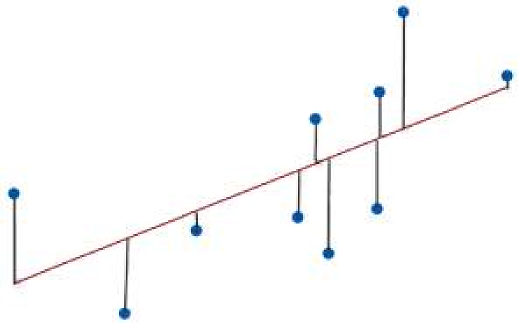


Simple Regression:

Evaluation Metric

Anand Paul

Stock Price Prediction

- How much did stock cost now and in the future ?

Data

input *output*

$(x_1 = \text{day}, y_1 = \$)$

$(x_2 = \text{day}, y_2 = \$)$

$(x_3 = \text{day}, y_3 = \$)$

$(x_4 = \text{day}, y_4 = \$)$

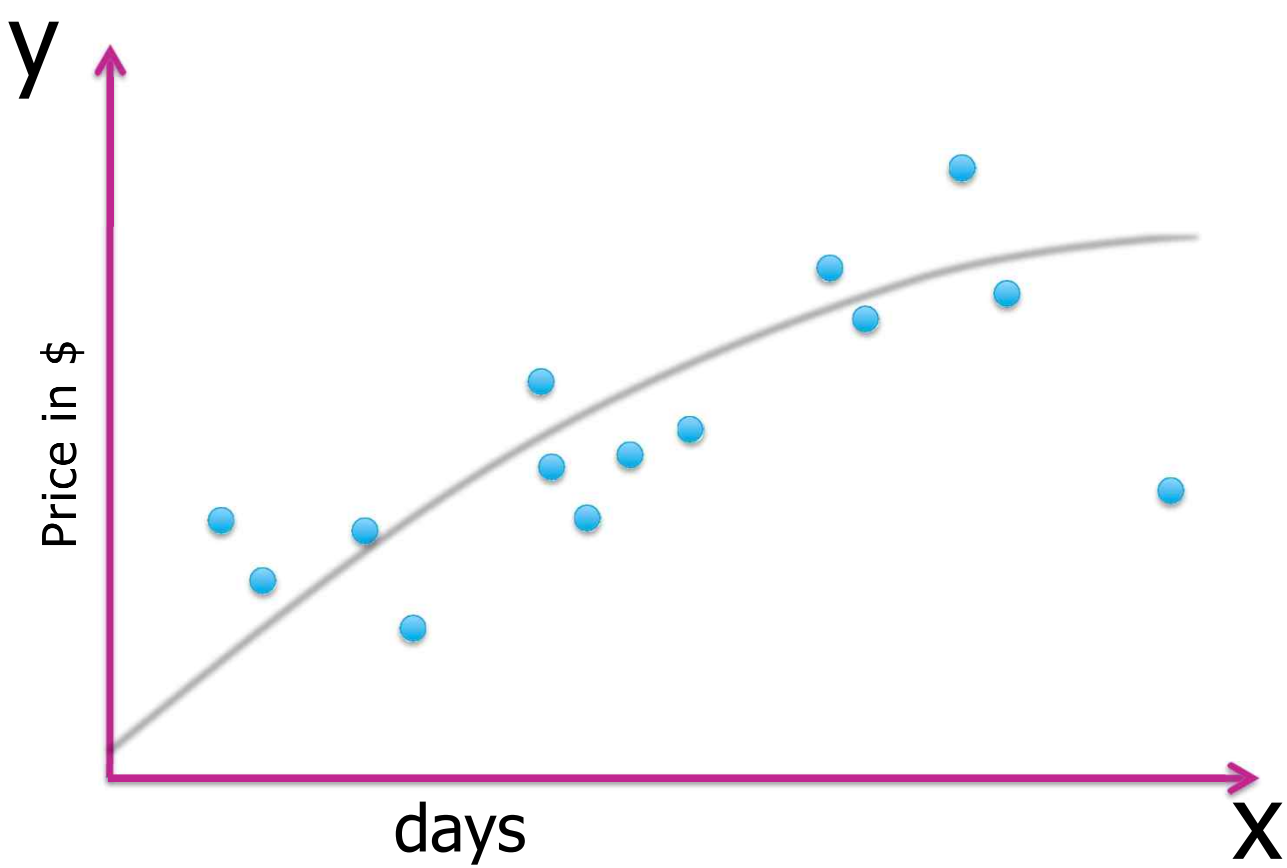
$(x_5 = \text{day}, y_5 = \$)$

⋮

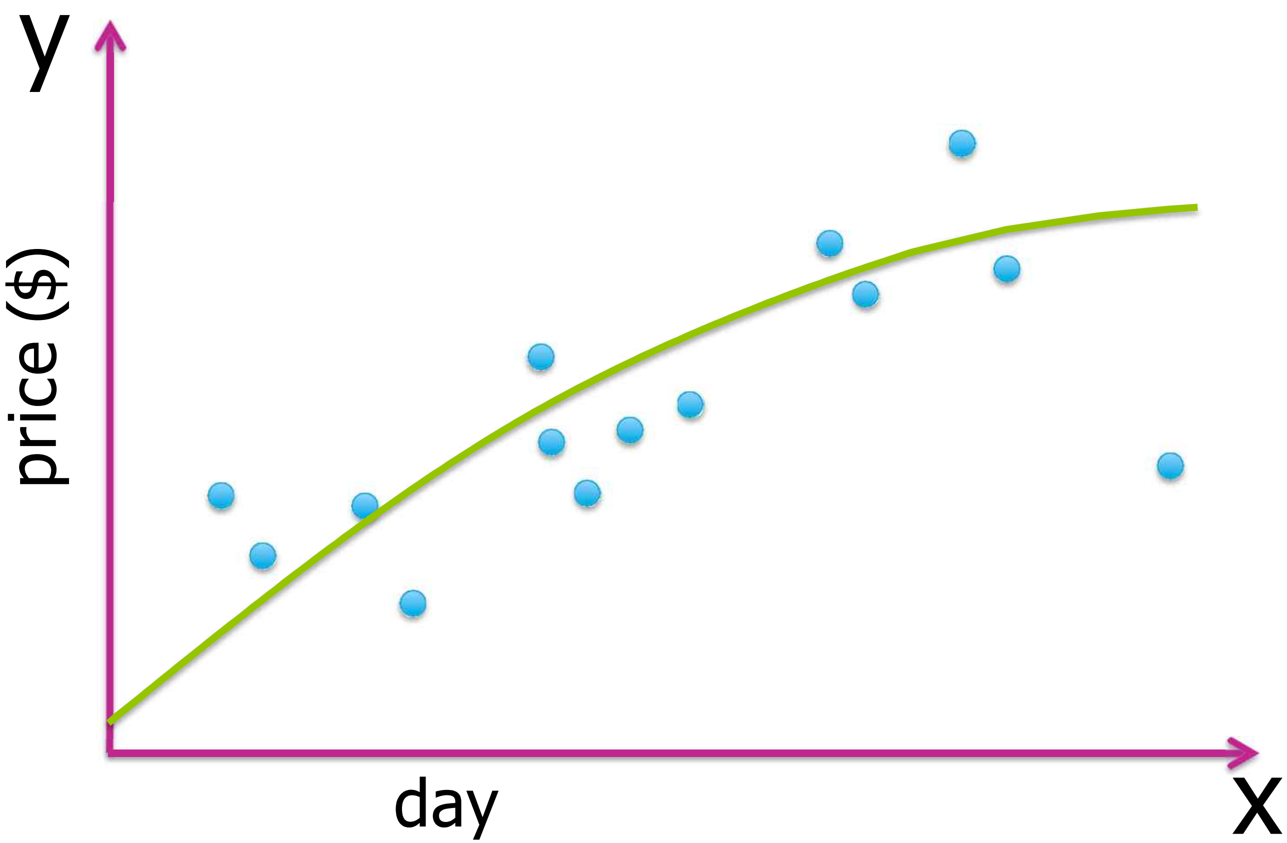
Input vs. Output:

- y is the quantity of interest
- assume y can be predicted from x

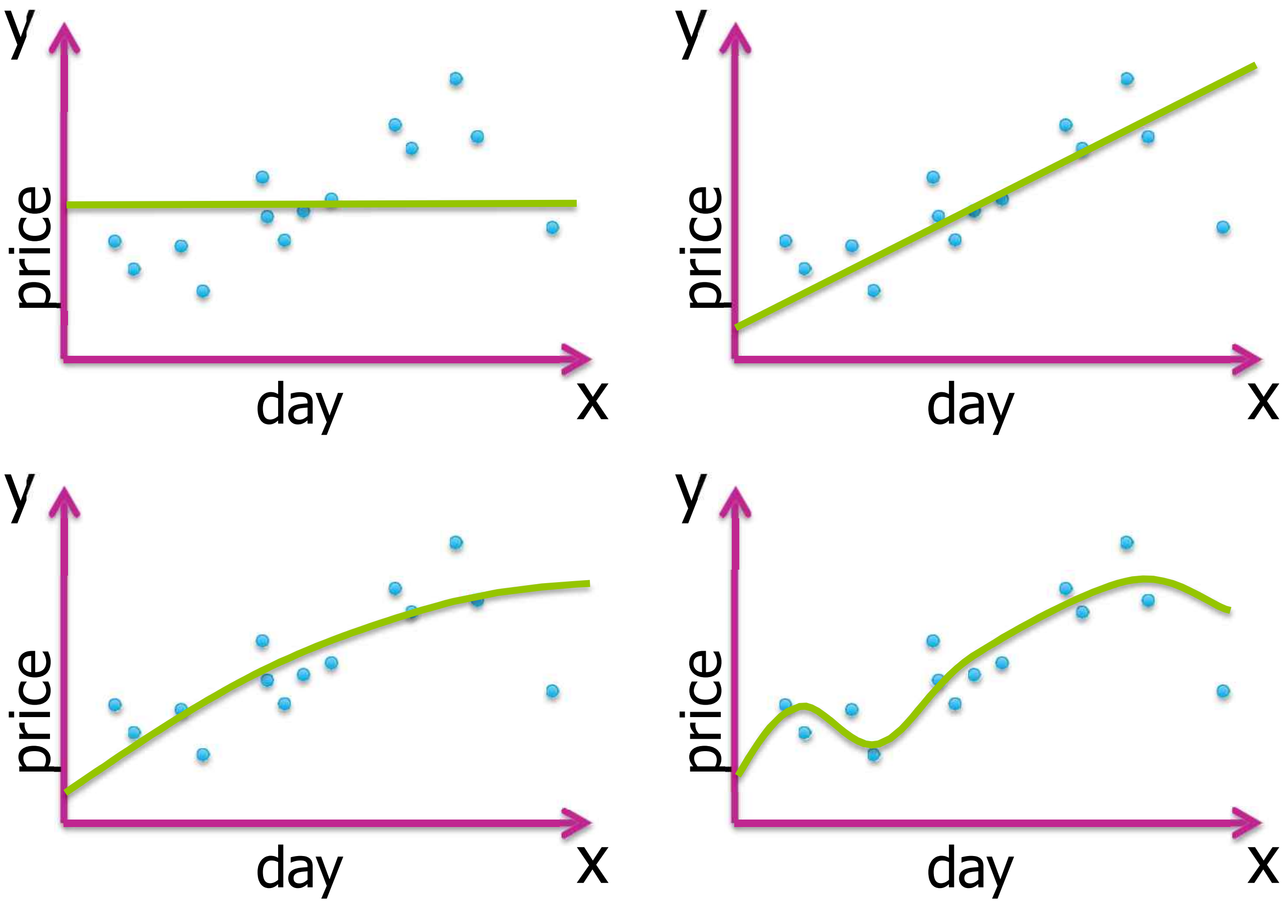
Regression Model –



