Regression Models

Richard Lui

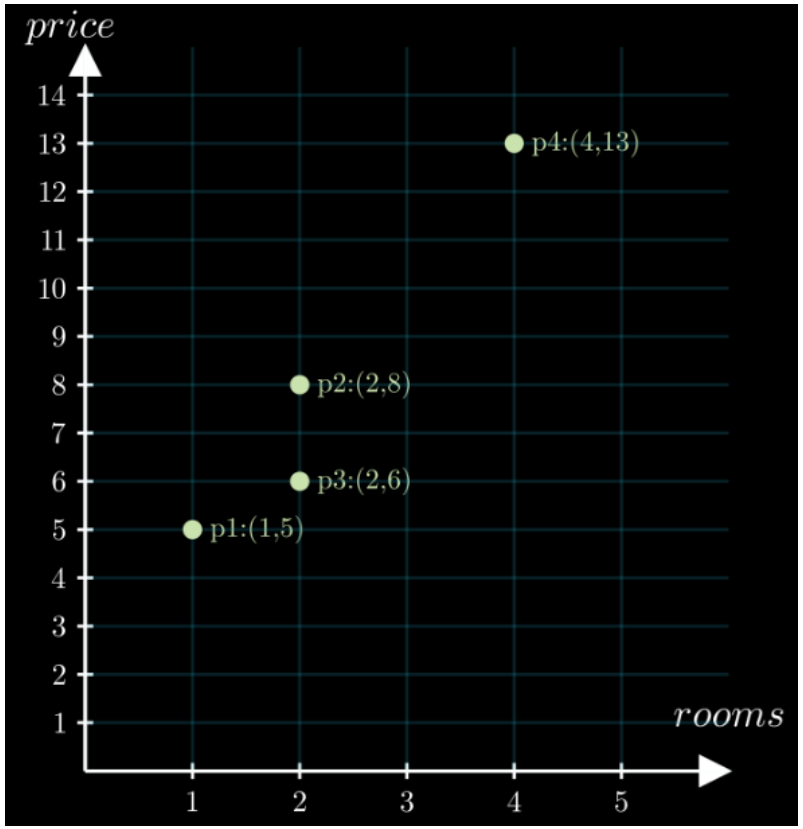
Overview of Regression Analysis

- Researchers are interested in the relationship between variables, e.g.
  - Do people with more education tend to have higher income?
  - What's the relationship between the number of rooms and the price of houses?
- In linear regression, we want to fit a straight line through a set of observations.

Simple linear regression

Rooms	1	2	2	4
Price	5	8	6	13

- Simple linear regression only considers one independent variable (X) and dependent variable (Y).
- E.g. The housing price data set
  - X: the size of the house (independent variable/predictor)
  - Y: the housing price (the dependent variable)
    - *Remark: The price is in million dollars*

