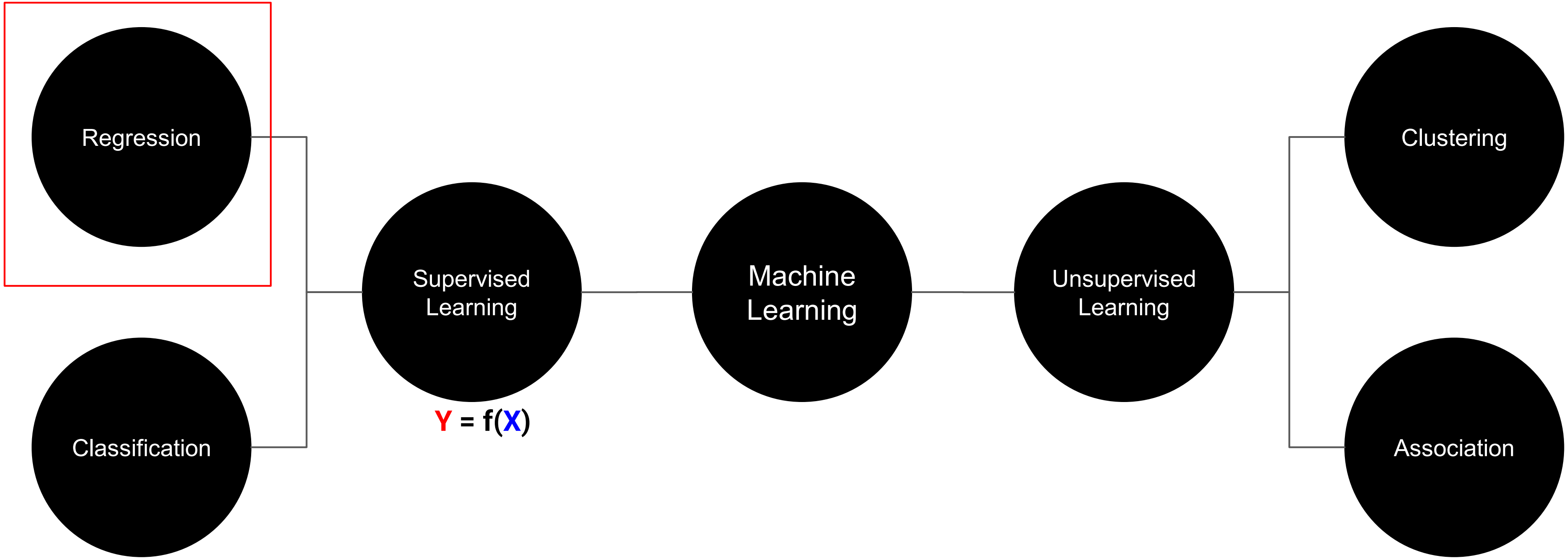
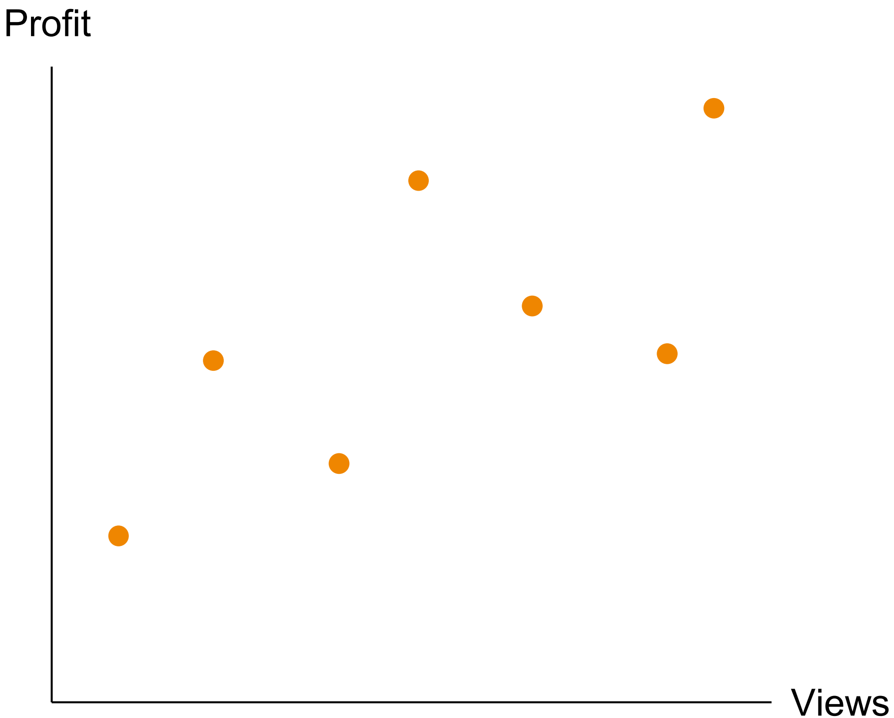
Data Prediction Model and Machine Learning

Online course #7

Regression: Linear Regression


Regression
Classification
Supervised
Learning
Y = f(X)
Machine
Learning
Unsupervised
Learning
Clustering
Association

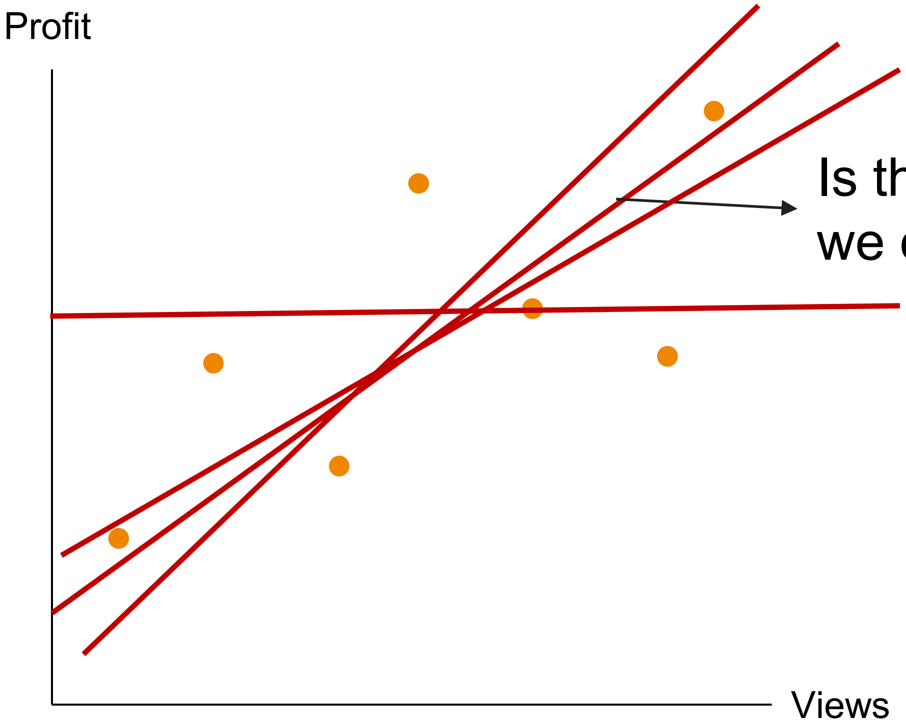
Linear Regression?