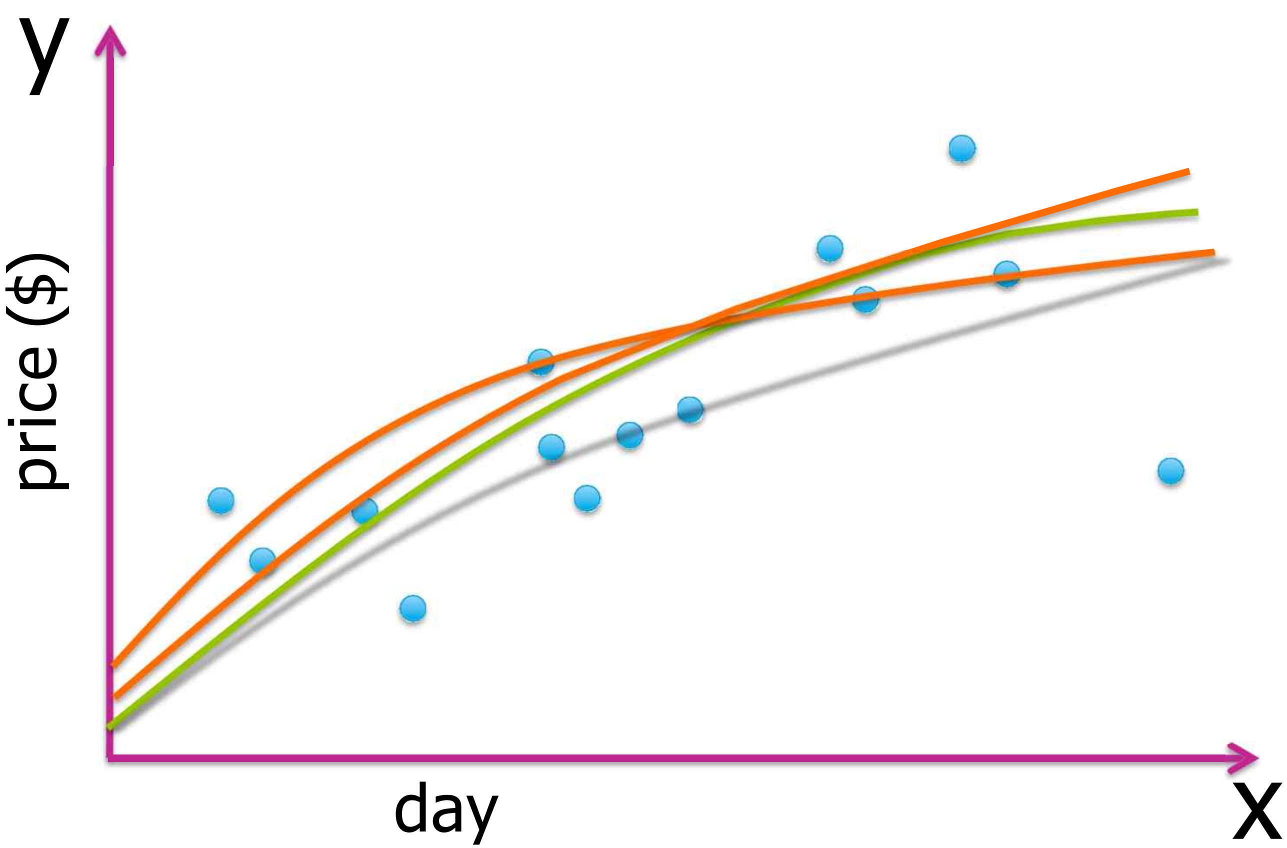
Model –
How we *assume* the world works



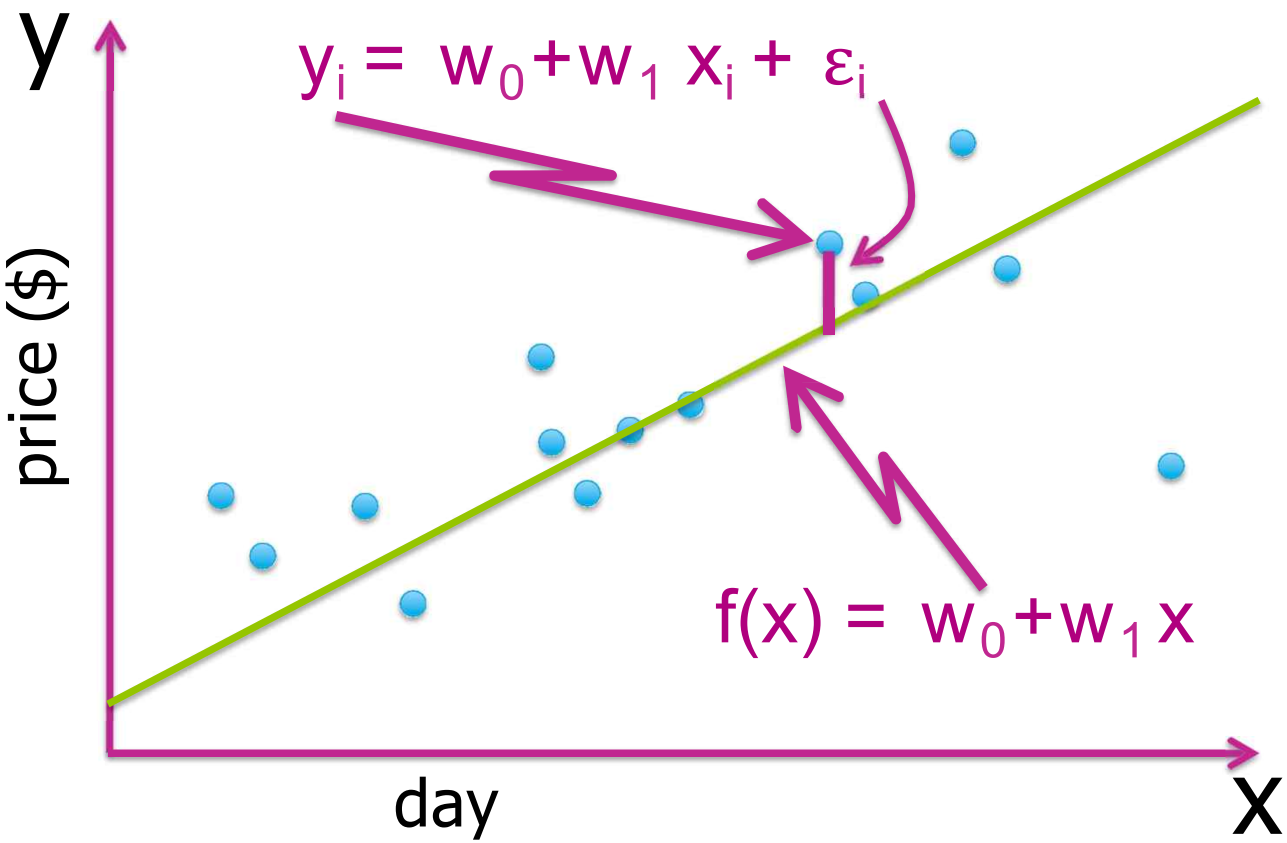
Task 1–
Which model f(x)?