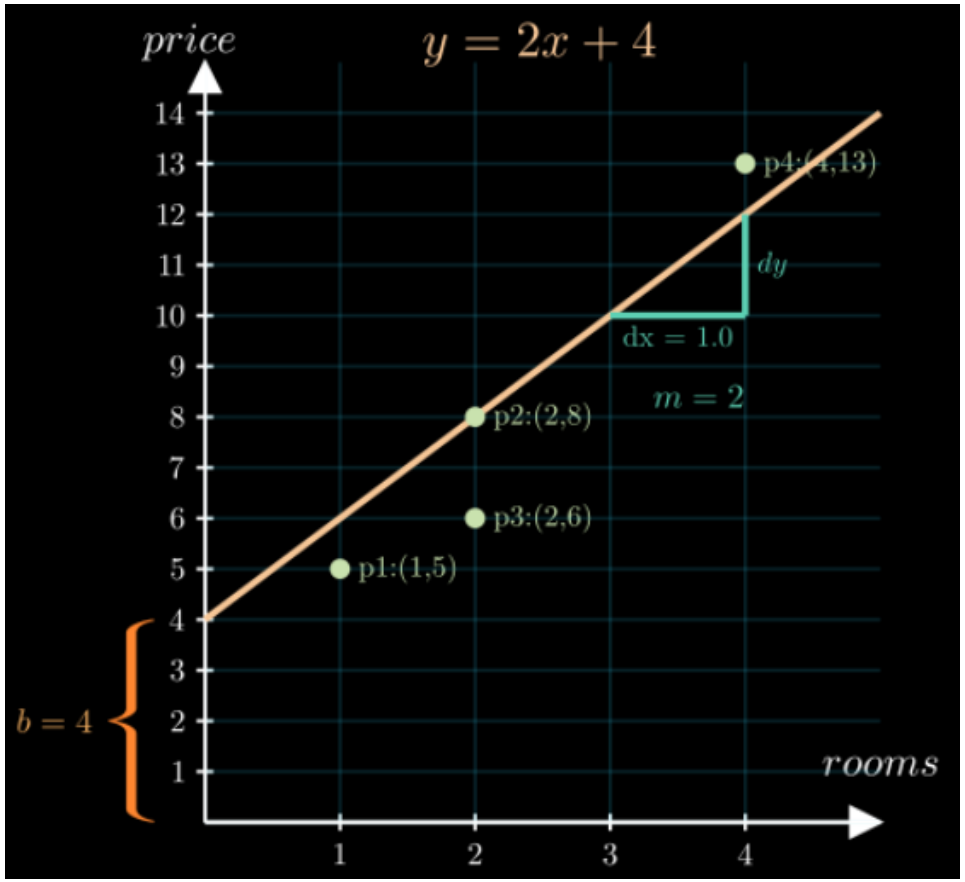


Fitting a straight line

$$Y = b + mX$$

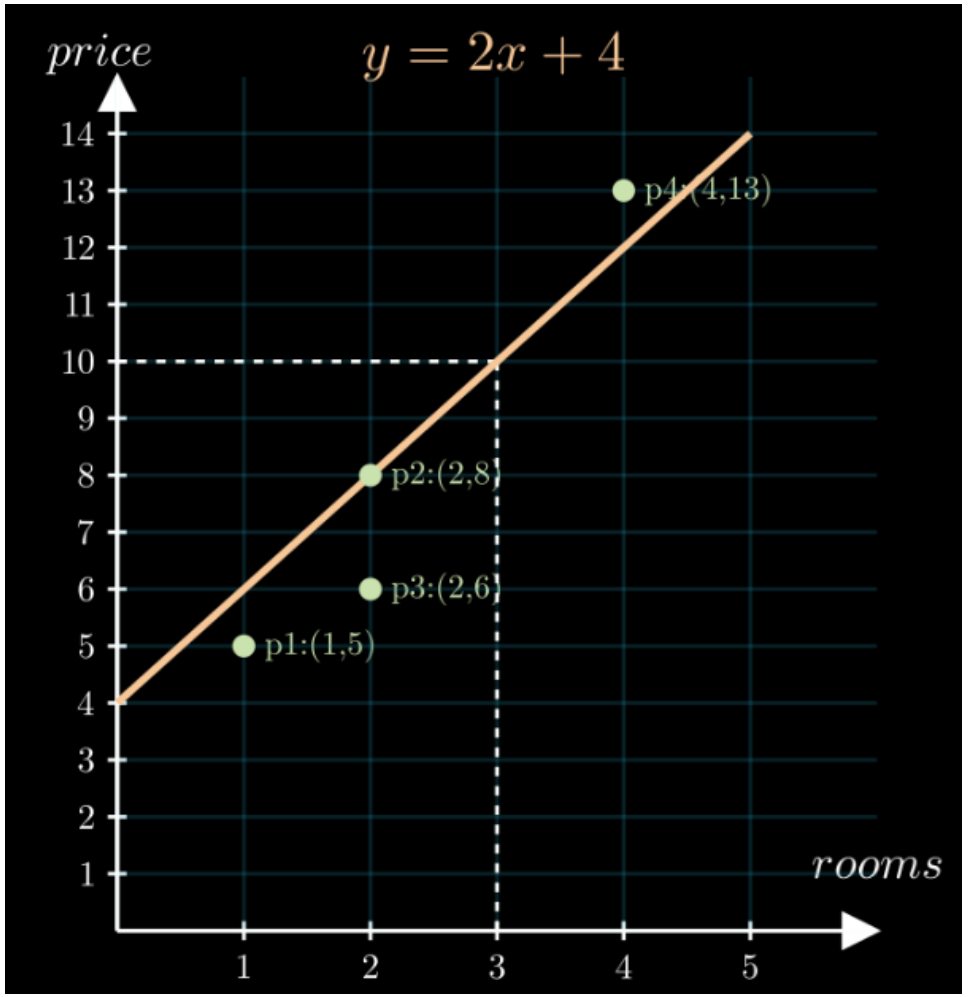
- The *y*-intercept **b** tells us where the line crosses the *y*-axis (the vertical axis)
- The slope **m** tells us how steep the line is

When the slope is 2, it means that when we walk along this line, for every unit that we move to the right, we are moving 2 units up.



House Price Prediction

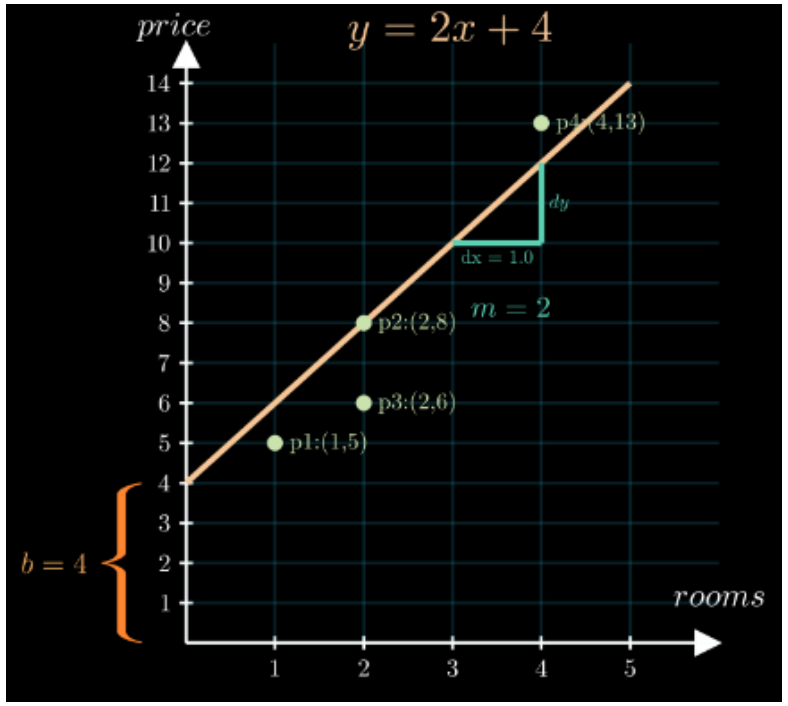
- What is the predicted price of houses with 3 rooms?
- What is the predicted price of houses with 5 rooms?
- How good is the model?

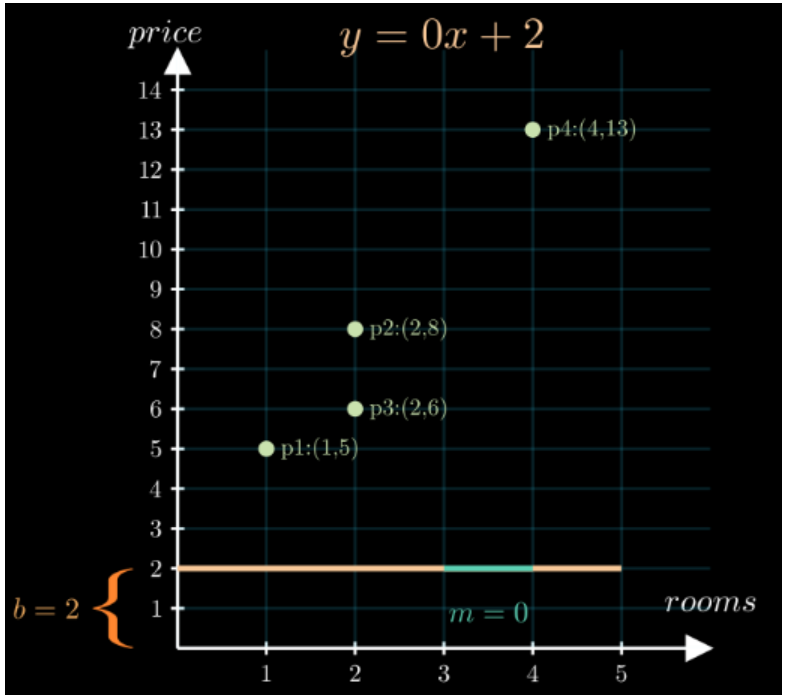


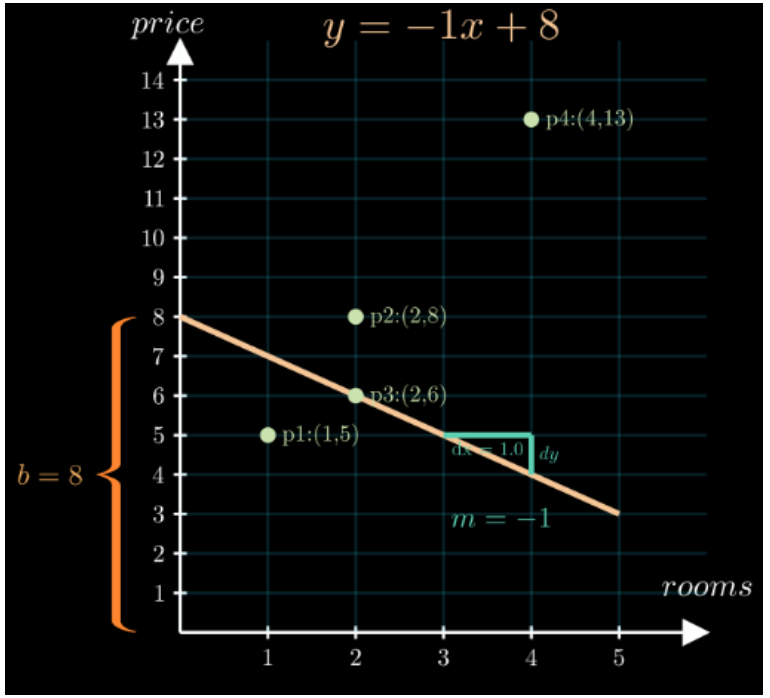
Which model is better?

$$Y = \boldsymbol{b} + \boldsymbol{m}X$$

- We can create different models with different values of b and m.