Profit of the video ∝ the number of views

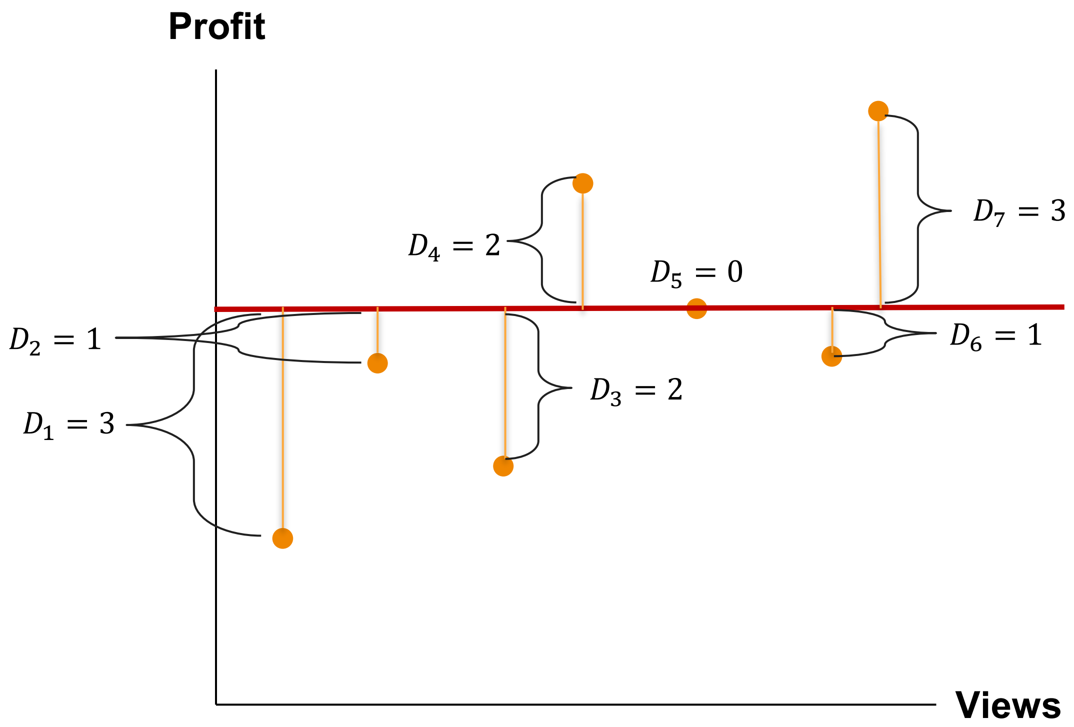
Video	Views	Profit
1	1	3
2	2	6
3	3	5
4	4	8
5	5	7
6	6	6
7	7	9

Linear Regression? Fitting a line to data



Is this line the best we can use?

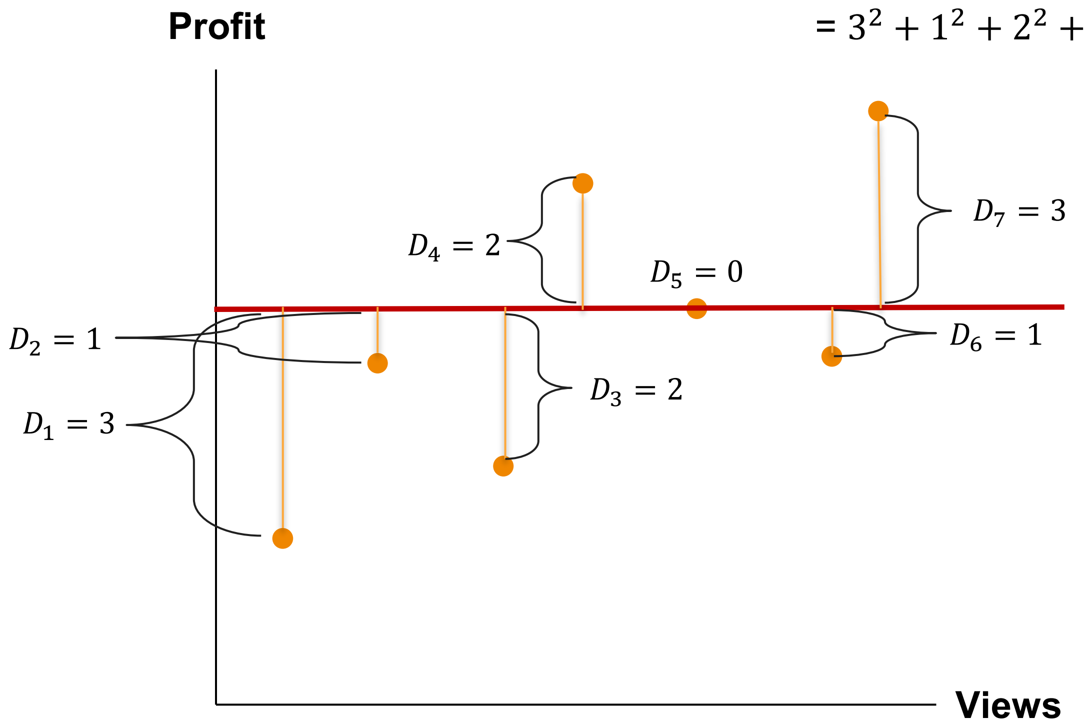
Linear Regression? Fitting a line to data



Linear Regression? Fitting a line to data

Sum of Distance squared = $D_1{}^2 + D_2{}^2 + D_3{}^2 + D_4{}^2 + D_5{}^2 + D_6{}^2 + D_7{}^2$

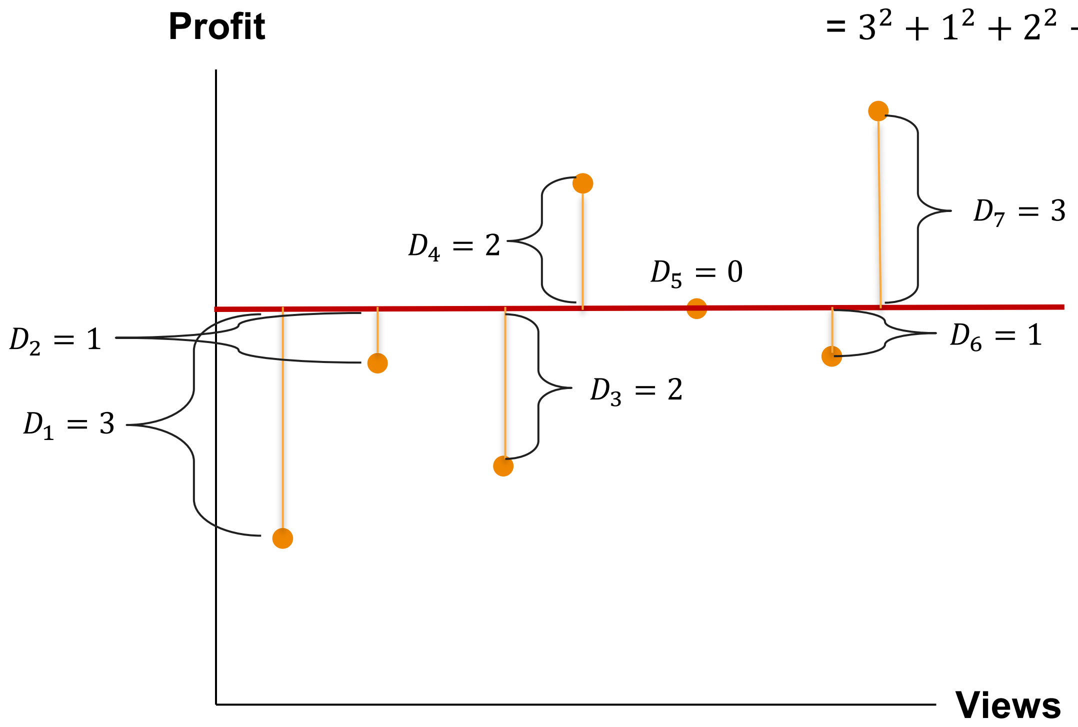
$= 3^2 + 1^2 + 2^2 + 2^2 + 0^2 + 1^2 + 3^2 = 28$