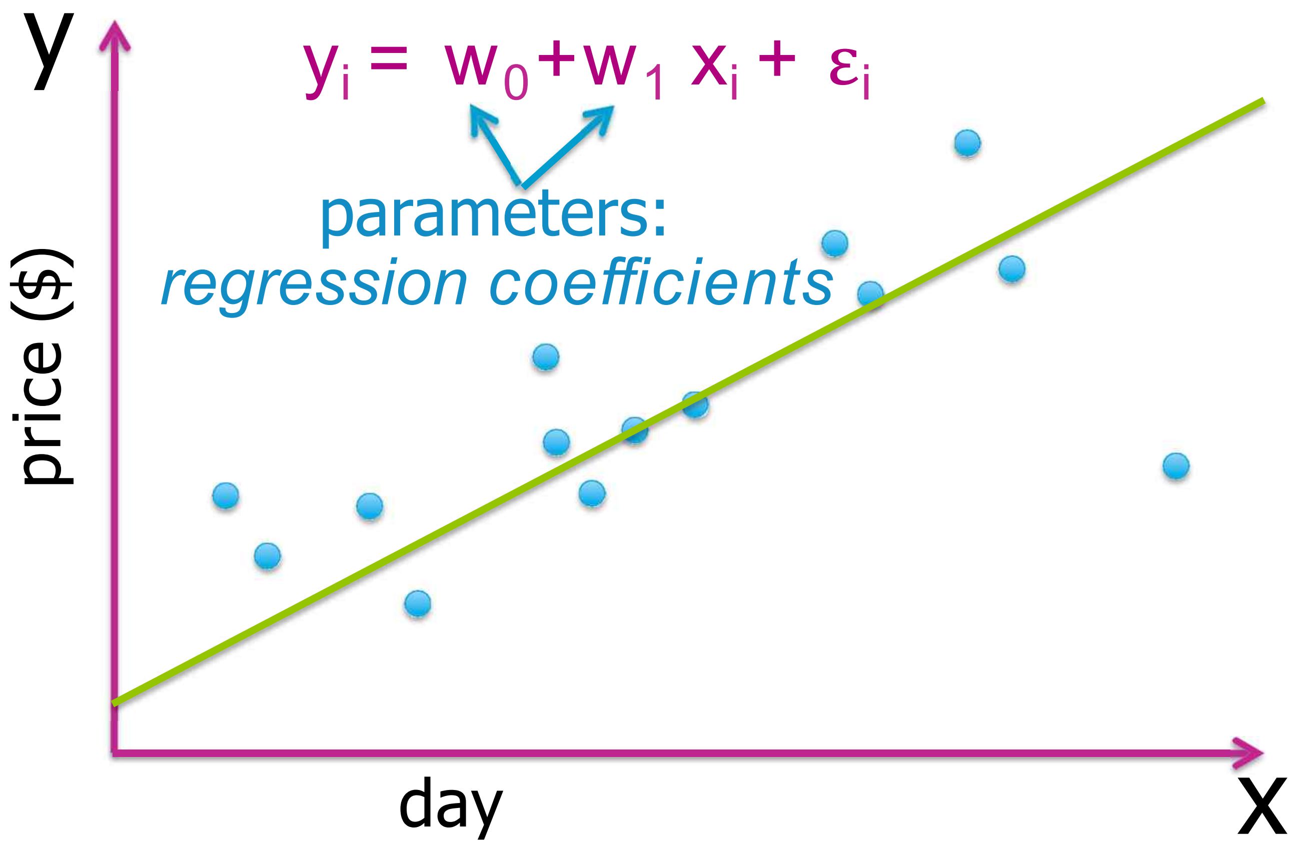
For a given model f(x), estimate function which model fit the data