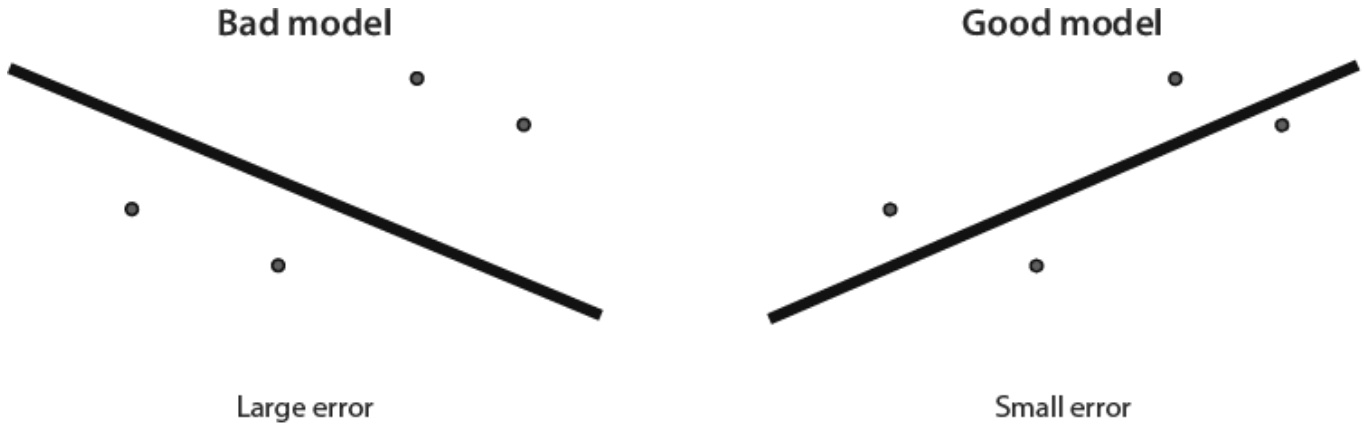






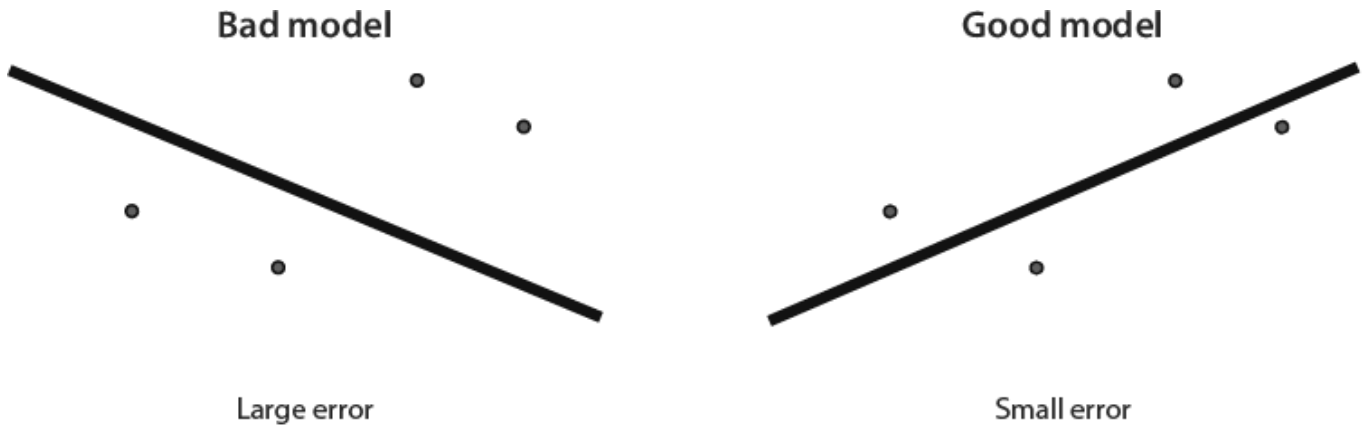
How good is a model?

- A good linear regression model is one where the line is close to the points.
- What does "*close*" means?



Error functions

- An error function is a metric that tells us how our model is doing.
- Sometimes called ***loss functions*** or ***cost functions***
- Assign a large value to the bad model on the left and a small value to the good model on the right.