Linear Regression? Fitting a line to data

Sum of Squared Distance = ${D_1}^2 + {D_2}^2 + {D_3}^2 + {D_4}^2 + {D_5}^2 + {D_6}^2 + {D_7}^2$

$= 3^2 + 1^2 + 2^2 + 2^2 + 0^2 + 1^2 + 3^2 = 28$

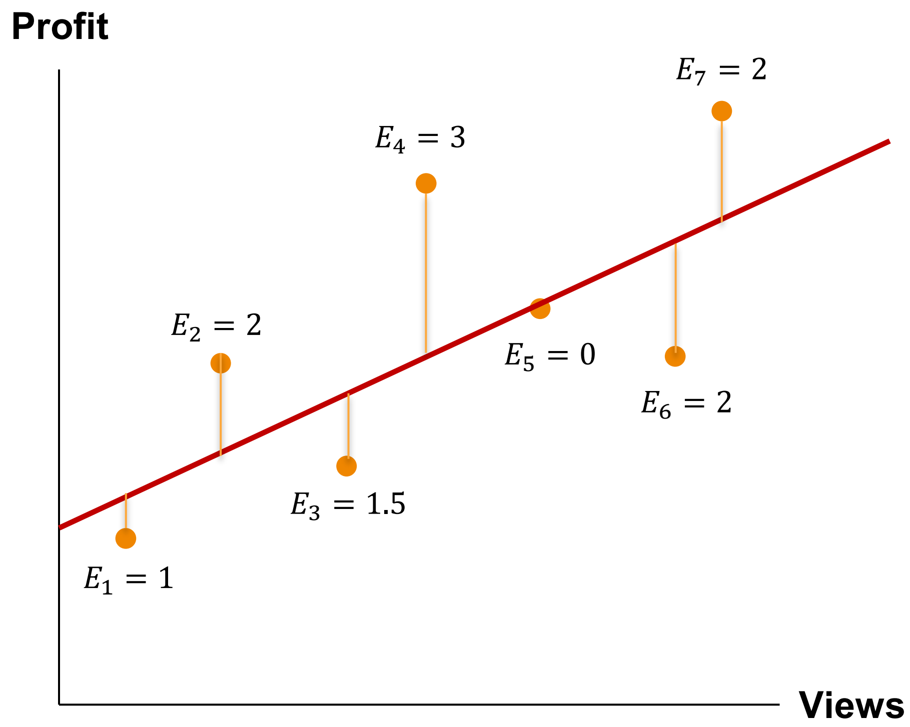


How well this line fits the date

Distance → Error

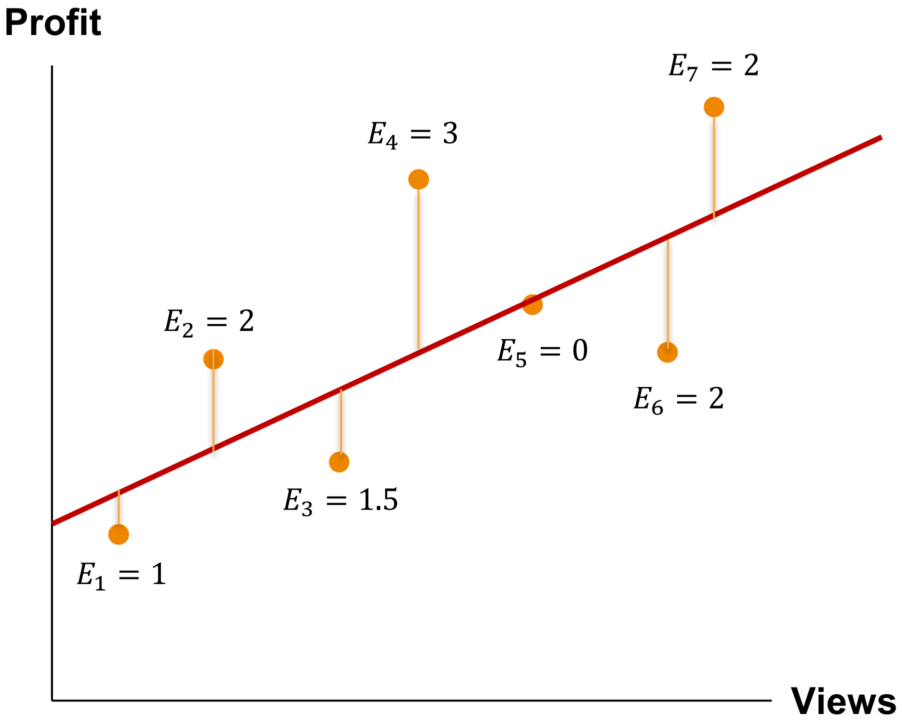
Sum of Squared Distance → Sum of Squared Errors (SSE)

Linear Regression? Fitting a line to data



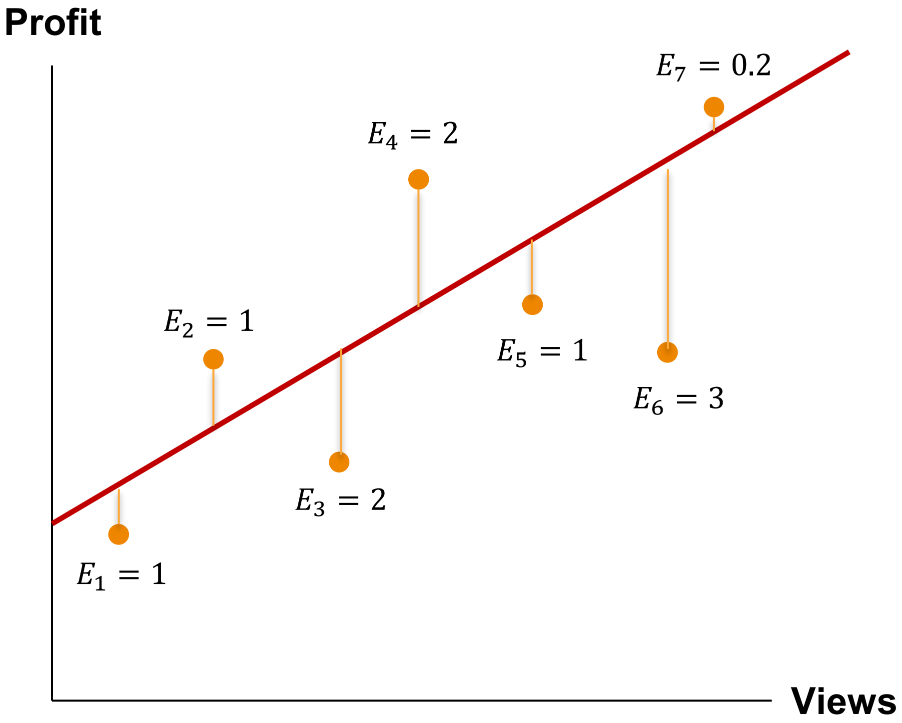
$$\text{Sum of Squared Errors} = E_1{}^2 + E_2{}^2 + E_3{}^2 + E_4{}^2 + E_5{}^2 + E_6{}^2 + E_7{}^2$$

$$SSE = \sum_{i=1}^{7} E_i{}^2 = 1^2 + 2^2 + 1.5^2 + 3^2 + 0^2 + 2^2 + 2^2 = 24.25$$