Simple linear regression model

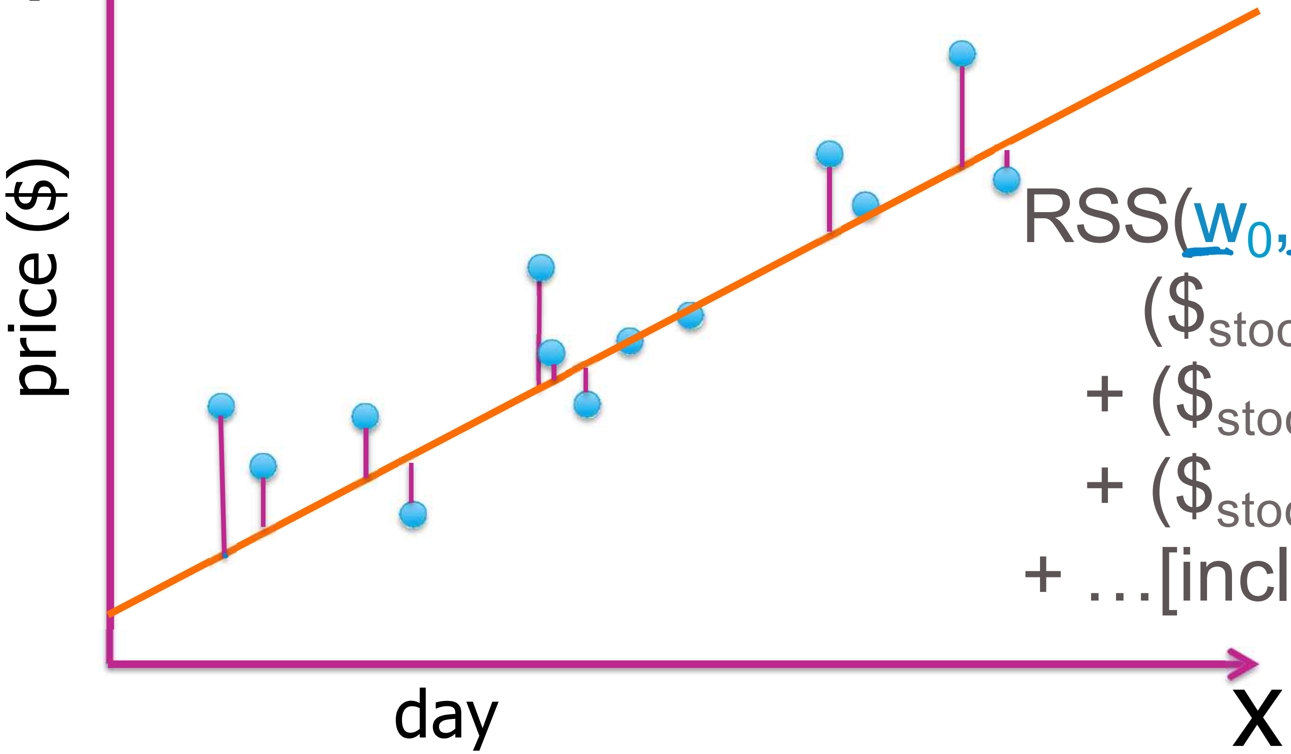


Simple linear regression model



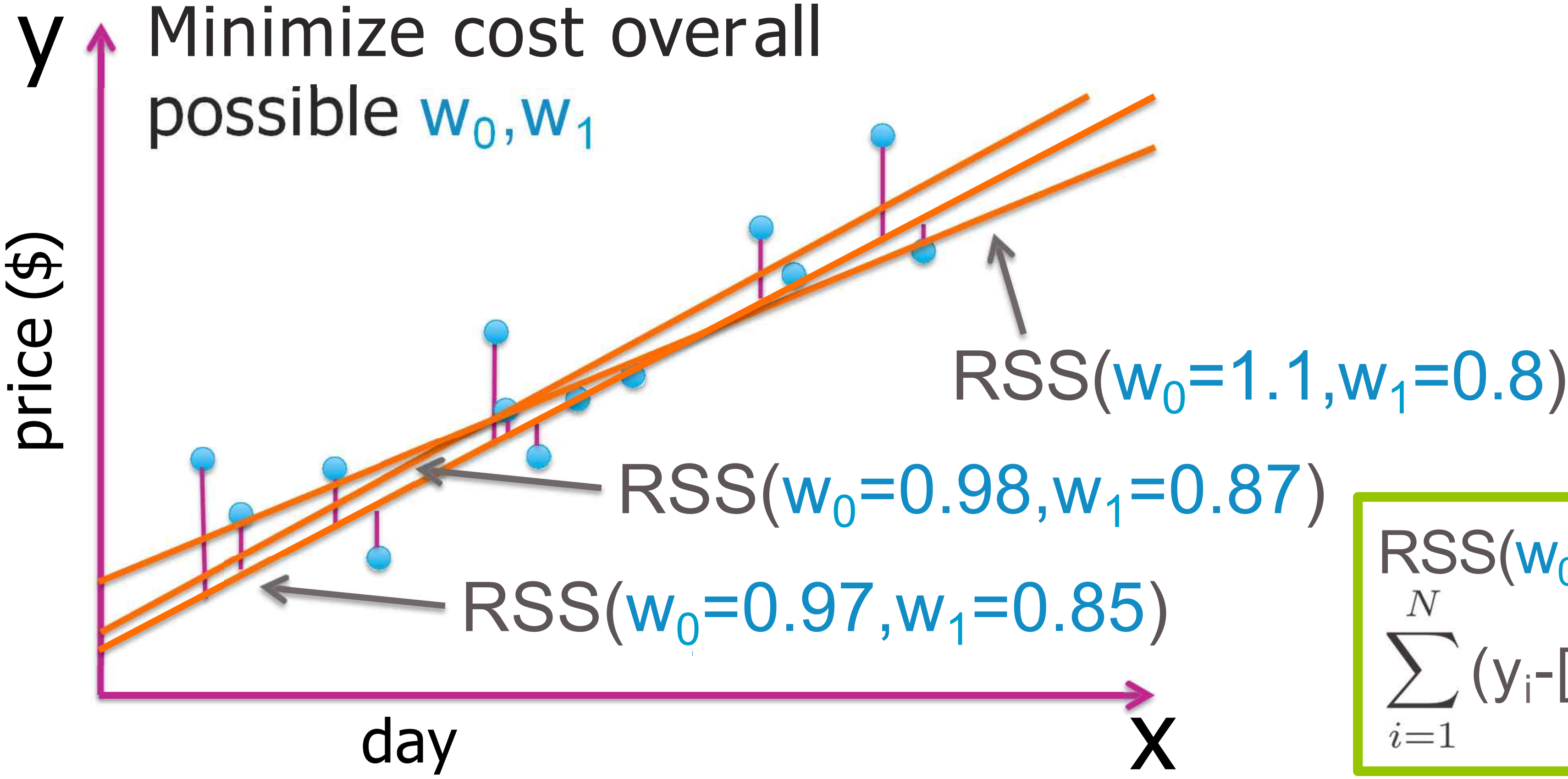
"Cost" of using a given line

Residual sum of squares (RSS)



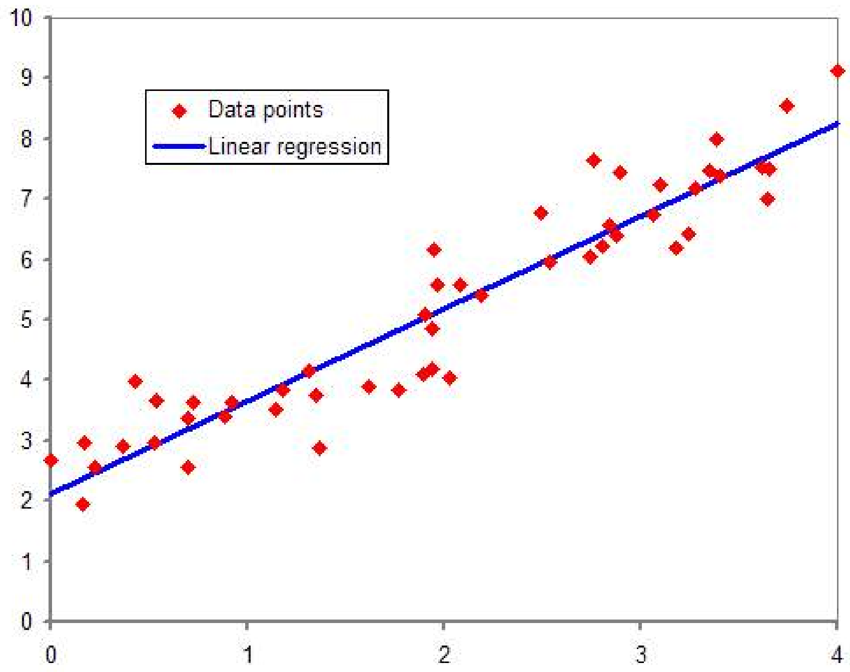
$$\mathrm{RSS}(w_0,w_1) = (\$_{\text{stock }1} - [w_0 + w_1 \mathrm{Day}._{\text{stock }1}])^2 + (\$_{\text{stock }2} - [w_0 + w_1 \mathrm{Day}._{\text{stock }2}])^2 + (\$_{\text{stock }3} - [w_0 + w_1 \mathrm{Day}._{\text{stock }3}])^2 + \ldots \text{[include all training stocks]}$$

Find “best” line



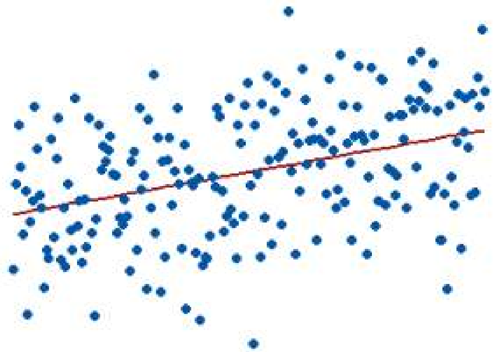
$$\text{RSS}(w_0,w_1) = \sum_{i=1}^{N}(y_i-[w_0+w_1x_i])^2$$

Coffee Consumption Vs Mortality



coefficient of determination (R^2)

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$



- 0% represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model.

- 100% represents a model that explains all of the variation in the response variable around its mean.

Linear Regression Examples

You have travel cost data given below:

Total Miles (X)	Total payment for Gas (Y)
390	36.66
403	37.02
396	34.71
383	32.5
321	31.1
391	34.45
386	36.79
371	37.44
404	38.09
392	38.09

If we drive for 630 miles, how much shall we pay for gas?

