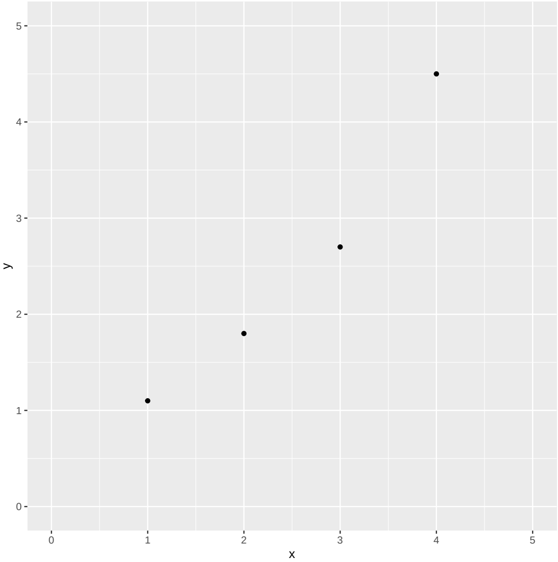
Simple Linear Model

1. Motivation

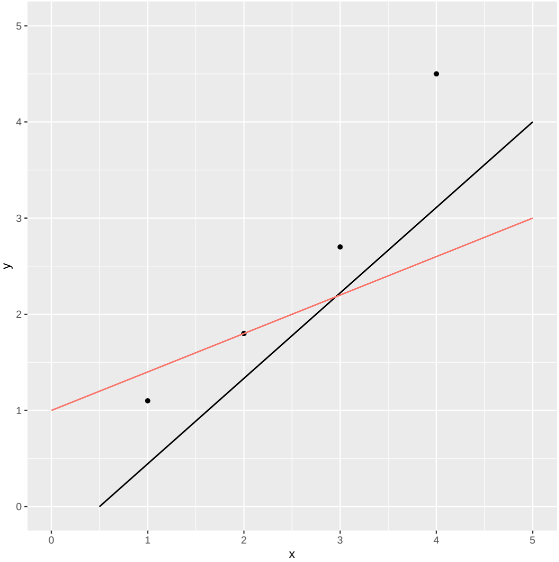
- Given the data

x	y
1	1.1
2	1.8
3	2.7
4	4.5

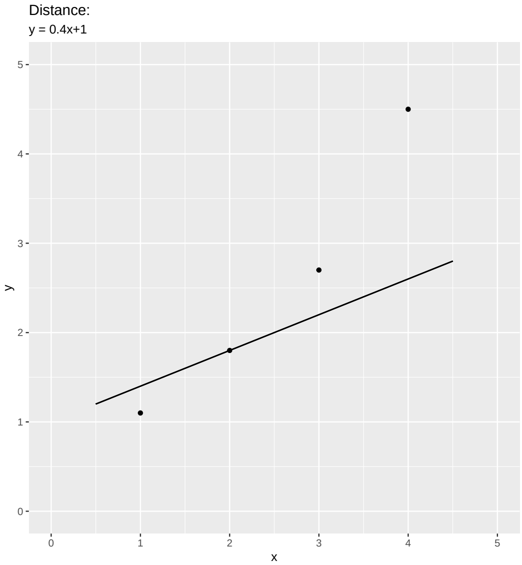
Scatter plot



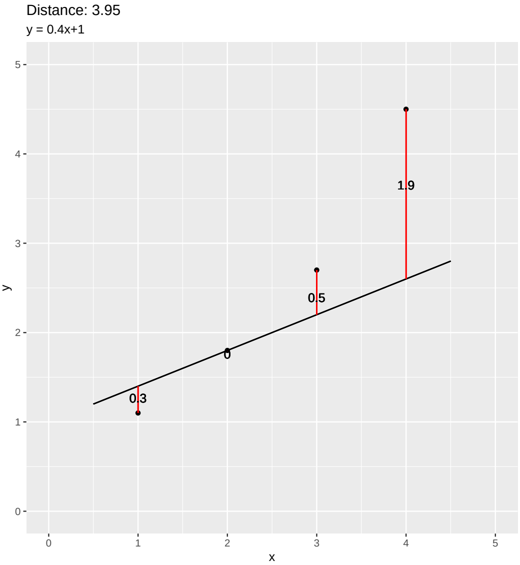
Which line is closer to the points?



