goal: Find the values of a and b that give the "best fitting line"
- Best fit: (**least squares**)
  - The values of a and b that minimize the total squared error.

Linear regression: finding the best fitting polynomial



- Bivariate data $(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$.
- Linear regression
  - fit a parabola to the data $f(x_i) = ax_i^2 + bx_i + c$
  - $y_i = f(x_i)\ +\ E_i = ax_i^2 + bx_i + c + E_i$ where $E_i \sim N\left(0, \sigma^2\right)$
  - and where σ is a fixed value, the same for all data points.
  - a, b and c are called the regression coefficients
- Total squared error: $\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \left(y_i - ax_i^2 - bx_i - c\right)^2$
- Goal: Find the values of a, b, c that give the "best fitting parabola"
- Best fit: (**least squares**)
  - The values of a, b, c that minimize the total squared error.
- Can also fit higher order polynomials

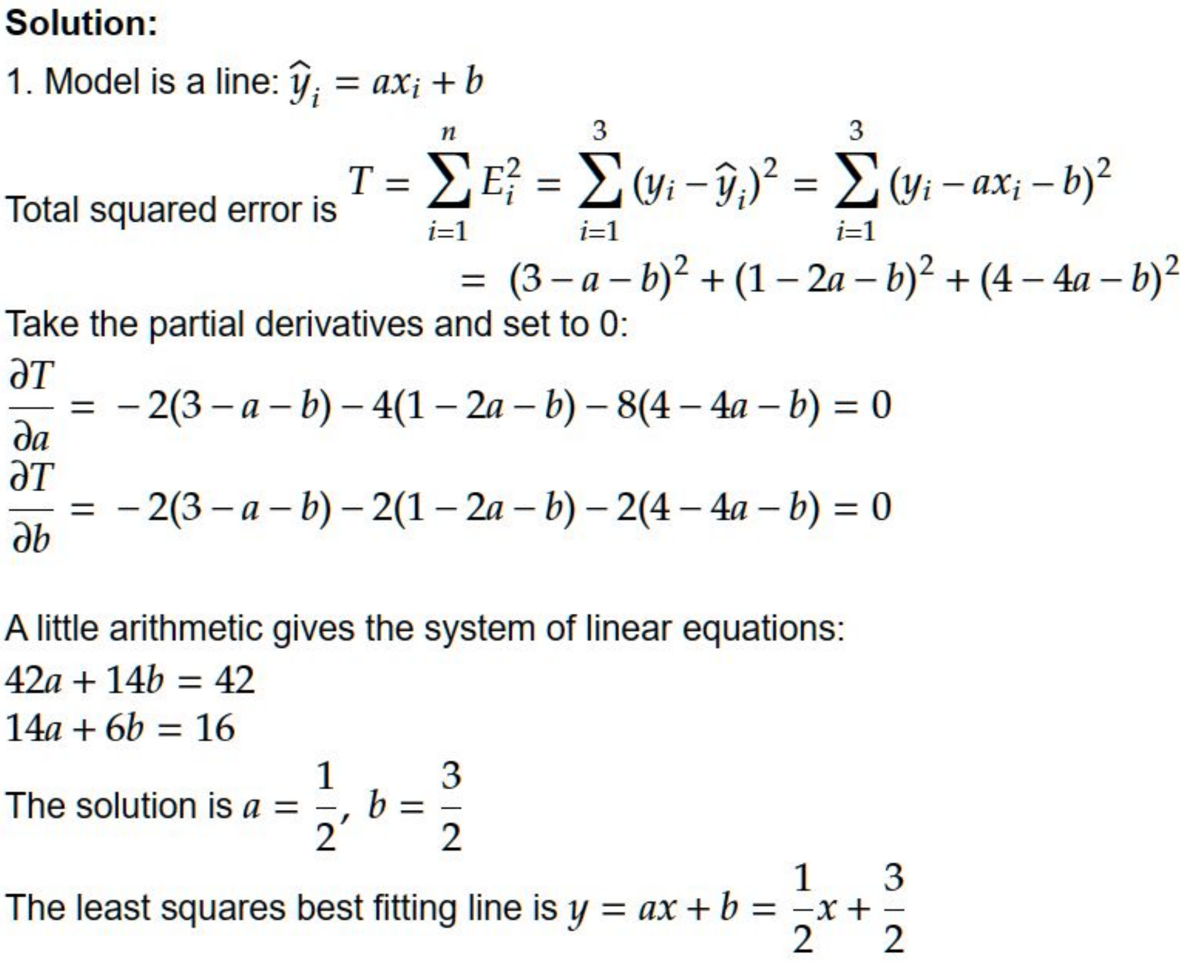
Example



Given a bivariate data: $(1,3),\ (2,1),\ (4,4)$

1. Do simple linear regression to find the best fitting line.
2. Do inear regression to find the best fitting parabola.
3. Find th best fitting exponential $y = e^{ax+b}$.

Example



Solution:

1. Model is a line: $\widehat{y}_i = ax_i + b$

Total squared error is $T = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{3} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{3} (y_i - ax_i - b)^2$

$$= (3 - a - b)^2 + (1 - 2a - b)^2 + (4 - 4a - b)^2$$

Take the partial derivatives and set to 0:

$$\frac{\partial T}{\partial a} = -2(3 - a - b) - 4(1 - 2a - b) - 8(4 - 4a - b) = 0$$

$$\frac{\partial T}{\partial b} = -2(3 - a - b) - 2(1 - 2a - b) - 2(4 - 4a - b) = 0$$

A little arithmetic gives the system of linear equations:

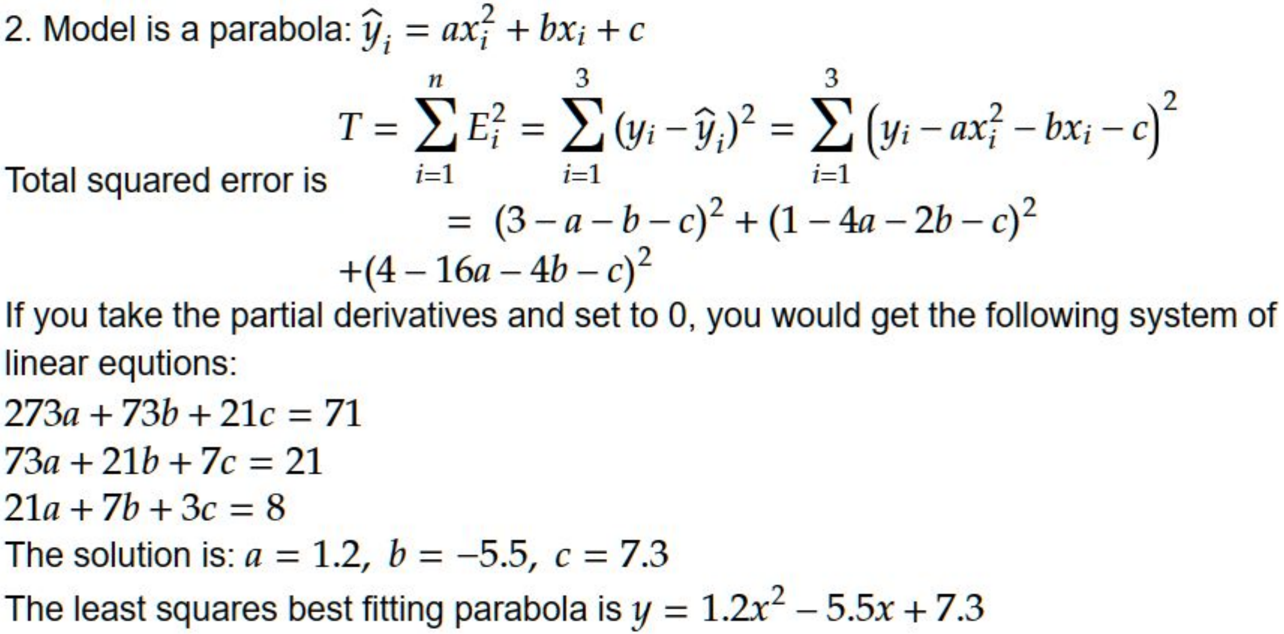
$$42a + 14b = 42$$

$$14a + 6b = 16$$

The solution is $a = \frac{1}{2}, \ b = \frac{3}{2}$

The least squares best fitting line is $y = ax + b = \frac{1}{2}x + \frac{3}{2}$

Example



2. Model is a parabola: $\widehat{y}_i = ax_i^2 + bx_i + c$

Total squared error is
$$T = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{3} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{3} \left(y_i - ax_i^2 - bx_i - c\right)^2$$
$$= (3 - a - b - c)^2 + (1 - 4a - 2b - c)^2$$
$$+(4 - 16a - 4b - c)^2$$

If you take the partial derivatives and set to 0, you would get the following system of linear equtions:

$273a + 73b + 21c = 71$

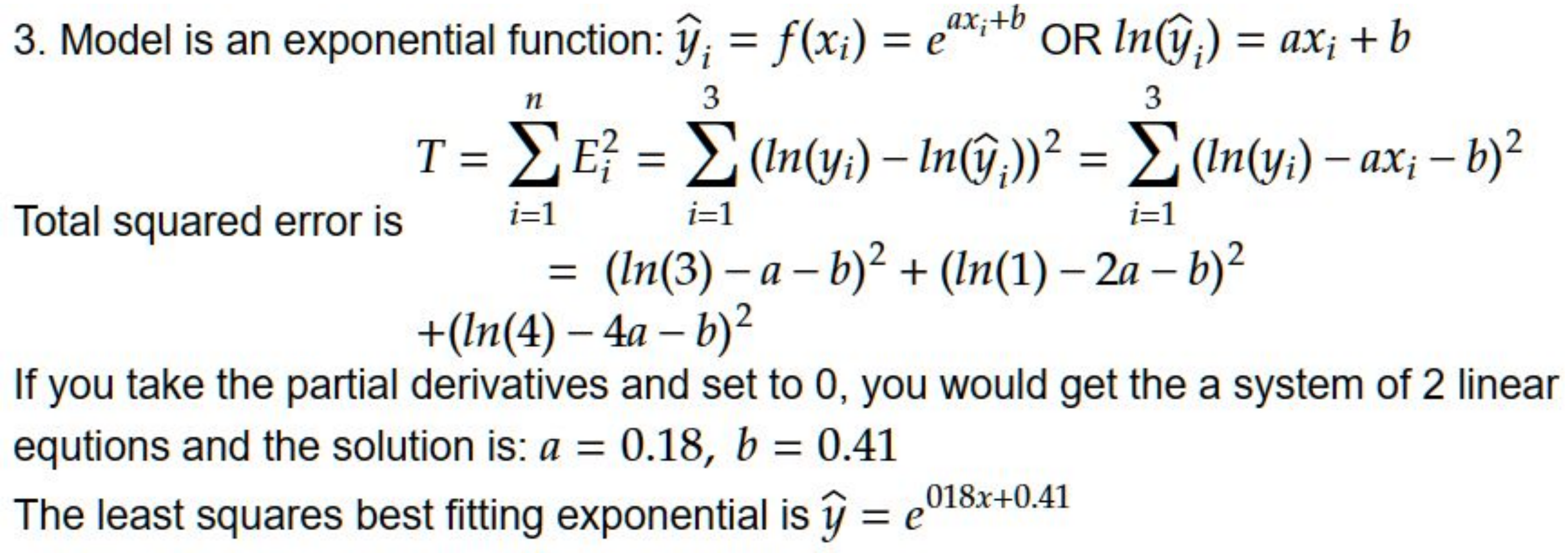
$73a + 21b + 7c = 21$

$21a + 7b + 3c = 8$

The solution is: $a = 1.2,\ b = -5.5,\ c = 7.3$

The least squares best fitting parabola is $y = 1.2x^2 - 5.5x + 7.3$

Example



3. Model is an exponential function: $\widehat{y}_i = f(x_i) = e^{ax_i+b}$ OR $ln(\widehat{y}_i) = ax_i + b$

Total squared error is $T = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{3} (ln(y_i) - ln(\widehat{y}_i))^2 = \sum_{i=1}^{3} (ln(y_i) - ax_i - b)^2$