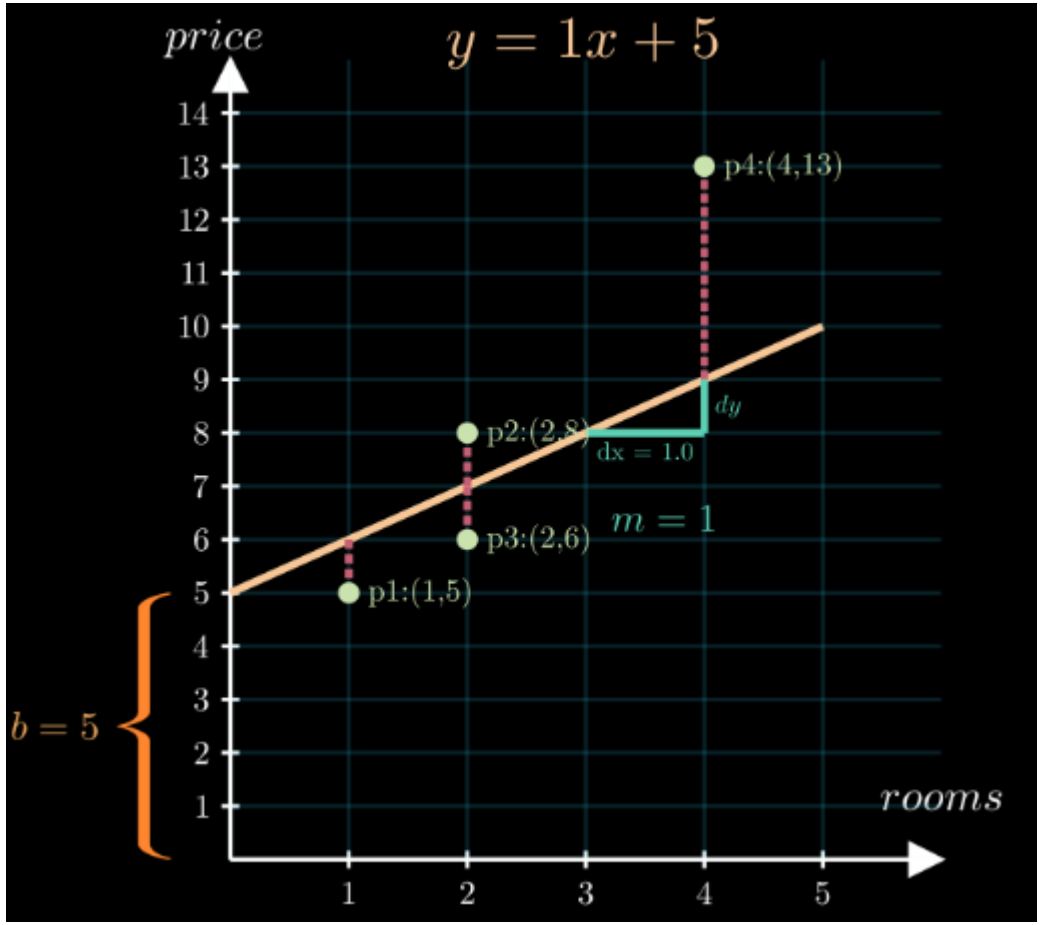


Residuals

- The residual for an observation is the difference between the observed value and predicted value.

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

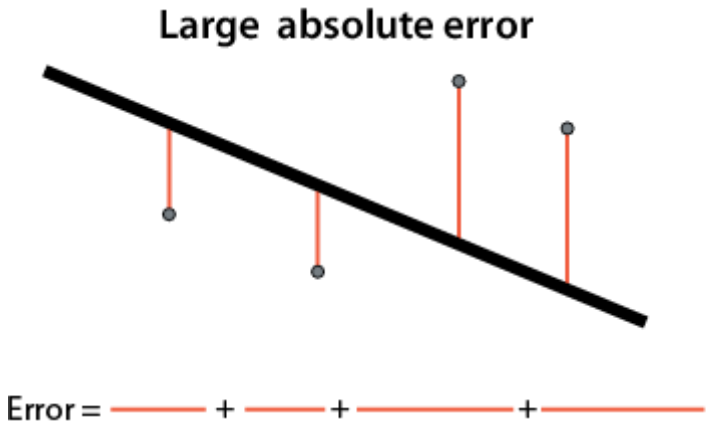
$$e_i = y_i - (\boldsymbol{m}x_i + \boldsymbol{b})$$

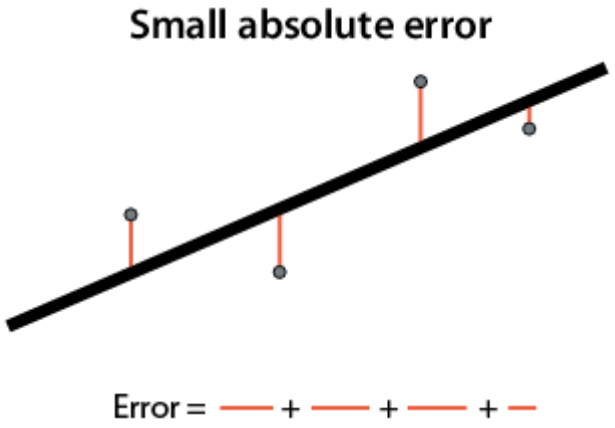


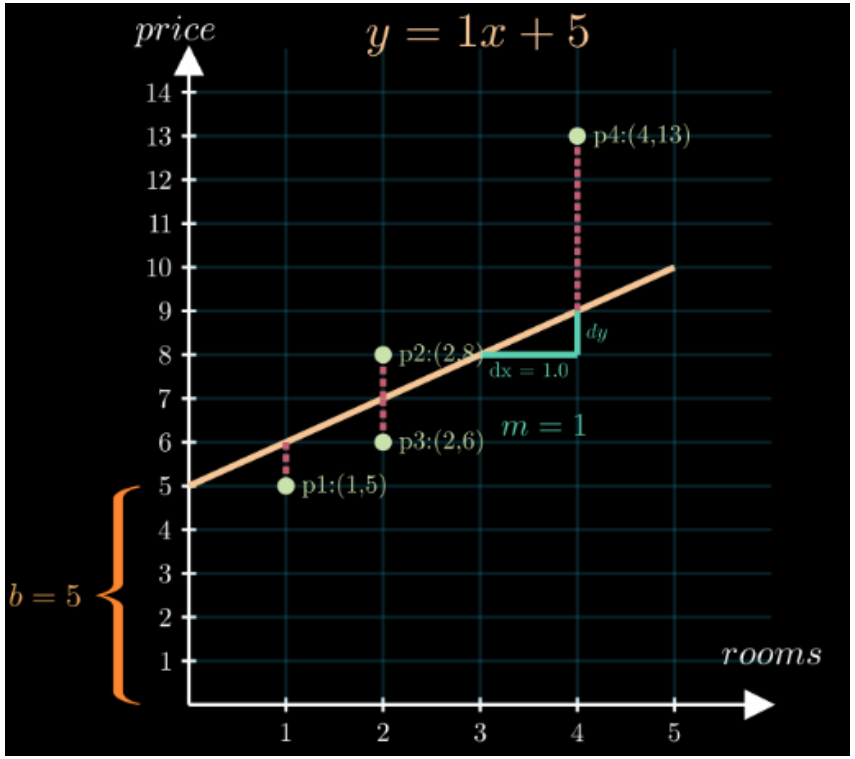
What is the residual for p1, p2, p3 and p4?

Absolute error

- A metric that tells us how good our model is by adding distances between predicted and actual values of the dependent variable

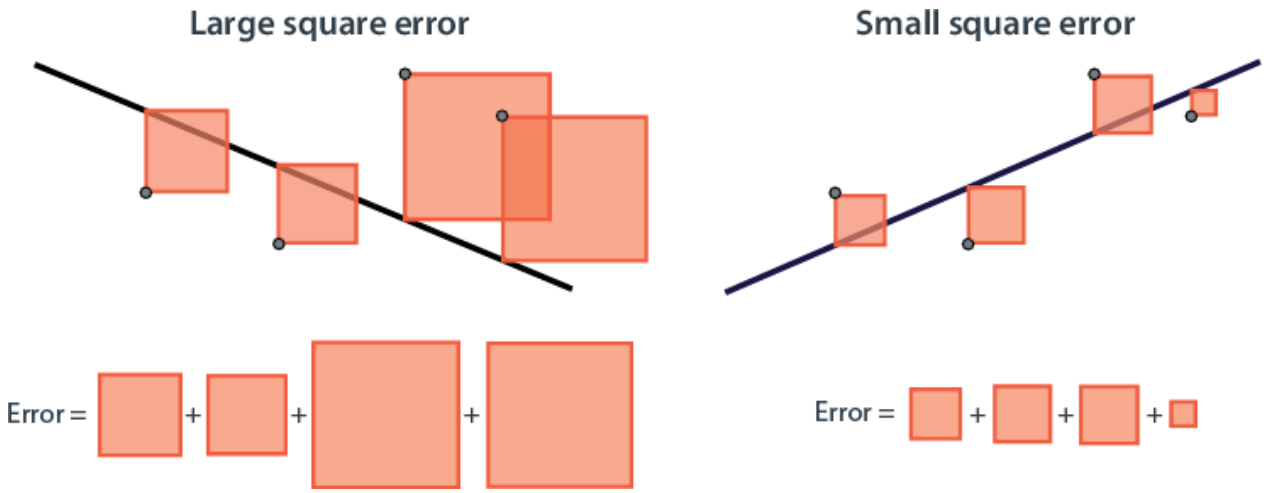






Square Error

- The ***square error*** is a metric that tells us how good our model is by adding squares of residuals
- The goal of linear regression is to fit a straight line through the points so as to *minimize* the sum of the square of the residuals (difference between the observed and predicted values of Y).