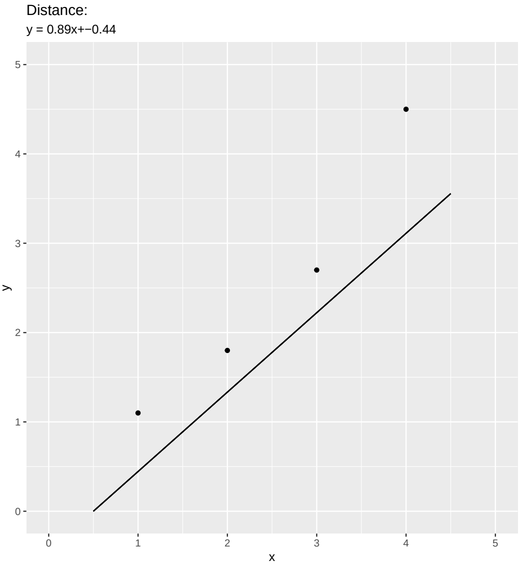
Squared Distance between a line and points



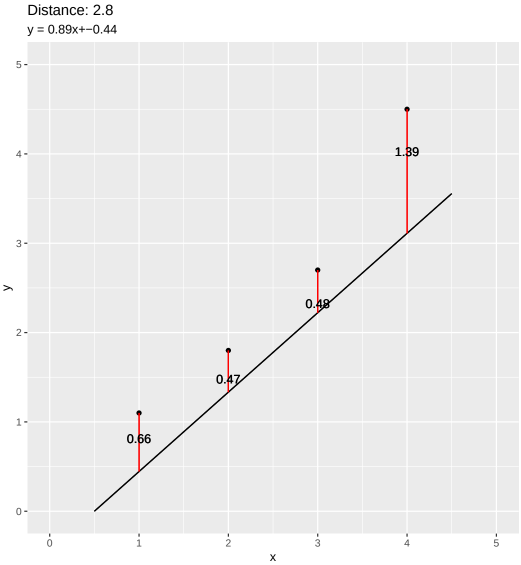
Squared Distance between a line and points



Squared Distance between a line and points



Squared Distance between a line and points



Linear Model

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

- Model Assumptions
  - (A1) The response variable y_i is a random variable and the predictor x_i is non-random
  - (A2) $\epsilon_i \sim^{iid} N(0, \sigma^2)$

Parameters Estimation

The best fitted line

- The least squared methods give us the formula for the closest line or the best fitted line:

$$y = \hat{\beta}_1 x + \hat{\beta}_0$$

- The estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ are

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

Example: Calculate from Data

x	y
1	1.1
2	1.8
3	2.7
4	4.5

	x	y	xy	x^2
	1	1.1		
	2	1.8		
	3	2.7		
	4	4.5		
$\sum$				

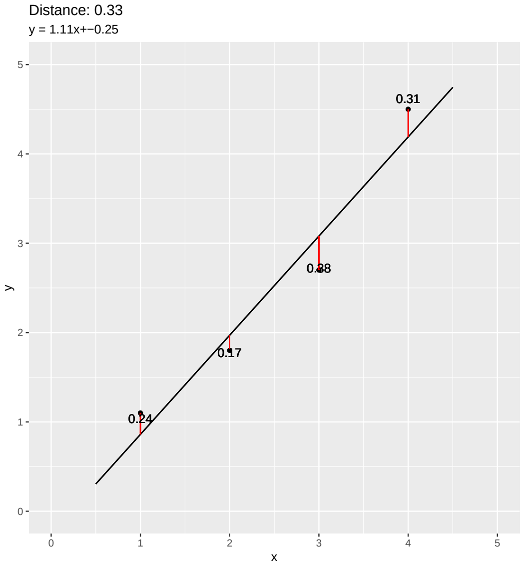
	x	y	xy	x^2
	1	1.1		
	2	1.8		
	3	2.7		
	4	4.5		
$\sum$				

- $\bar{x} = \frac{1+2+3+4}{4} = 2.5$
- $\bar{y} = \frac{1.1+1.8+2.4+4.5}{4} = 2.525$