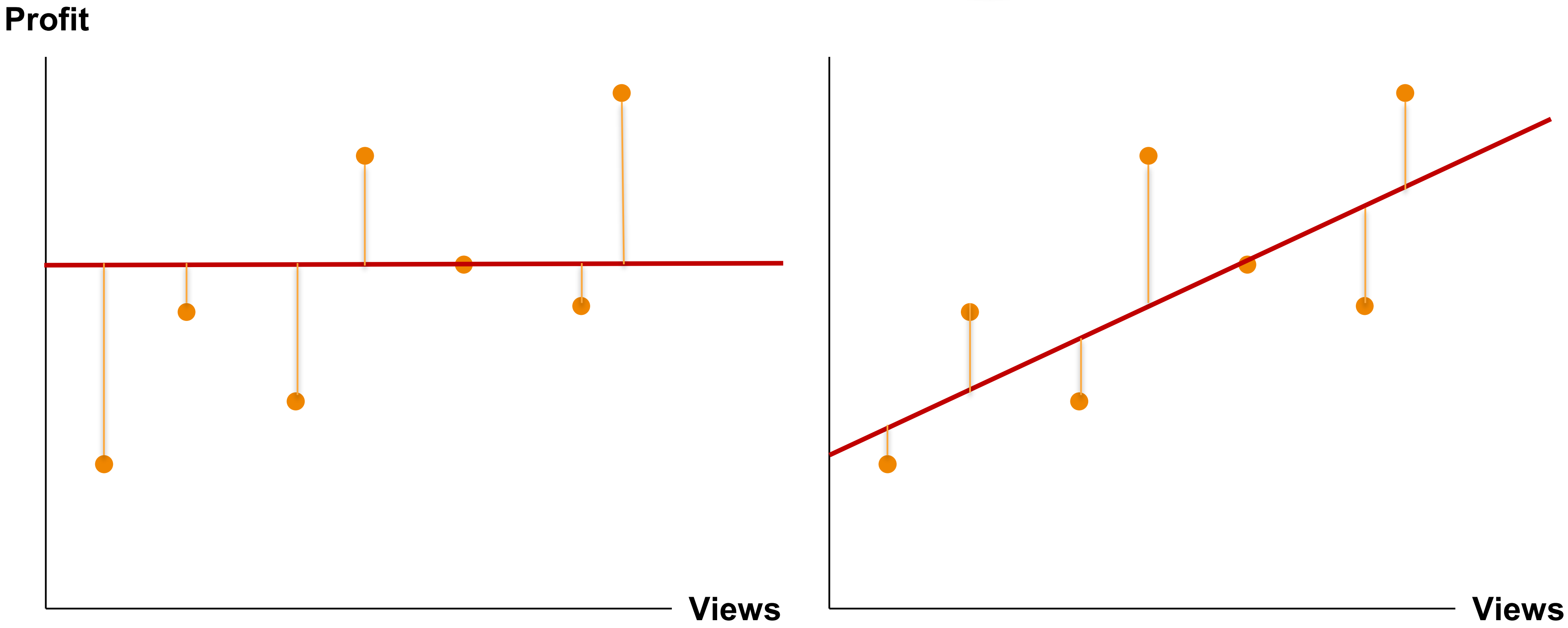
Sum of Squared Errors = ${E_1}^2 + {E_2}^2 + {E_3}^2 + {E_4}^2 + {E_5}^2 + {E_6}^2 + {E_7}^2$

$$SSE = \sum_{i=1}^{7} {E_i}^2 = 1^2 + 1^2 + 2^2 + 2^2 + 1^2 + 3^2 + 0.2^2 = 20.04$$