Exercise: Square Errors

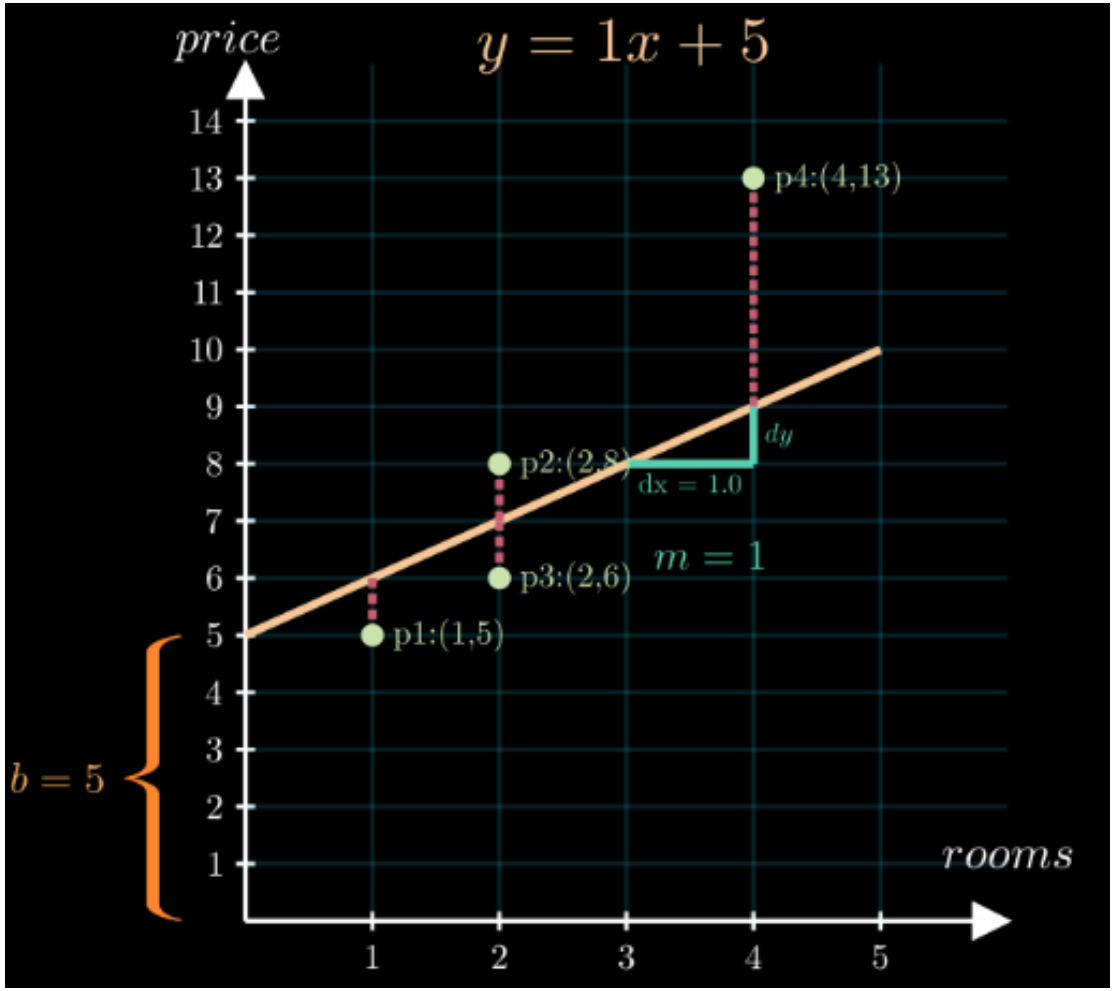
$$e_i = y_i - \hat{y}_i$$

$p1$: $(5-6)^2 = 1$

$p2$: _____

$p3$: _____

$p4$: _____



Error functions

- **Sum of Squared Estimate of Errors (SSE)**
  - the sum of the squares of residuals (deviations of the predicted from actual values of data)

Sum over all examples

Actual

predicted

$$SSE = \sum_{i=1}^{N} \left(y_i - (mx_i + b)\right)^2$$

- **Mean Square Error (MSE)**
  - the average of the squares of these distances.

$$\mathrm{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - (mx_i + b)\right)^2$$