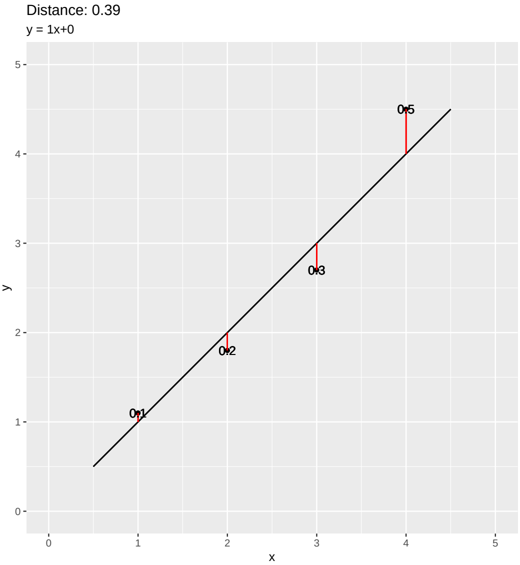
	x	y	xy	x^2
	1	1.1	1.1	1
	2	1.8	3.6	4
	3	2.7	8.1	9
	4	4.5	18	16
$\sum$			30.8	30

- $\hat{\beta}_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = 1.11$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -0.25$

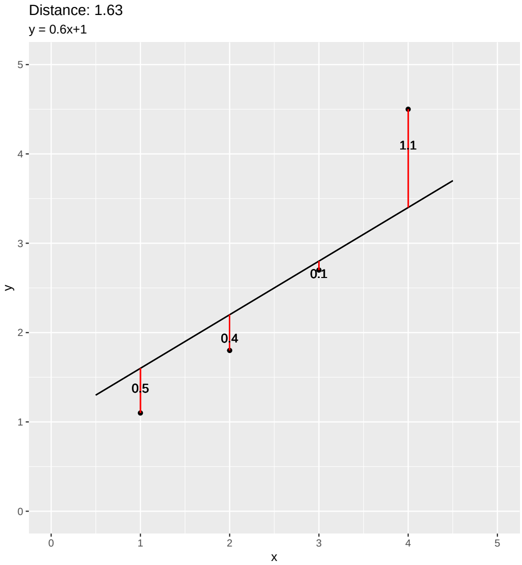
Best fitted line



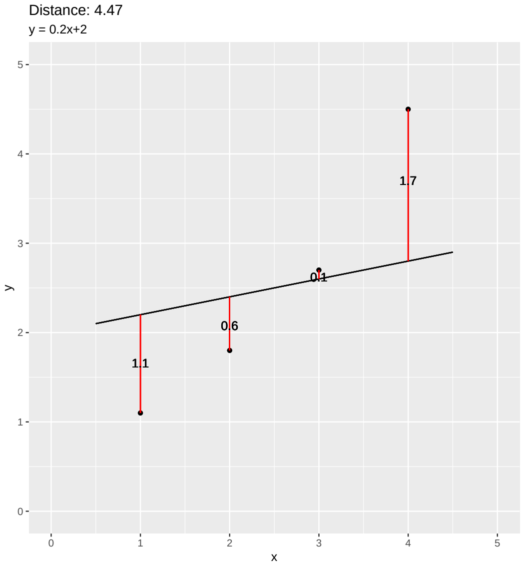
Some other lines



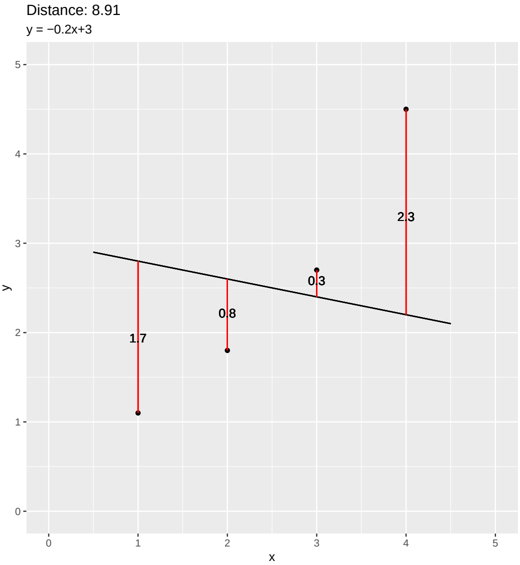
Some other lines



Some other lines



Some other lines



Example: Calculate from Sumations

The regression model is $y = \beta_0 + \beta_1 x + \epsilon$. There are six observations. The summary statistics are

$$\sum y_i = 42,$$
$$\sum x_i = 21,$$
$$\sum x_i^2 = 91,$$
$$\sum x_i y_i = 187,$$
$$\sum y_i^2 = 390$$

Calculate the least squares estimate of β_1.