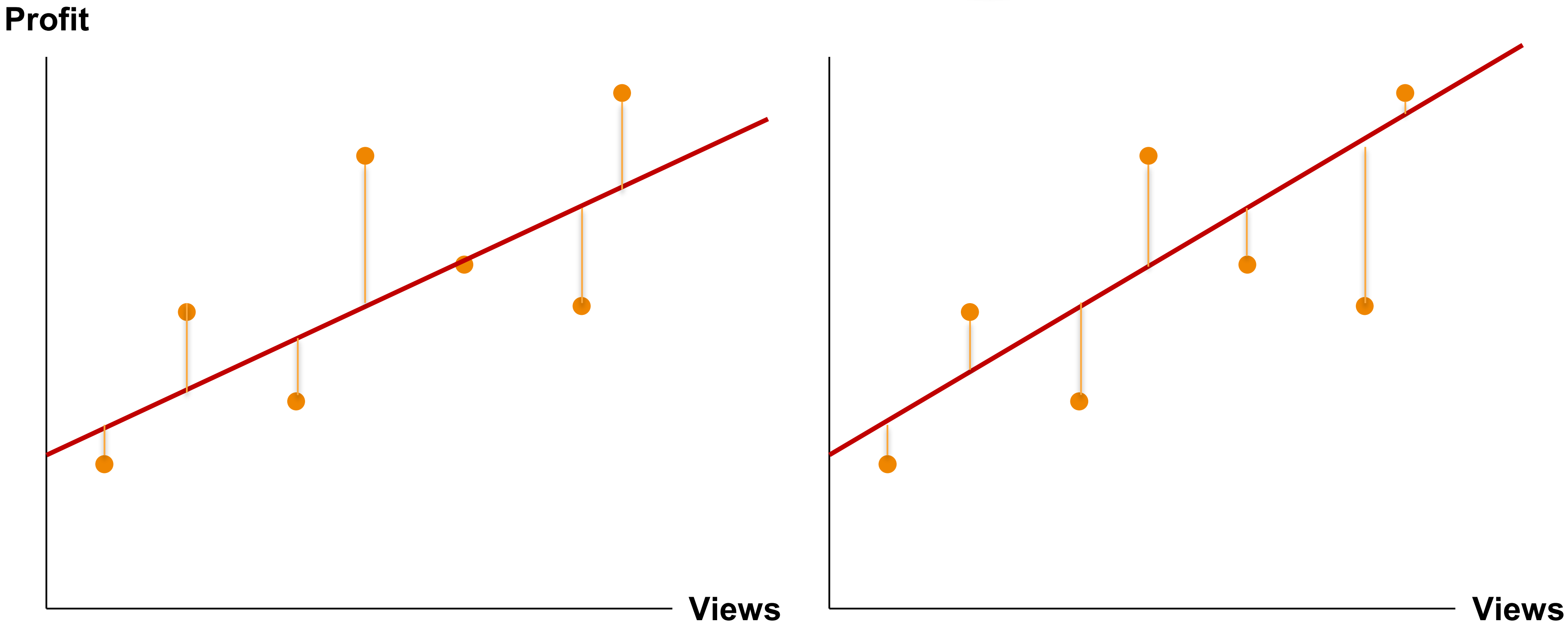
To find a better fitting line

SSE=28 > SSE=24.25



To find a better fitting line

SSE=24.25 > SSE=20.04



To minimize the SSE

Objective: Find the line which **minimizes** the sum of squared error

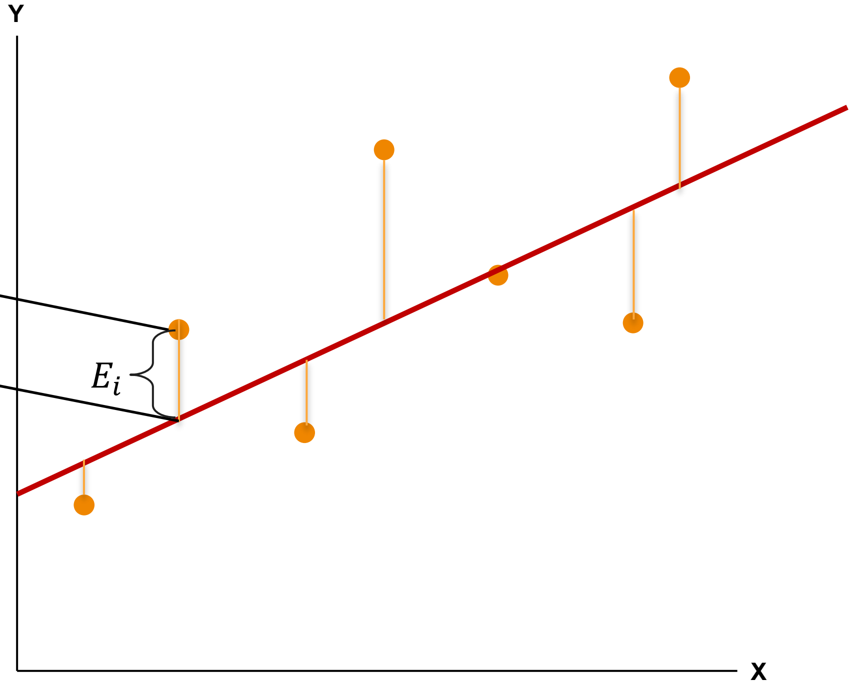
Min(SSE)

$$SSE = \sum_i {E_i}^2 = \sum_i \{Y_i - (\alpha + \beta X_i)\}^2$$

$$E_i = Y_i - \widehat{Y}_i \begin{cases} Y_i \\ \widehat{Y}_i \end{cases}$$

$$E_i = Y_i - (\alpha + \beta X_i)$$

$$\widehat{Y}_i = \alpha + \beta X_i$$



To minimize the SSE

Objective: Find the line which **minimizes** the sum of squared error

Min(SSE)

$$SSE = \sum_i {E_i}^2 = \sum_i \{Y_i - (\alpha + \beta X_i)\}^2$$

$$\frac{\partial}{\partial \alpha} \sum_i \{Y_i - (\alpha + \beta X_i)\}^2 = 0$$
$$\frac{\partial}{\partial \beta} \sum_i \{Y_i - (\alpha + \beta X_i)\}^2 = 0$$

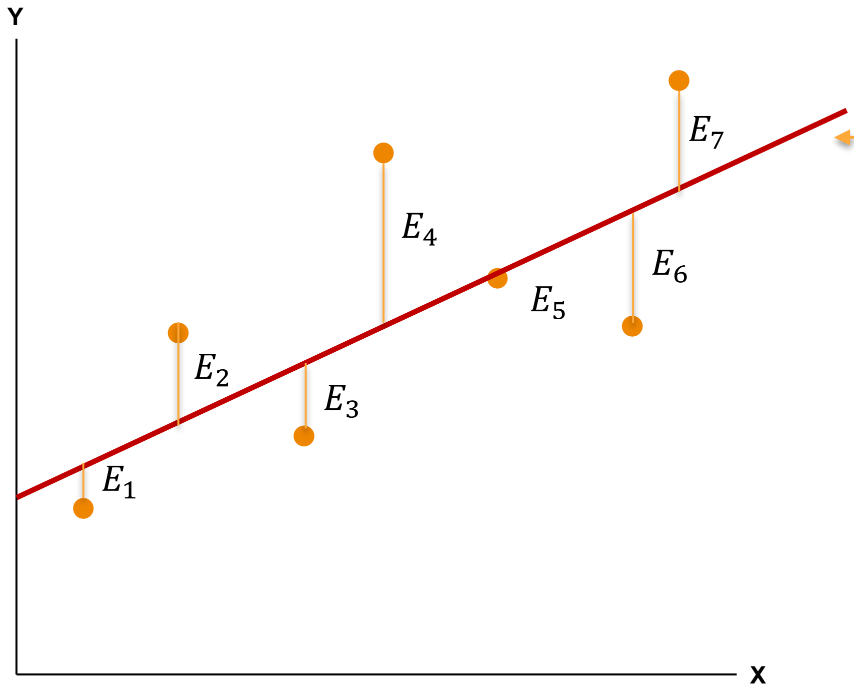
Given

$$SS_{xx} = \sum_i (X_i - \overline{X}_i)^2$$
$$SS_{xy} = \sum_i (X_i - \overline{X}_i) \times (Y_i - \overline{Y}_i)$$

$$\beta = \frac{SS_{xy}}{SS_{xx}}$$

$$\alpha = \overline{Y}_i - \beta \overline{X}_i$$

In Sum!



$$\sum_i {E_i}^2$$

$$Min(\sum_i {E_i}^2)$$

$$\hat{Y}_i = \alpha + \beta X_i$$

$$Min(\sum_i (Y_i - \hat{Y}_i)^2)$$

$$Min(\sum_i (Y_i - \alpha - \beta X_i)^2)$$

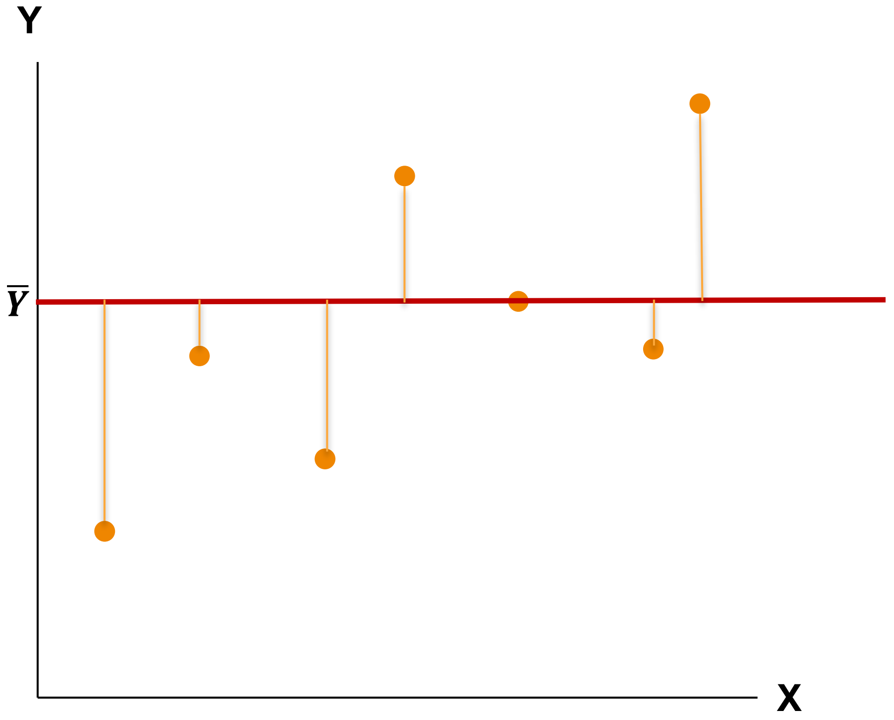
$$\alpha = \bar{Y}_i - \beta \bar{X}_i \qquad \beta = \frac{SS_{xy}}{SS_{xx}}$$