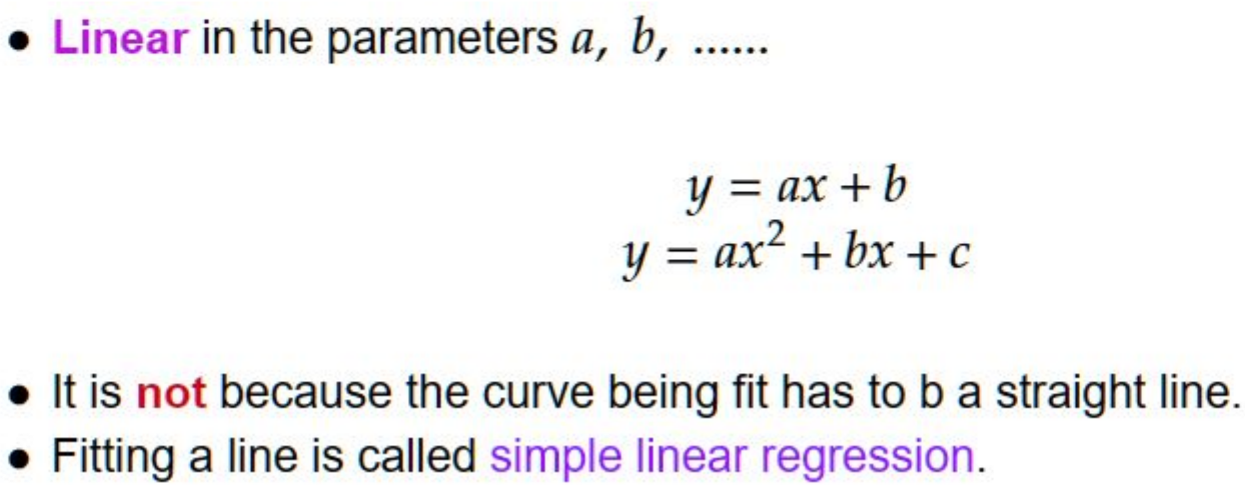
$$= (ln(3) - a - b)^2 + (ln(1) - 2a - b)^2$$

$$+(ln(4) - 4a - b)^2$$

If you take the partial derivatives and set to 0, you would get the a system of 2 linear equtions and the solution is: $a = 0.18, \ b = 0.41$

The least squares best fitting exponential is $\widehat{y} = e^{018x+0.41}$

What is linear about linear regression?



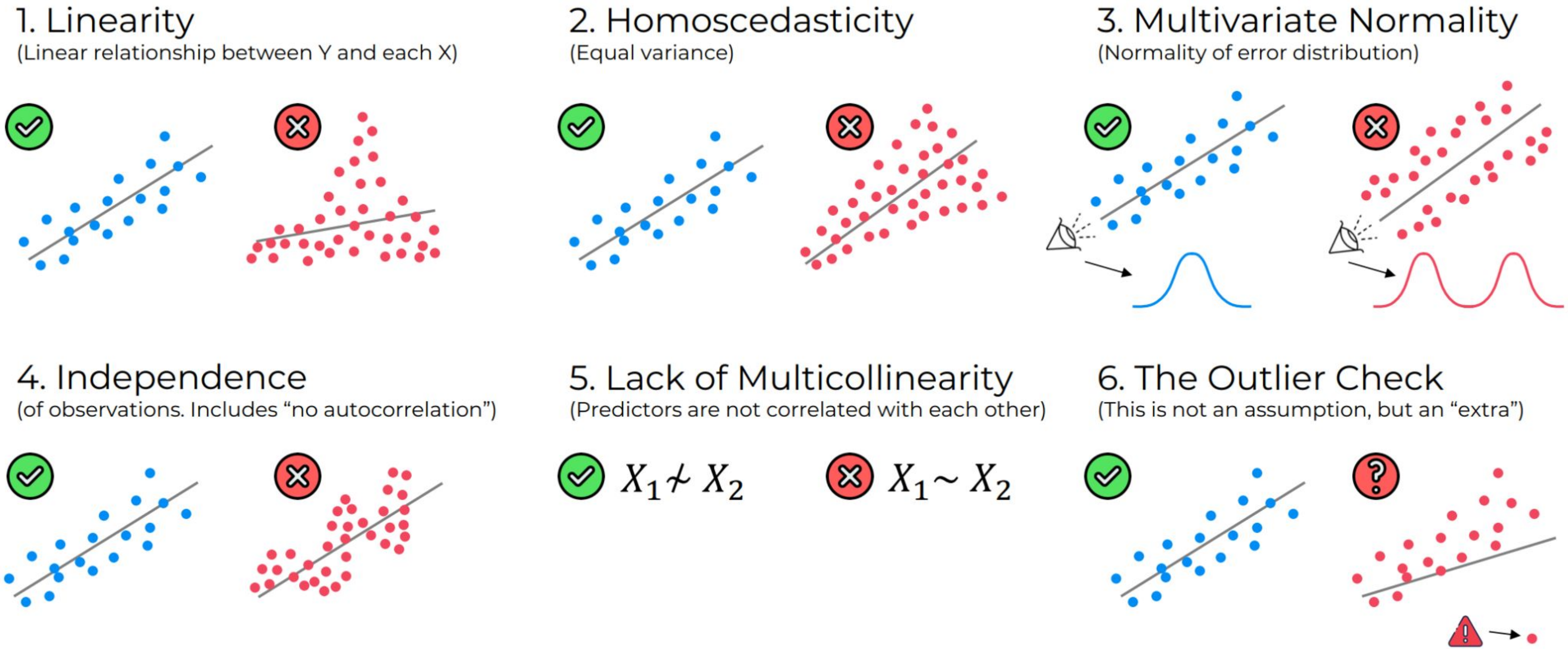
- **Linear** in the parameters $a,\ b,\ \ldots\ldots$

$$y = ax + b$$
$$y = ax^2 + bx + c$$

- It is **not** because the curve being fit has to b a straight line.
- Fitting a line is called simple linear regression.