Example: Calculate from Sumations

The regression model is $y = \beta_0 + \beta_1 x + \epsilon$. There are five observations. The summary statistics are

$$\sum y_i = 30,$$
$$\sum x_i = 15,$$
$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 25,$$
$$\sum (x_i - \bar{x})^2 = 10,$$
$$\sum (y_i - \bar{y})^2 = 64,$$

Write the equation of the best fitted line using the least squares method.

Goodness of Fit

Coefficient of Determination

- Baseline model:

$$y = \beta_0 + \epsilon$$

  - In this model, y_i is estimated by one number, $\bar{y}$
- Linear Model:

$$y = \beta_0 + \beta_1 x + \epsilon$$

  - In this model, y_i is estimated by

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

$$
\begin{aligned}
\sum_{i=1}^{n}(y_i-\bar{y})^2 &= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})} \\
&= \sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2.
\end{aligned}
$$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{RSS}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Reg SS}}$$

Coefficient of Determination

$$R^2 = 1 - \frac{RSS}{TSS}$$

- R^2 runs from 0 to 1. The larger R^2, the better the model

Example

You are given the following results from a regression model.

Observation number (i)	y_i	$\hat{f}(x_i)$
1	1	1
2	2	3
3	3	7
4	5	9
5	9	10

Calculate the sum of squared errors (SSE) , the total sum squares (TSS), and the regression sum squares, and the R^2 of the model.

For a simple linear regression model the total sum of squares (TSS) is 150 and the R^2 statistic is 0.7. Calculate the sum of squares of the residuals for this model.

F-test

- i.i.d model (Baseline Model)

$$y = \beta_0 + \epsilon$$

- SLR model

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$\underbrace{\mathrm{H}_0 : \beta_1 = 0}_{\text{i.i.d. model}} \quad \text{vs.} \quad \underbrace{\mathrm{H}_a : \beta_1 \neq 0}_{\text{SLR model}},$$

$$F = \frac{\text{Reg SS}/1}{\text{RSS}/(n-2)}$$

- The smaller p-value (the larger F-statistics) supports H_1
- Small p-value ($\leq .05$): We reject H_0. The linear model is a significant improvement over the baseline model.
- Large p-value ($> .05$): Fail to reject H_0

Example

Two actuaries are analyzing car accident claims for a group of $n = 52$ participants. The predictor x is driving experience (years).

Actuary 1 uses the following regression model:

$$Y = \beta + \epsilon$$

Actuary 2 uses the following regression model:

$$Y = \beta_0 + \beta_1 \times x + \epsilon$$

The residual sum of squares for the regression of Actuary 2 is 120 and the total sum of squares is 150.

Calculate the F-statistic to test whether the model of Actuary 2 is a significant improvement over the model of Actuary 1.

t-test

- We use t-test to test the value of β_1 and β_0

$$t(\hat{\beta}_j) = \frac{\hat{\beta}_j - d}{\mathrm{SE}(\hat{\beta}_j)}, \quad j = 0,1,$$

$$\mathrm{SE}(\hat{\beta}_0) = \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} = \sqrt{\frac{s^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}$$

and $$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{s^2}{S_{xx}}}.$$

$$s^2 = \frac{\mathrm{RSS}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2},$$

Example

The results of fitting five observations by the regression model, $y = \beta_0 + \beta_1 x + \epsilon$, are given below.

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-1.5	0.7416	-2.023	0.13631
x	2.5	0.2236	11.180	0.00153

Determine the test results of the hypothesis $H_0 : \beta_1 = 0$ against $H_\alpha : \beta_1 \neq 0$.