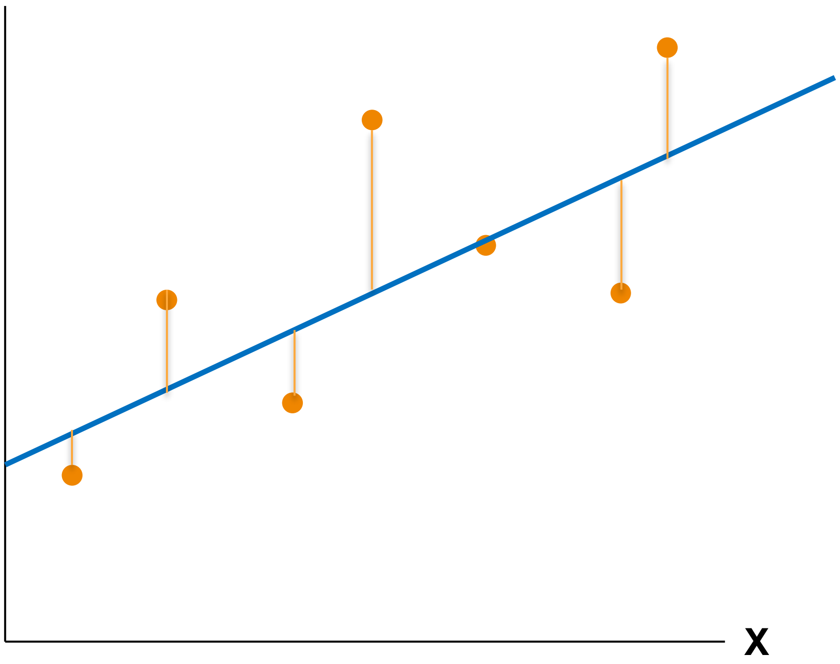
Model: $\mathbf{Y = 3.4 + 0.7X}$

How good our model is: The R squared

$$Var(Y) = \sum_{i=1}^{7} (Y - \bar{Y})^2$$

$$Var(line) = \sum_{i=1}^{7} (Y - \hat{Y}_i)^2$$



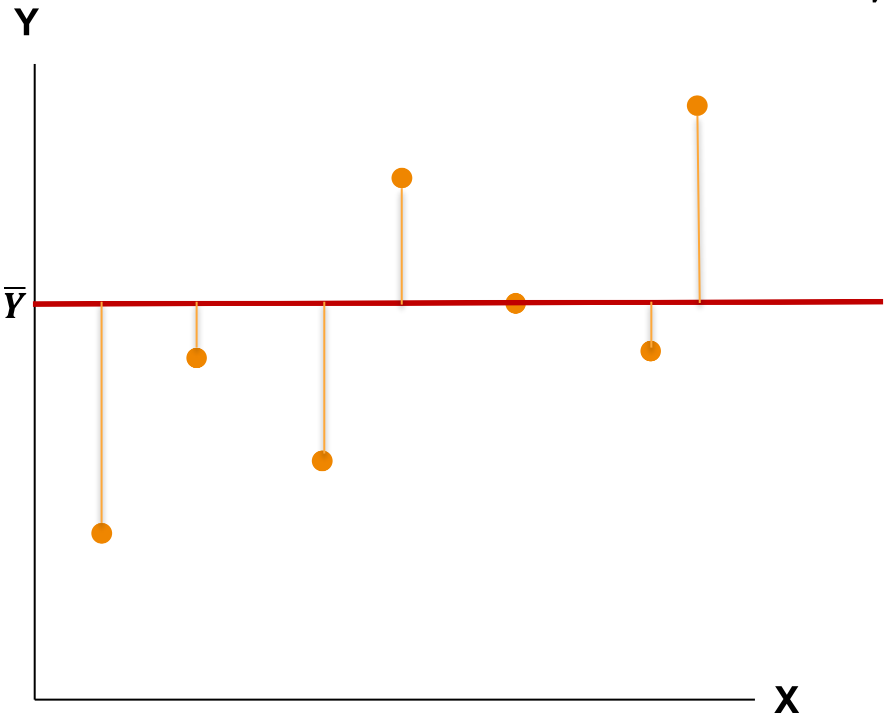


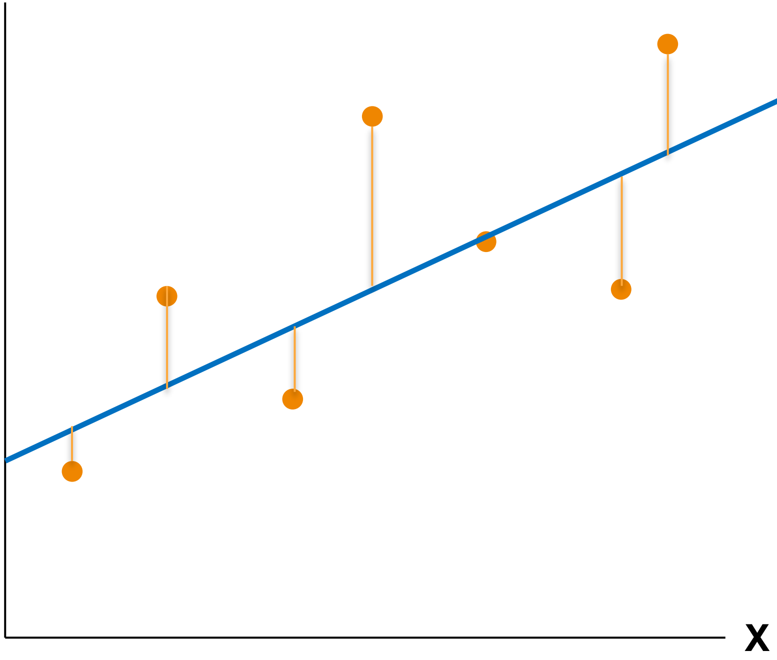
How good our model is: The R squared

$$Var(Y) = \sum_{i=1}^{7} (Y - \bar{Y})^2$$

$$R^2 = \frac{Var(Y) - Var(line)}{Var(Y)}$$

$$Var(line) = \sum_{i=1}^{7} (Y - \hat{Y}_i)^2$$





How good our model is: The R squared