- source1
- source2
- residual analysis - colab notebook

Big Assumptions in linear regression



1. Linearity
(Linear relationship between Y and each X)

2. Homoscedasticity
(Equal variance)

3. Multivariate Normality
(Normality of error distribution)

4. Independence
(of observations. Includes “no autocorrelation”)

5. Lack of Multicollinearity
(Predictors are not correlated with each other)

✓ $X_1 \nsim X_2$ ✗ $X_1 \sim X_2$

6. The Outlier Check
(This is not an assumption, but an “extra”)

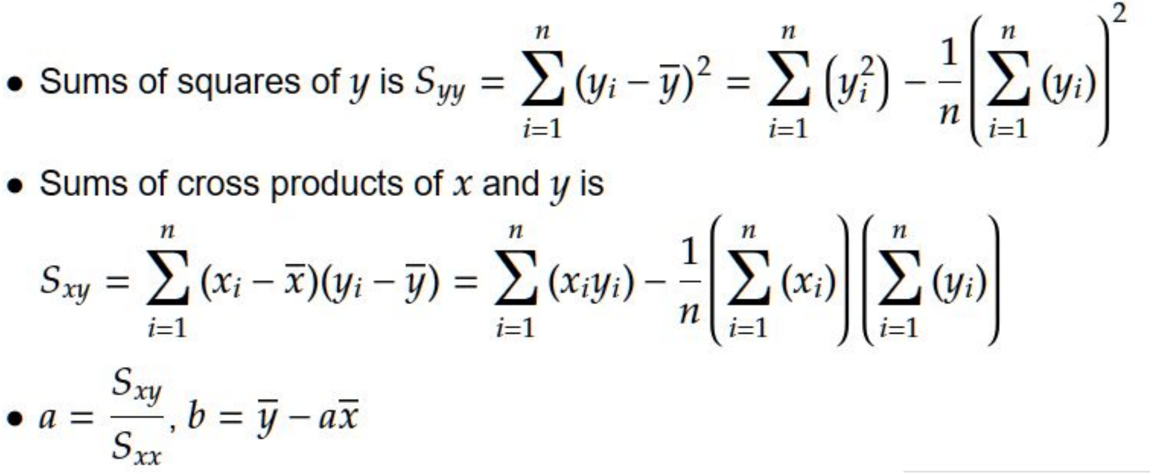
Formulas for simple linear regression

- Bivariate data $(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_n, y_n)$.
- Model:

$$y_i = ax_i + b + E_i \ \ where \ \ E_i \sim N\left(0, \sigma^2\right)$$

- The sample mean for x is $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
- The sample mean for y is $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$
- Sums of squares of x is $S_{xx} = \sum\limits_{i=1}^{n} (x_i - \overline{x})^2 = \sum\limits_{i=1}^{n} \left(x_i^2\right) - \dfrac{1}{n}\left(\sum\limits_{i=1}^{n} (x_i)\right)^2$

Formulas for simple linear regression

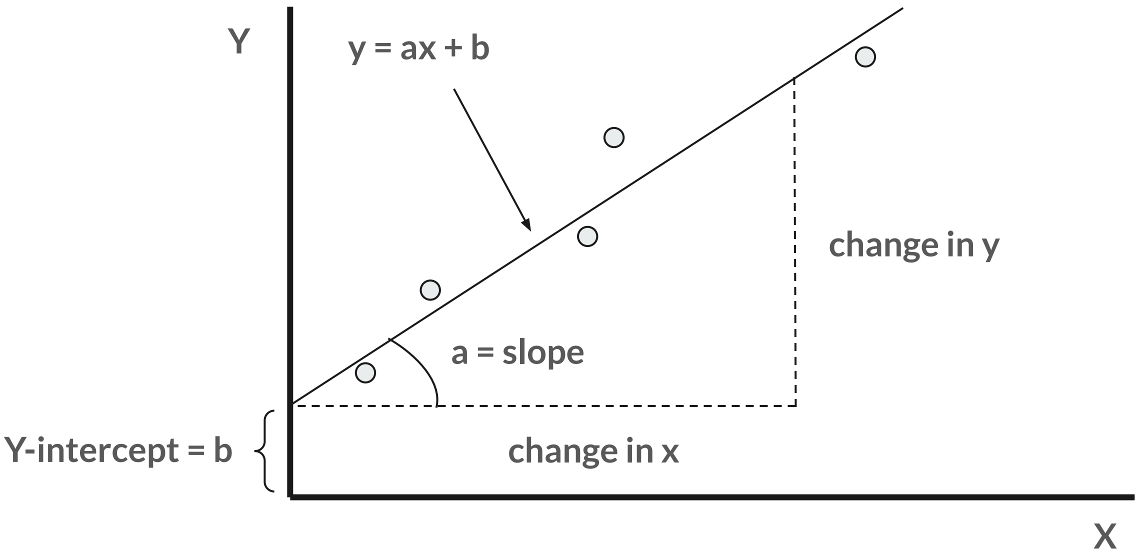


- Sums of squares of y is $S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} \left(y_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} (y_i)\right)^2$
- Sums of cross products of x and y is

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} (x_i y_i) - \frac{1}{n}\left(\sum_{i=1}^{n} (x_i)\right)\left(\sum_{i=1}^{n} (y_i)\right)$$

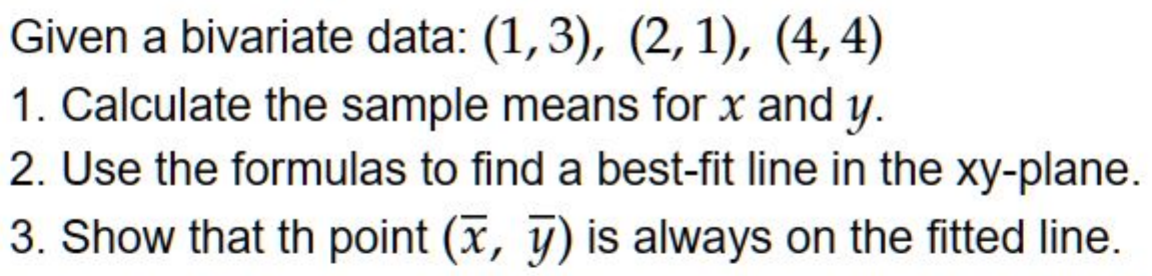
- $a = \frac{S_{xy}}{S_{xx}}, \; b = \overline{y} - a\overline{x}$

Graphical representation of simple linear regression model



source

Example



Given a bivariate data: $(1, 3),\ (2, 1),\ (4, 4)$

1. Calculate the sample means for x and y.
2. Use the formulas to find a best-fit line in the xy-plane.
3. Show that th point $(\overline{x},\ \overline{y})$ is always on the fitted line.