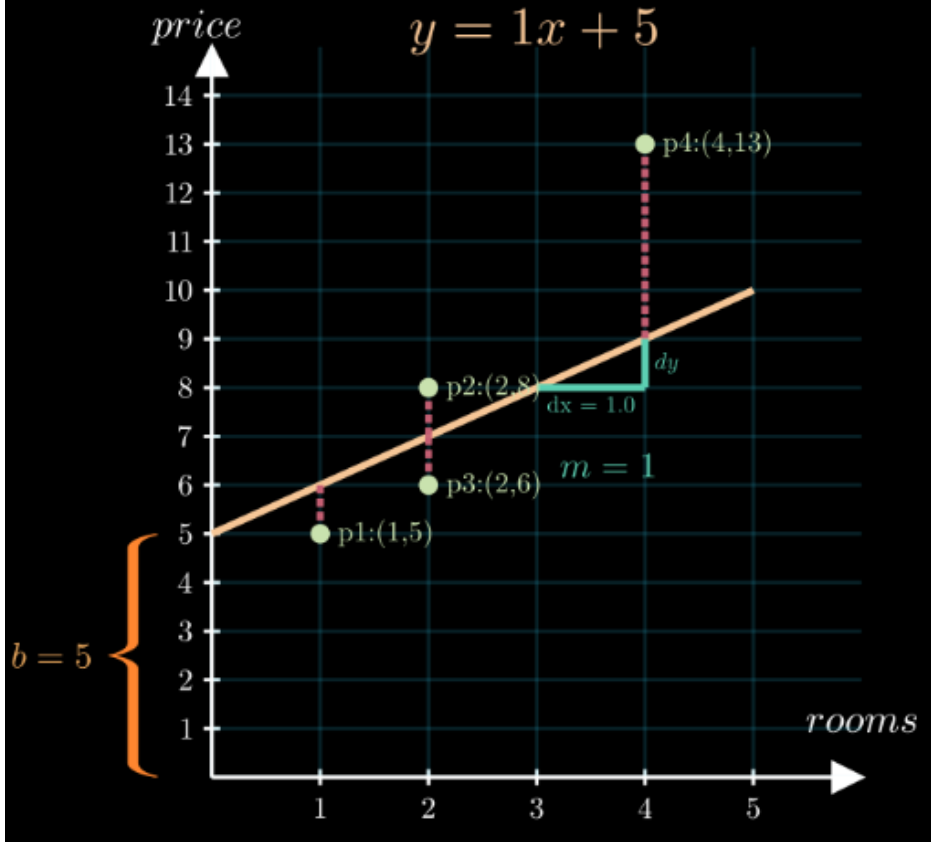
- **Root Mean Square Error (RMSE)**

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_i - (mx_i + b)\right)^2}$$

Example: SSE

$$SSE = \sum_{i=1}^{N} \left(y_i - (mx_i + b)\right)^2$$

Rooms	1	2	2	4
Price	5	8	6	13



$$SSE = \begin{matrix} (y_1 - (mx_1 + b))^2 \\ + \ (y_2 - (mx_2 + b))^2 \\ + \ (y_3 - (mx_3 + b))^2 \\ + \ (y_4 - (mx_4 + b))^2 \end{matrix} = \begin{matrix} (5 - (1(1) + 5))^2 & \Rightarrow 1 \\ + \ (8 - (1(2) + 5))^2 & \Rightarrow 1 \\ + \ (6 - (1(2) + 5))^2 & \Rightarrow 1 \\ + \ (13 - (1(4) + 5))^2 & \Rightarrow 16 \end{matrix} = 19$$

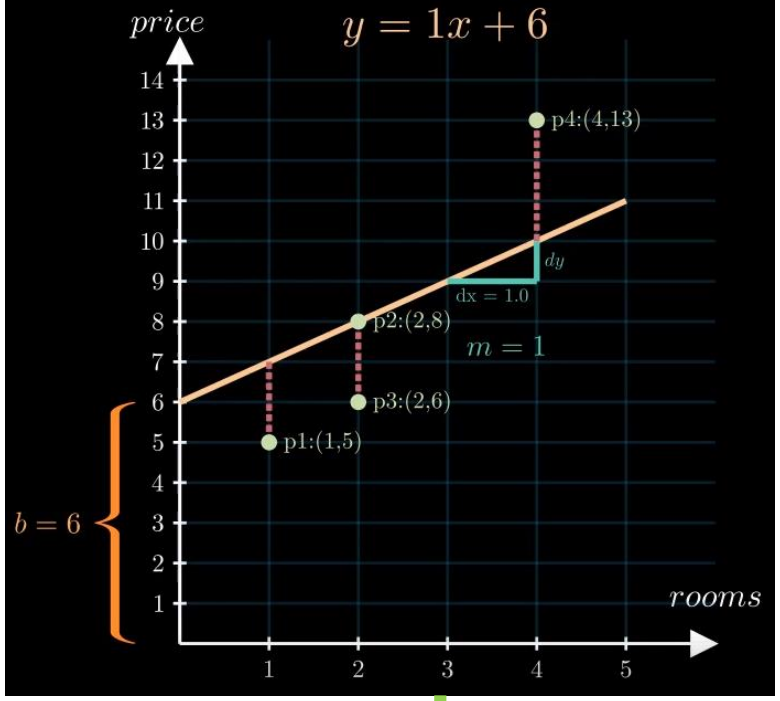
Is there a better model?

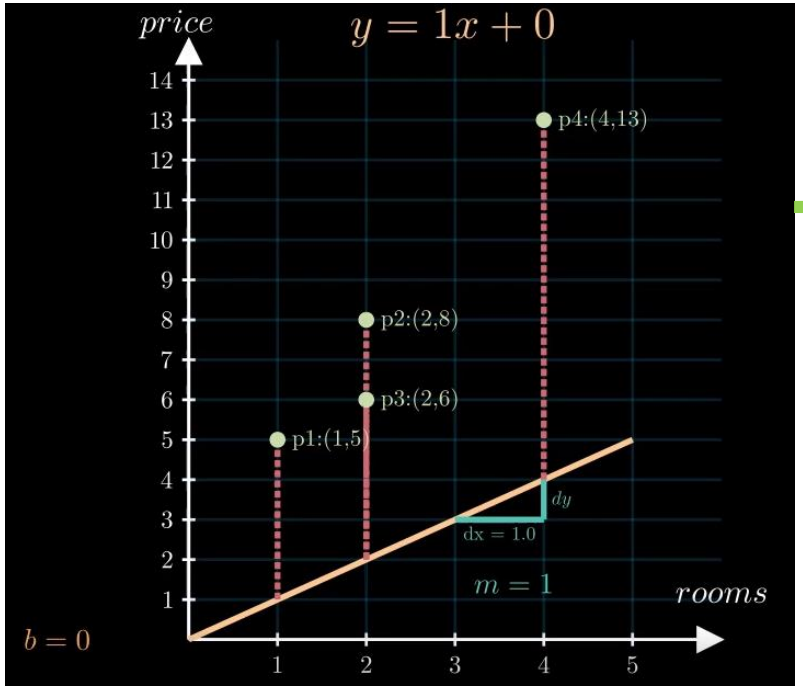
Which model is better?