$$Var(Y) = \sum_{i=1}^{7} (Y - \bar{Y})^2$$

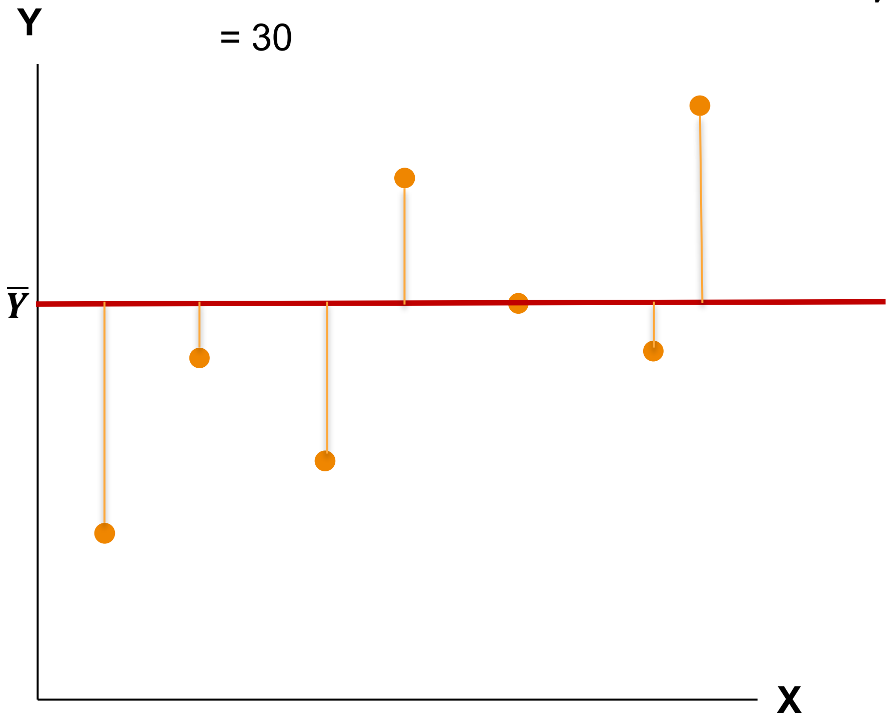
= 30

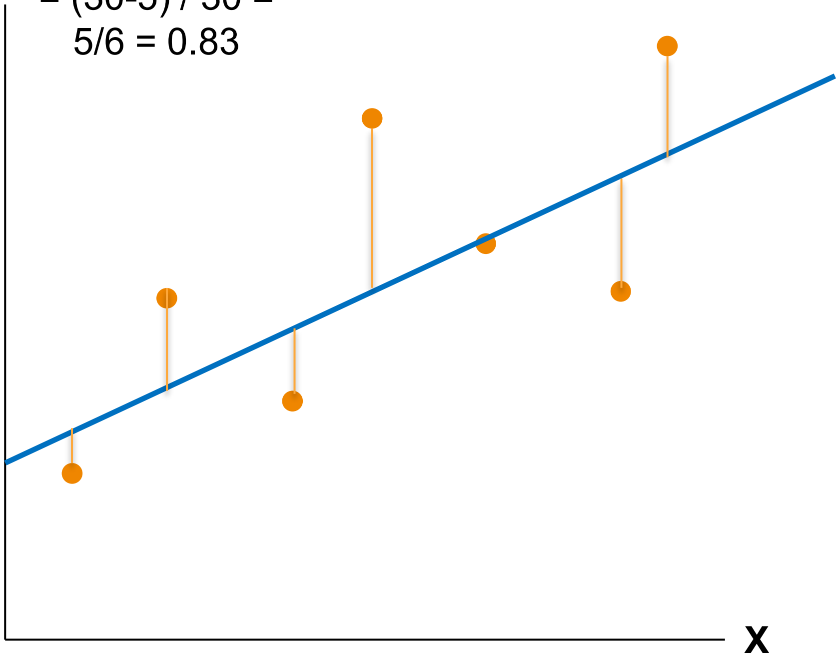
$$R^2 = \frac{Var(Y) - Var(line)}{Var(Y)}$$

= (30-5) / 30 = 5/6 = 0.83

$$Var(line) = \sum_{i=1}^{7} (Y - \hat{Y}_i)^2$$

= 5

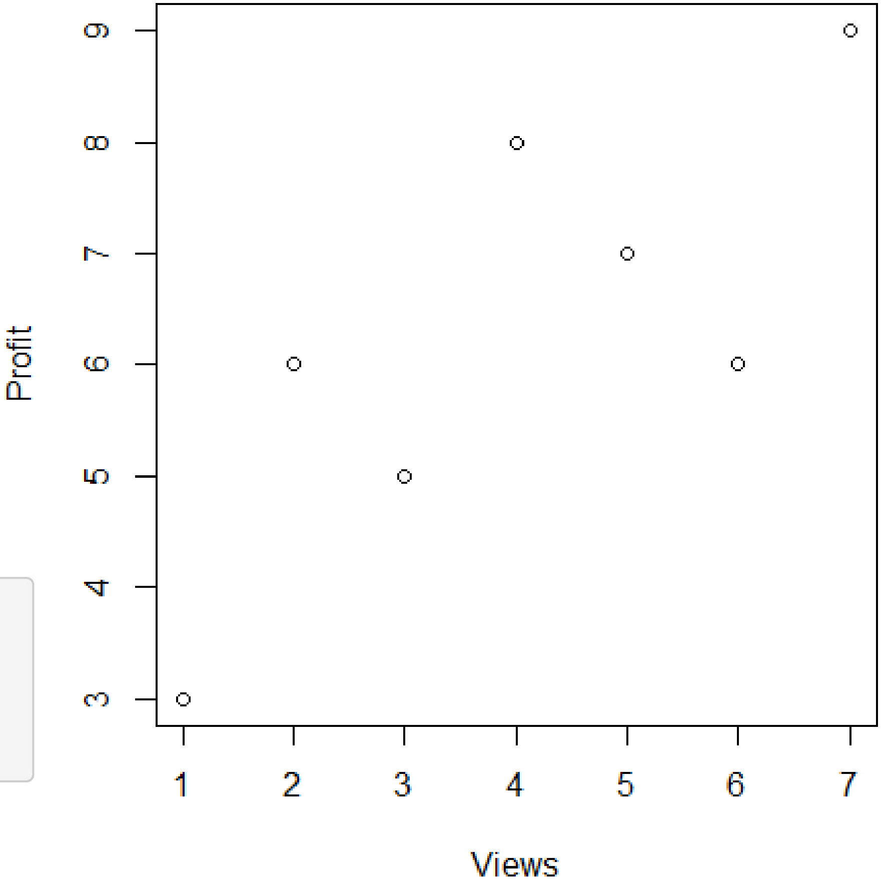




Let's get back to the first example

Video	Views	Profit
1	1	3
2	2	6
3	3	5
4	4	8
5	5	7
6	6	6
7	7	9

```
x=c(1:7)
y=c(3,6,5,8,7,6,9)
par(pty="s")
plot(x,y)
```



Learning model with linear model function (lm)