Example

Solution:

1. $\bar{x} = \frac{1+2+4}{3} = \frac{7}{3}, \ \bar{y} = \frac{8}{3}$

2. $S_{xx} = \sum_{i=1}^{n}\left(x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n}(x_i)\right)^2 = \left(1^2+2^2+4^2\right) - \frac{1}{3}(1+2+4)^2 = 21 - \frac{49}{3} = \frac{14}{3}$

$$S_{xy} = \sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\left(\sum_{i=1}^{n}(x_i)\right)\left(\sum_{i=1}^{n}(y_i)\right) = (1*3+2*1+4*4) - \frac{1}{3}(1+2+4)(3+1+4)$$

$$= 21 - \frac{56}{3} = \frac{7}{3}$$

$$a = \frac{S_{xy}}{S_{xx}} = \frac{7}{14} = \frac{1}{2}, \ b = \bar{y} - a\bar{x} = \frac{8}{3} - \frac{1}{2}*\frac{7}{3} = \frac{9}{6} = \frac{3}{2}$$

The best fitting line is $y = \frac{1}{2}x + \frac{3}{2}$

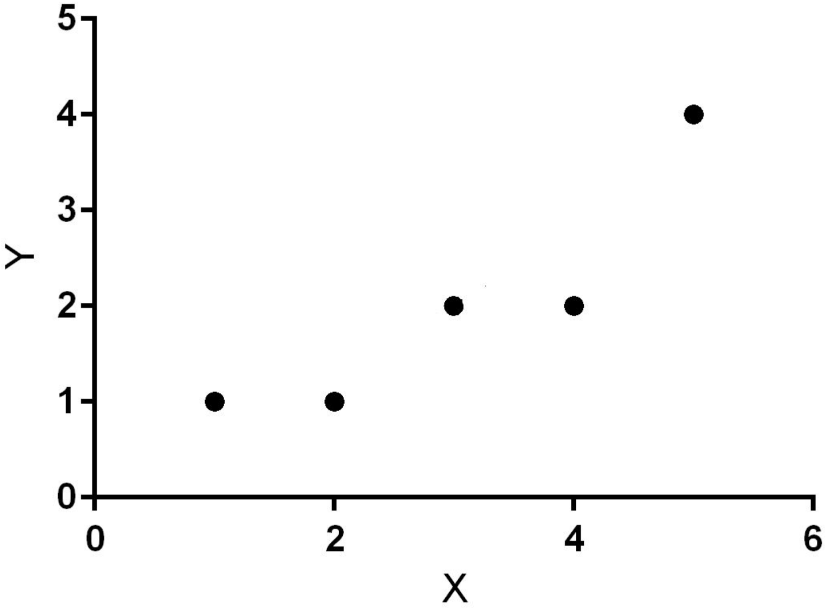
3. The formula $b = \bar{y} - a\bar{x}$ is exactly the same as $\bar{y} = a\bar{x} + b$. That is the point $(\bar{x}, \bar{y})$ is always on the line $y = ax + b$.

Example

- You are a marketing analyst in an e-commerce company. You gather the following data:
- The independent variable x
  - Ad expenditure (predictor)
- The dependent variable y
  - Sales revenue (response)

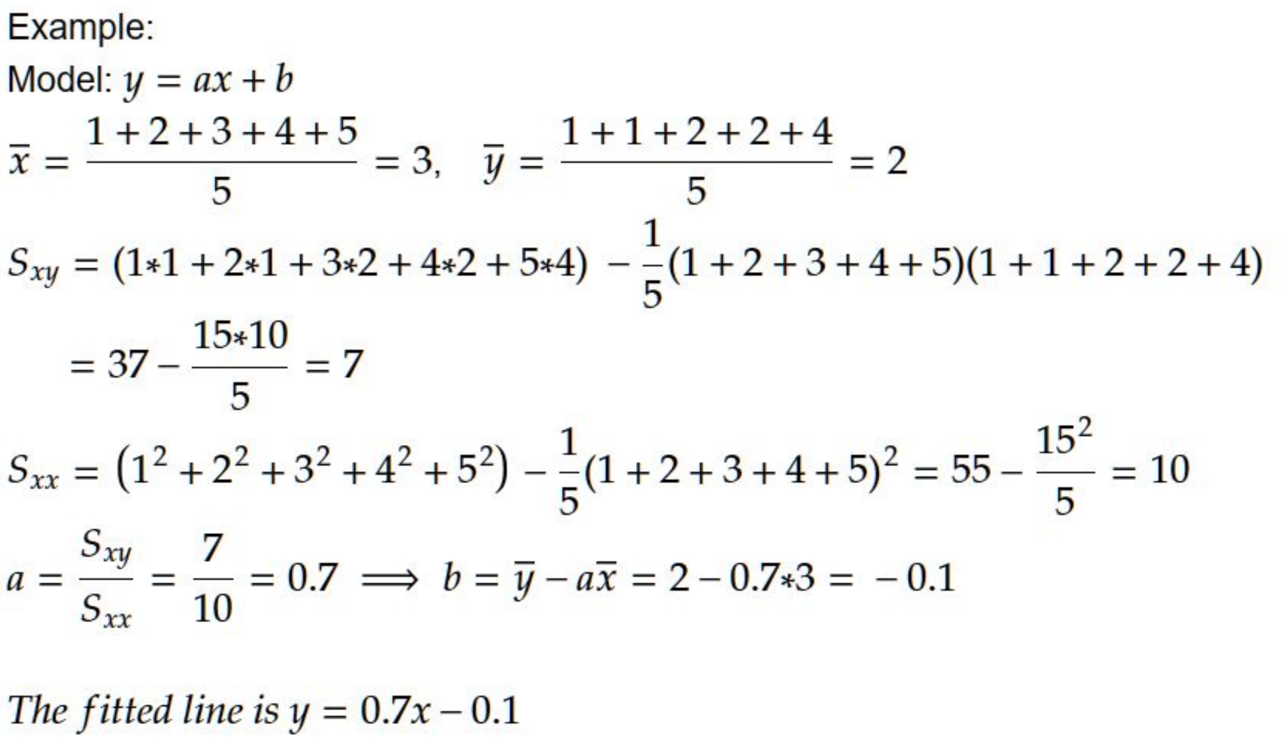
Ad expenditure ($100)	Sales revenue ($1000)
1	1
2	1
3	2
4	2
5	4