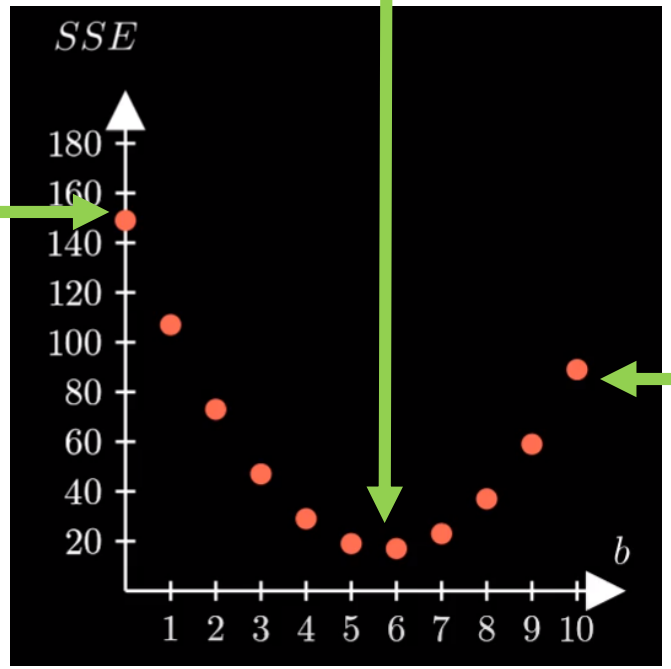
$$y = 1x + b$$

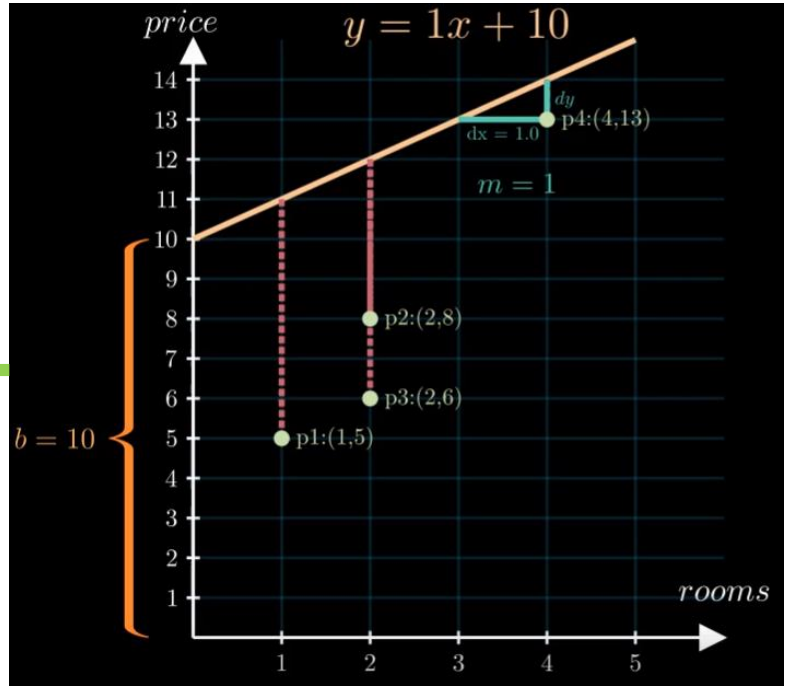
- With m=1, compute SSE with b ={1,2, 3, 4, 6}.
- Plot the SSE for the different b values. Find the best model.

$$y = 1x + b$$


price
$y = 1x + 6$
p4:(4,13)
p2:(2,8)
p3:(2,6)
p1:(1,5)
dy
dx = 1.0
$m = 1$
$b = 6$
rooms


price
$y = 1x + 0$
p4:(4,13)
p2:(2,8)
p3:(2,6)
p1:(1,5)
dy
dx = 1.0
$m = 1$
$b = 0$
rooms

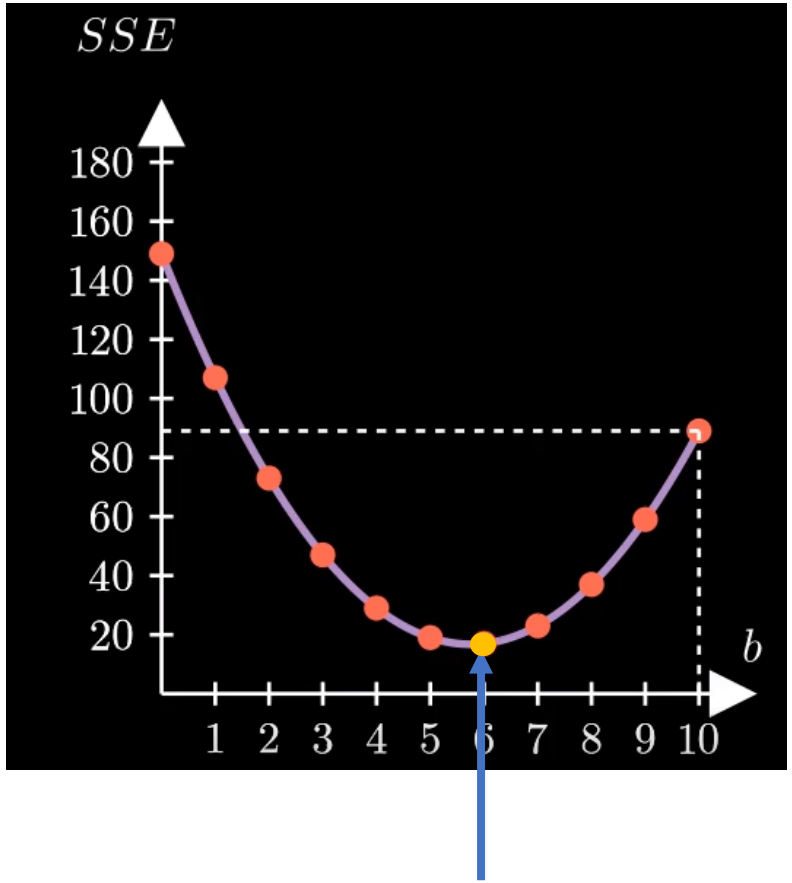

SSE
b


price
$y = 1x + 10$
dy
dx = 1.0
p4:(4,13)
$m = 1$
p2:(2,8)
p3:(2,6)
p1:(1,5)
$b = 10$
rooms

What is the best model?

- When m=1, what is the b which minimizes the SSE?