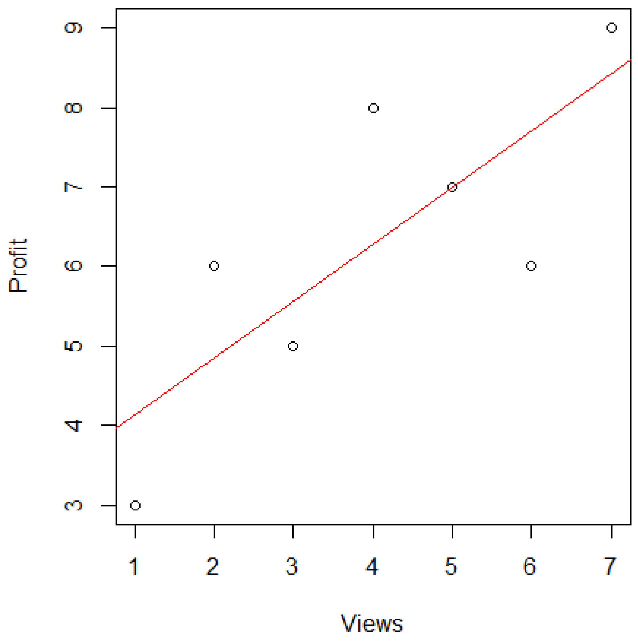
Video	Views	Profit
1	1	3
2	2	6
3	3	5
4	4	8
5	5	7
6	6	6
7	7	9

```
lr.model<-lm(y~x)
summary(lr.model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##          1          2          3          4          5          6          7
## -1.143e+00  1.143e+00 -5.714e-01  1.714e+00  1.110e-16 -1.714e+00  5.714e-01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Interc  α=  3.4286     1.1429   3.000   0.0301 *
## x        β=  0.7143     0.2556   2.795   0.0382 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.352 on 5 degrees of freedom
## Multiple R-squared:  0.6098, Adjusted R-squared:  0.5317
## F-statistic: 7.813 on 1 and 5 DF,  p-value: 0.03821
```

E_i

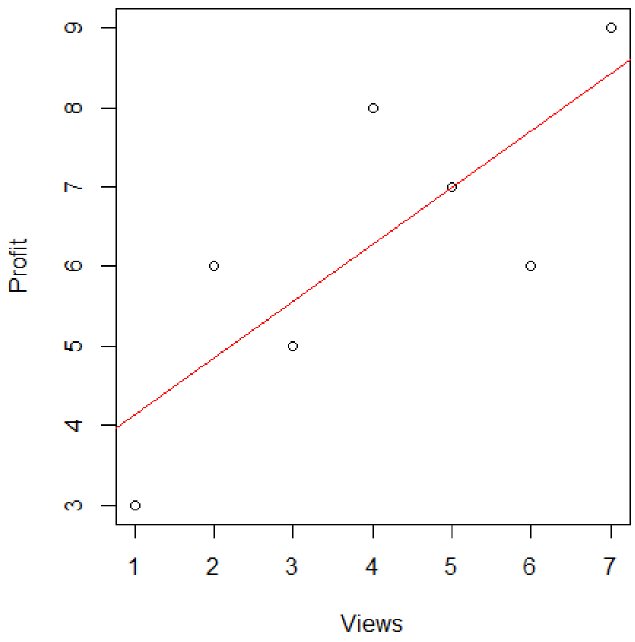
Model interpretation




```
##
## Call:
## lm(formula = y ~ x)
##                                            E_i
## Residuals:
##          1          2          3          4          5          6          7
## -1.143e+00  1.143e+00 -5.714e-01  1.714e+00  1.110e-16 -1.714e+00  5.714e-01
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Interc   α=  3.4286     1.1429   3.000   0.0301 *
## x         β=  0.7143     0.2556   2.795   0.0382 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.352 on 5 degrees of freedom
## Multiple R-squared:  0.6098, Adjusted R-squared:  0.5317
## F-statistic: 7.813 on 1 and 5 DF,  p-value: 0.03821
```

Model: $Y = 3.4 + 0.7X$

Prediction



Views	Profit
8	$3.4 + 0.7 * 8 = 9$
9	$3.4 + 0.7 * 9 = 9.7$
10	$3.4 + 0.7 * 10 = 10.4$

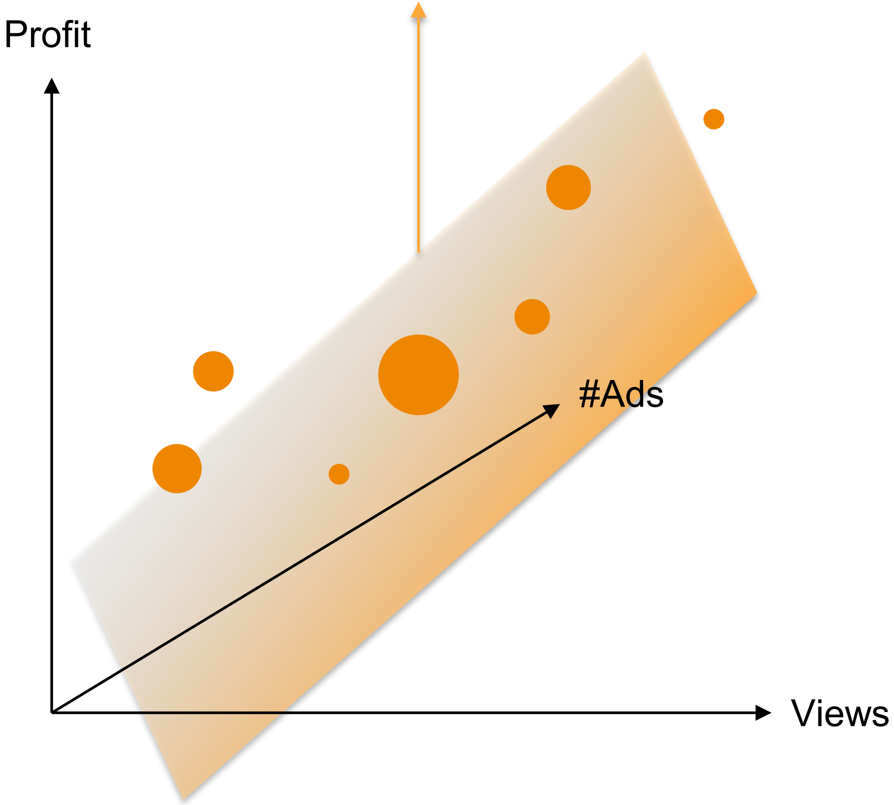
```
new.data=data.frame(x=c(8:10))
predict(lr.model, new.data)
```

```
##         1         2         3
##  9.142857  9.857143 10.571429
```

Model: $Y = 3.4 + 0.7X$

When the number of features is two

$$Profit_i = \alpha + \beta Views_i + \gamma Ads_i + \varepsilon_i$$



Profit of the video ∝ the number of views

Views	Ads	Profit
1	2	4
2	3	5
3	2	5
4	1	4
5	4	8
6	3	6
7	4	10

To minimize the SSE

$$Profit_i = \alpha + \beta Views_i + \gamma Ads_i + \varepsilon_i$$

Objective: Find the line which **minimizes** the sum of squared error

Min(SSE)

$$SSE = \sum_i E_i{}^2 = \sum_i \{Profit_i - (\alpha + \beta Views_i + \gamma Ads_i)\}^2$$

$$\frac{\partial}{\partial \alpha} \sum_i \{Profit_i - (\alpha + \beta Views_i + \gamma Ads_i)\}^2 = 0$$

$$\frac{\partial}{\partial \beta} \sum_i \{Profit_i - (\alpha + \beta Views_i + \gamma Ads_i)\}^2 = 0$$

$$\frac{\partial}{\partial \gamma} \sum_i \{Profit_i - (\alpha + \beta Views_i + \gamma Ads_i)\}^2 = 0$$

$$\Rightarrow \alpha, \beta, \gamma$$

$$\Downarrow$$

$$Profit = \alpha + \beta * View + \gamma * Ads$$

Linear regression Pros & Cons

Pros	Cons
Simple model	Overly-simplistic
Computationally efficient	Linearity assumption
Easy interpretability	Severely-affected by outliers
	Independence of variables
	Assumes Homoskedacity
	Inability to determine Feature importance