Example



Ad expenditure ($100) X	Sales revenue ($1000) Y
1	1
2	1
3	2
4	2
5	4

Example



Example:

Model: $y = ax + b$

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3, \quad \bar{y} = \frac{1+1+2+2+4}{5} = 2$$

$$S_{xy} = (1*1 + 2*1 + 3*2 + 4*2 + 5*4) - \frac{1}{5}(1+2+3+4+5)(1+1+2+2+4)$$

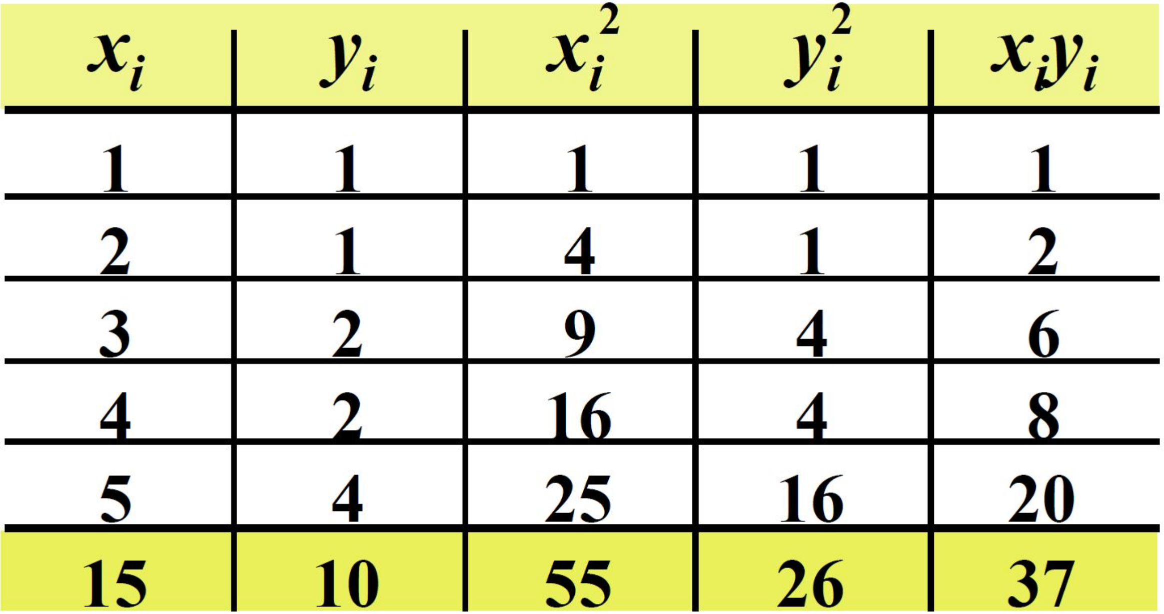
$$= 37 - \frac{15*10}{5} = 7$$

$$S_{xx} = (1^2 + 2^2 + 3^2 + 4^2 + 5^2) - \frac{1}{5}(1+2+3+4+5)^2 = 55 - \frac{15^2}{5} = 10$$

$$a = \frac{S_{xy}}{S_{xx}} = \frac{7}{10} = 0.7 \implies b = \bar{y} - a\bar{x} = 2 - 0.7*3 = -0.1$$

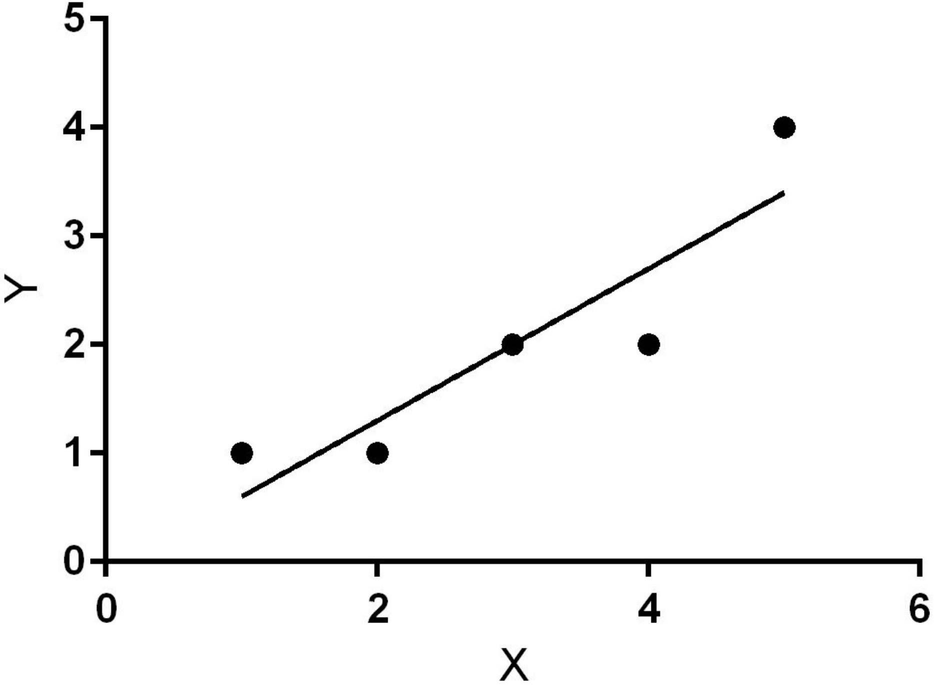
$\textit{The fitted line is } y = 0.7x - 0.1$