Linear Regression Examples

Solution:

Total Miles (X)	Total payment for Gas (Y)
390	36.66
403	37.02
396	34.71
383	32.5
321	31.1
391	34.45
386	36.79
371	37.44
404	38.09
392	38.09

The equation for any straight line can be written as:\
$\hat{Y} = b_0 + b_1X$
where:
b_0 = Y intercept, and
b_1 = regression coefficient = slope of the line

With the data provided, goal is to determine the regression equation

Linear Regression Examples

b1: We can solve b1 with given equation.

$$b_1 = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{\sum XY - \frac{(\sum X \sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

So for our given data we have to find b1.

Linear Regression Examples

Total Miles (X)	Total payment for Gas (Y)	XY	x^2	y^2
390	36.66	14297.4	152100	1343.95
403	37.02	14919.06	162409	1370.48
396	34.71	13769.80	156816	1204.78
383	32.5	12447.5	146689	1056.25
321	31.1	9983.1	103041	967.21
391	34.45	13469.95	152881	1186.80
386	36.79	14200.94	148996	1463.87
371	37.44	13890.24	137641	1401.75
404	38.09	15388.36	163216	1450.84
392	38.09	14931.28	153664	1450.84

$\Sigma X = 3837$

$\Sigma Y = 356.85$

$\Sigma XY = 137297.63$

$\Sigma x^2 = 1477453$

$\Sigma y^2 = 12896.77$

$$b_1 = \frac{\sum XY - \frac{(\sum X \sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

b1= 0.072

Linear Regression Examples

To complete the regression equation, we need to calculate bo.

Total Miles (X)	Total payment for Gas (Y)	XY	x^2	y^2
390	36.66	14297.4	152100	1343.95
403	37.02	14919.06	162409	1370.48
396	34.71	13769.80	156816	1204.78
383	32.5	12447.5	146689	1056.25
321	31.1	9983.1	103041	967.21
391	34.45	13469.95	152881	1186.80
386	36.79	14200.94	148996	1463.87
371	37.44	13890.24	137641	1401.75
404	38.09	15388.36	163216	1450.84
392	38.09	14931.28	153664	1450.84

bo = $\bar{y}$ - b $\bar{x}$

bo= 35.68 – 27.62

bo = 8.05

So our regression equation is

Y^ = 8.05 + 0.072X

= 8.05 + 0.072 (630)

= 53.41

Linear Regression Examples

Example #2:

Follow the very simple data and :

x	y
1	2
2	4
3	5
4	4
5	5

Find regression line and also find estimation error.

Linear Regression Examples

Solution:

x	y	x^2	xy
1	2	1	2
2	4	4	8
3	5	9	15
4	4	16	16
5	5	25	25
Σx: 15	**Σy: 20**	**Σx^2: 55**	**Σxy: 66**

The equation for any straight line can be written as:
$\hat{Y} = b_0 + b_1 X$
where:
b_0 = Y intercept, and
b_1 = regression coefficient = slope of the line

With the data provided, Goal is to determine the regression equation

Linear Regression Examples

Let us find the best m (slope)

x	y	x^2	xy
1	2	1	2
2	4	4	8
3	5	9	15
4	4	16	16
5	5	25	25
Σx: 15	**Σy: 20**	**Σx^2: 55**	**Σxy: 66**

Step 1
For each (x,y) calculate x^2 and xy

Step 2
Sum x,y,x^2 and xy (Gives us **Σx, Σy, Σx^2, Σxy**)

Step 3
Calculate slope by using this formula

$$b_1 = \frac{\sum XY - \frac{(\sum X \sum Y)}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}}$$

b_1= 0.6

Linear Regression Examples

b (y-intercept) that suits the data

x	y	x^2	xy
1	2	1	2
2	4	4	8
3	5	9	15
4	4	16	16
5	5	25	25
Σx: 15	**Σy: 20**	**Σx^2: 55**	**Σxy: 66**

Calculate y-intercept by using this formula

bo = $\bar{y}$ - b $\bar{x}$

$\bar{x}$ = 3 and $\bar{y}$ = 4

b_o **= 4 – 0.6 * 3**

= 2.2

So, our regression equation is

$\hat{Y}$ = 2.2 + 0.6X

Linear Regression Examples

Let's find **Error**:

x	y	Estimated y	Est y – Actual y	(Est y – Act y)²
1	2	2.8	0.8	0.64
2	4	3.4	- 0.6	0.36
3	5	4	- 1	1
4	4	4.6	0.6	0.36
5	5	5.2	0.2	0.04
				Σ: 2.4

Error formula

$$\sqrt{\frac{\Sigma(\text{Est y} - \text{Actual y})2}{n-2}}$$

$$= \sqrt{\frac{2.4}{5-2}} = \sqrt{\frac{2.4}{3}}$$

$$= \sqrt{0.8} = 0.89$$