Example



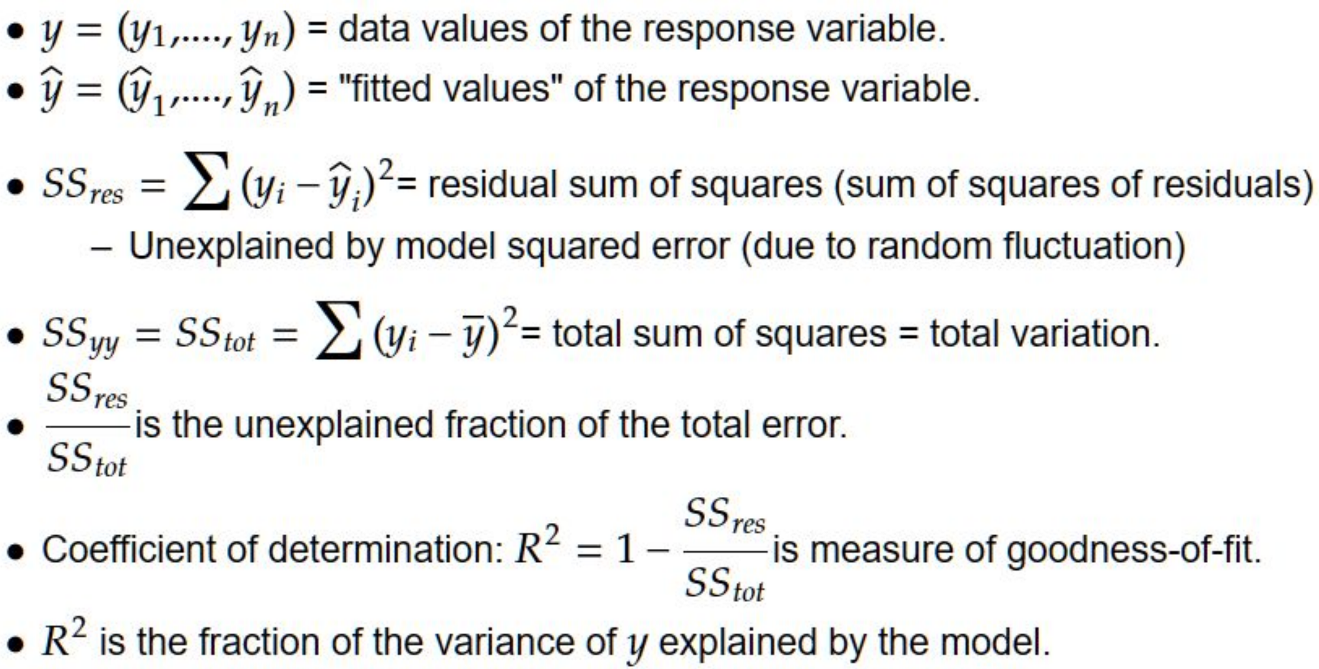
x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

Example



SS_res is a non-standardized measure

Measuring the fit – R^2



- $y = (y_1, \ldots, y_n)$ = data values of the response variable.
- $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$ = "fitted values" of the response variable.
- $SS_{res} = \sum (y_i - \hat{y}_i)^2$= residual sum of squares (sum of squares of residuals)
  - Unexplained by model squared error (due to random fluctuation)
- $SS_{yy} = SS_{tot} = \sum (y_i - \bar{y})^2$= total sum of squares = total variation.
- $\dfrac{SS_{res}}{SS_{tot}}$ is the unexplained fraction of the total error.
- Coefficient of determination: $R^2 = 1 - \dfrac{SS_{res}}{SS_{tot}}$ is measure of goodness-of-fit.
- R^2 is the fraction of the variance of y explained by the model.

Example

- You are a marketing analyst in an e-commerce company. You gather the following data:
- The independent variable x
  - Ad expenditure (predictor)
- The dependent variable y
  - Sales revenue (response)

Ad expenditure ($100)	Sales revenue ($1000)
1	1
2	1
3	2
4	2
5	4

Example

$$SS_{yy} = SS_{tot} = \sum (y_i - \bar{y})^2 = (1-2)^2 + (1-2)^2 + (2-2)^2 + (2-2)^2 + (4-2)^2$$
$$= 1 + 1 + 0 + 0 + 4 = 6$$

x_i	y_i	$\widehat{y}_i$	$y_i - \widehat{y}_i$	$(y_i - \widehat{y}_i)^2$
1	1	0.6	0.4	0.16
2	1	1.3	-0.3	0.09
3	2	2	0	0
4	2	2.7	-0.7	0.49
5	4	3.4	0.6	0.36

$$SS_{res} = \sum (y_i - \widehat{y}_i)^2 = 0.16 + 0.09 + 0 + 0.49 + 0.36 = 1.1$$

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = \frac{6 - 1.1}{6} = \frac{4.9}{6} = 0.8167 \simeq 81\%$$

The model "fitted line" can explain 81% of the variation in y.

Example

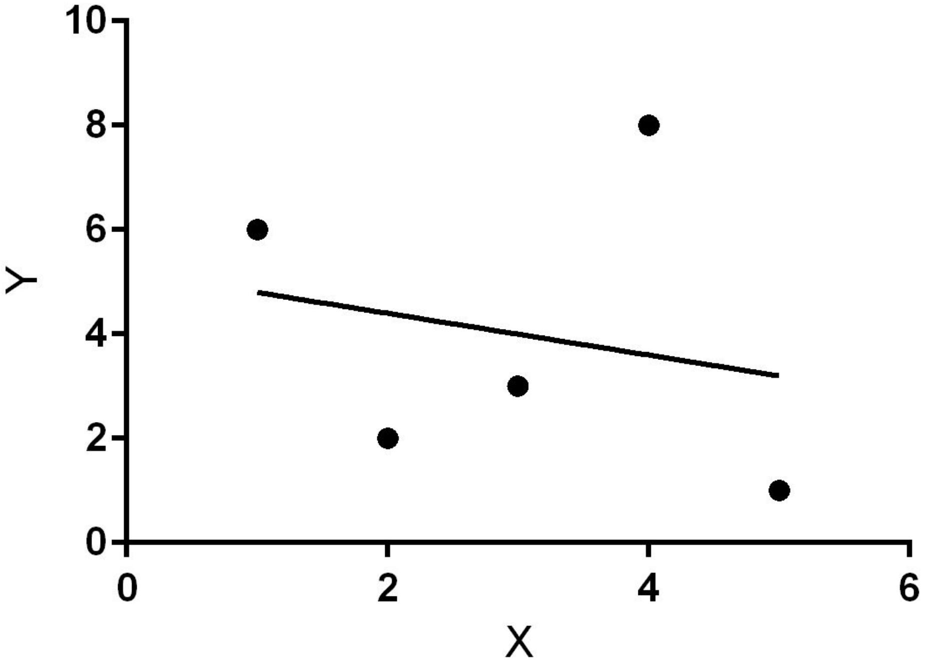
- You are a marketing analyst in an e-commerce company. You gather the following data:
- The independent variable x
  - Ad expenditure (predictor)
- The dependent variable y
  - Sales revenue (response)

Ad expenditure ($100)	Sales revenue ($1000)
1	6
2	2
3	3
4	8
5	1

Example

$$y = -0.4x + 5.2$$
$$R^2 = 0.04706 = 4\%$$

4% of the variation in y is explained by the model



Example

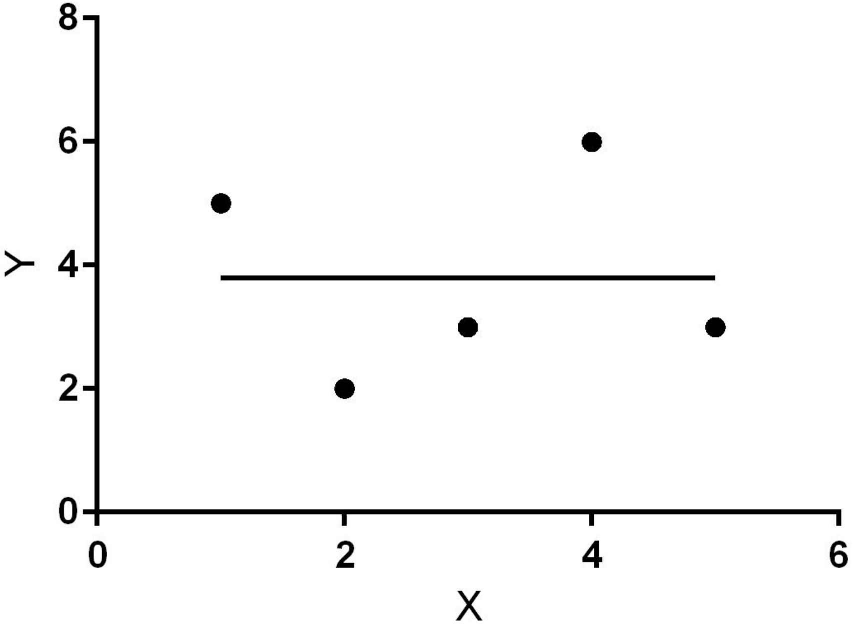
- You are a marketing analyst in an e-commerce company. You gather the following data:
- The independent variable x
  - Ad expenditure (predictor)
- The dependent variable y
  - Sales revenue (response)

Ad expenditure ($100)	Sales revenue ($1000)
1	5
2	2
3	3
4	6
5	3

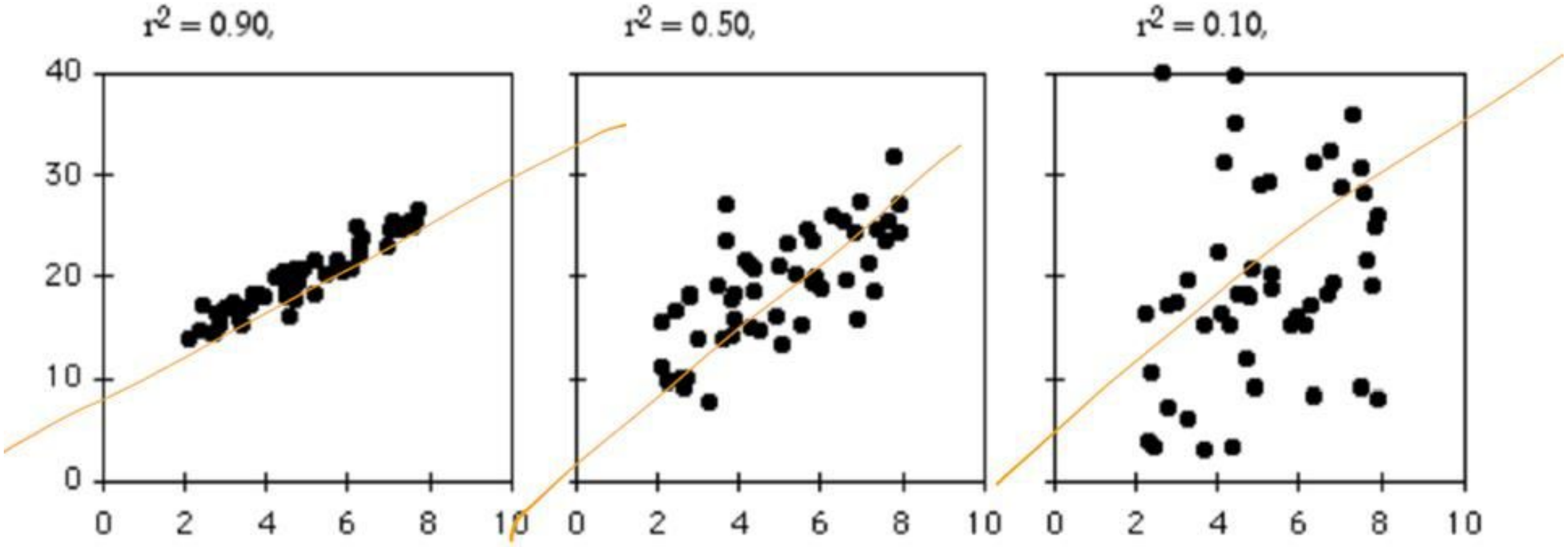
Example

$$y = 0x + 3.8 \implies y = 3.8$$
$$R^2 = 0 \implies R^2 = 0\%$$

The model cannot explain any variation in y and our model is useless.



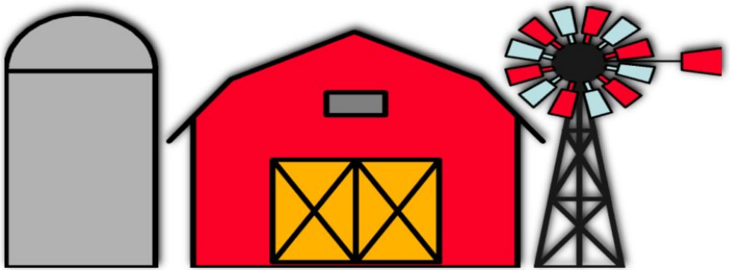
Coefficient of determination



Home assignment

You're an economist for the county cooperative. You gather the following data:

Fertilizer (lb.)	Yield (lb.)
4	3.0
6	5.5
10	6.5
12	9



Find the **least squares line** relating crop yield and fertilizer.

Attendance

https://baam.tatar

Questions?