$$y = 1x + b$$

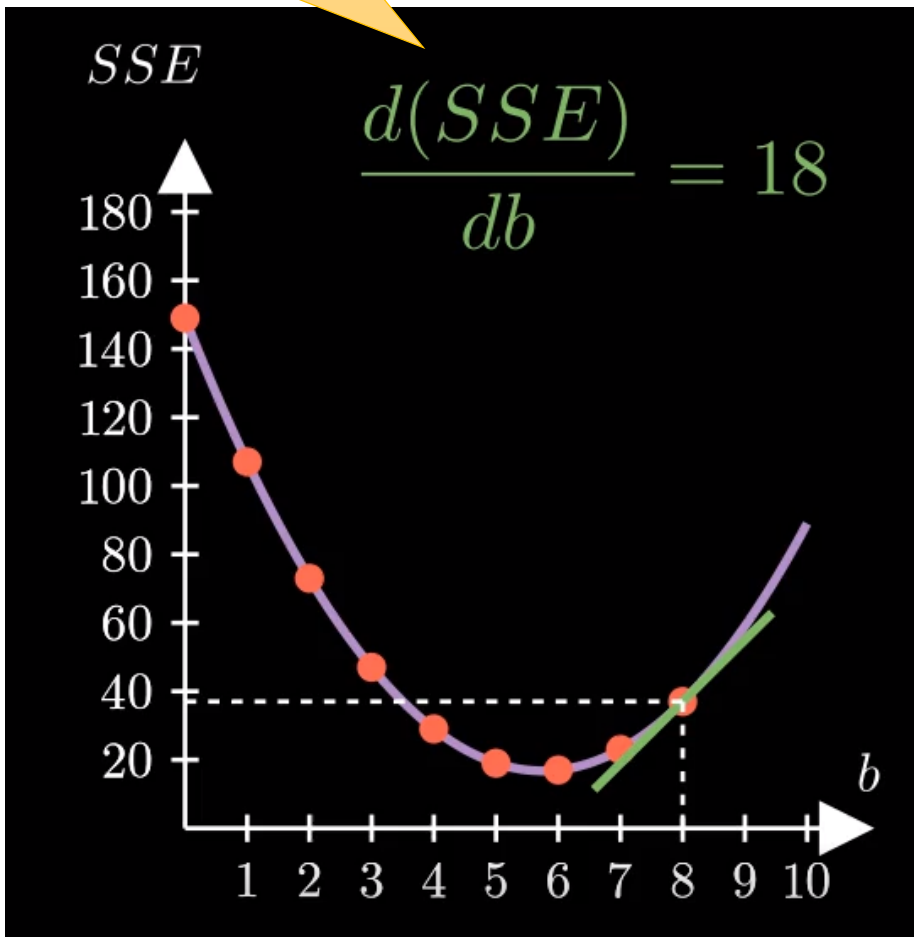


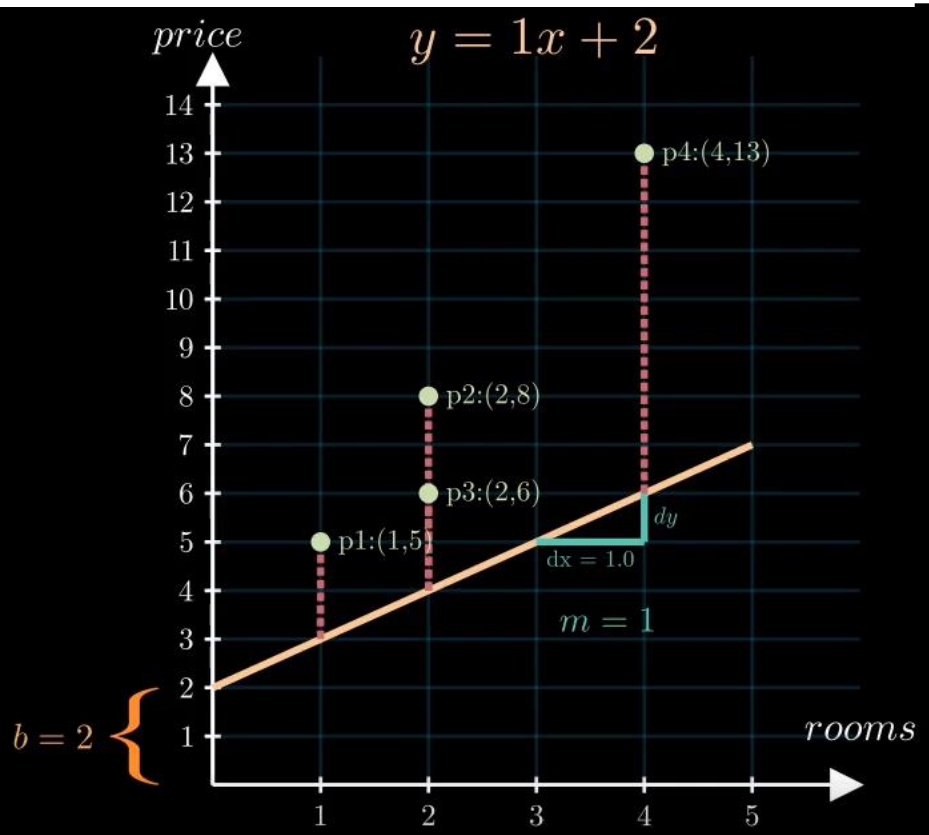
Tangent line

The derivative of the curve tells us the slope of the tangent line that touches the curve at a certain point.

- The tangent to a plane curve at a given point is the straight line that "just touches" the curve at that point.

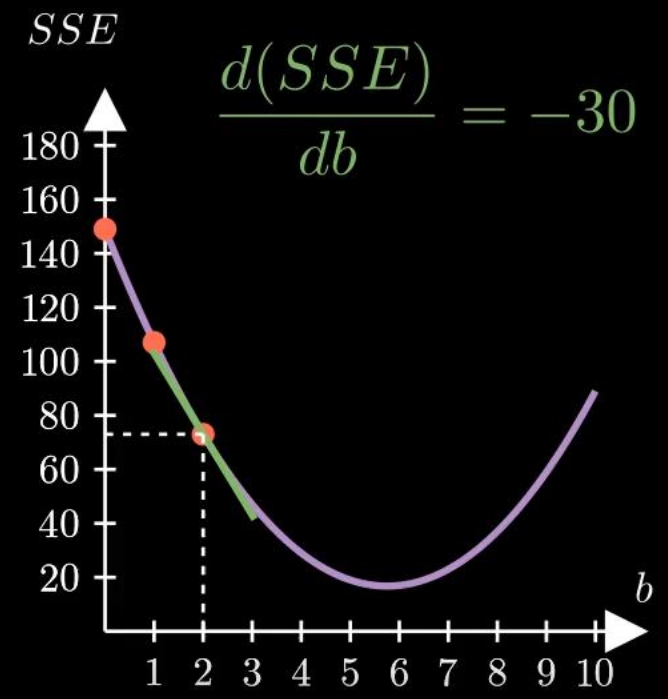
$$y = 1x + b$$

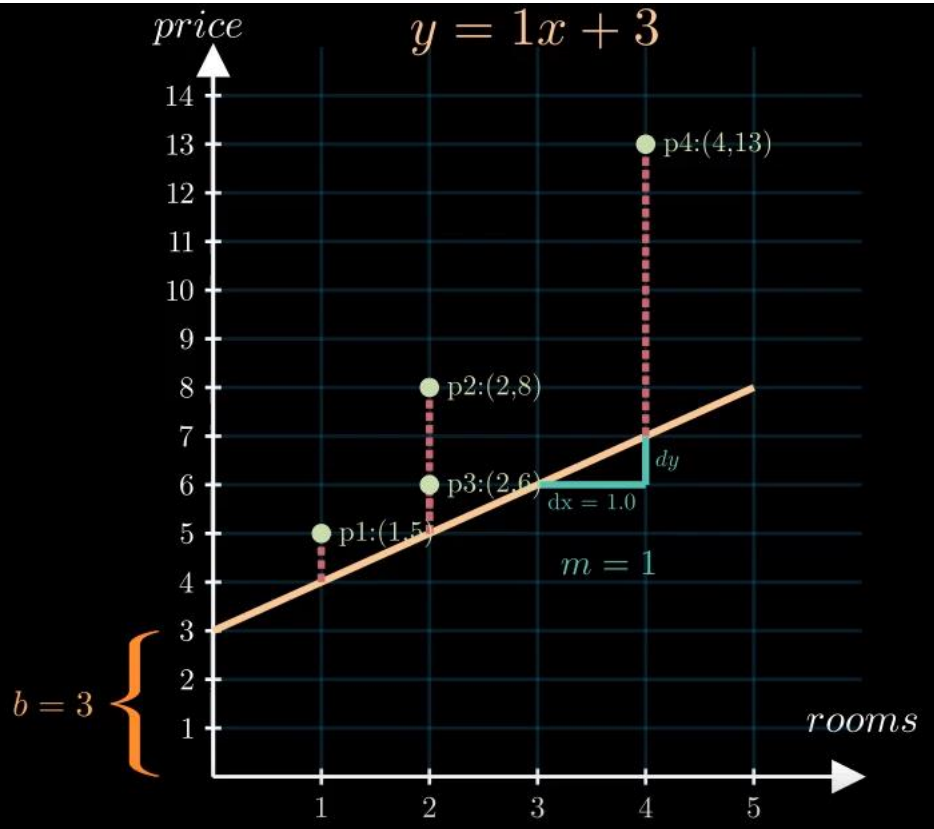




When $b = 2$, the slope of the tangent is negative.

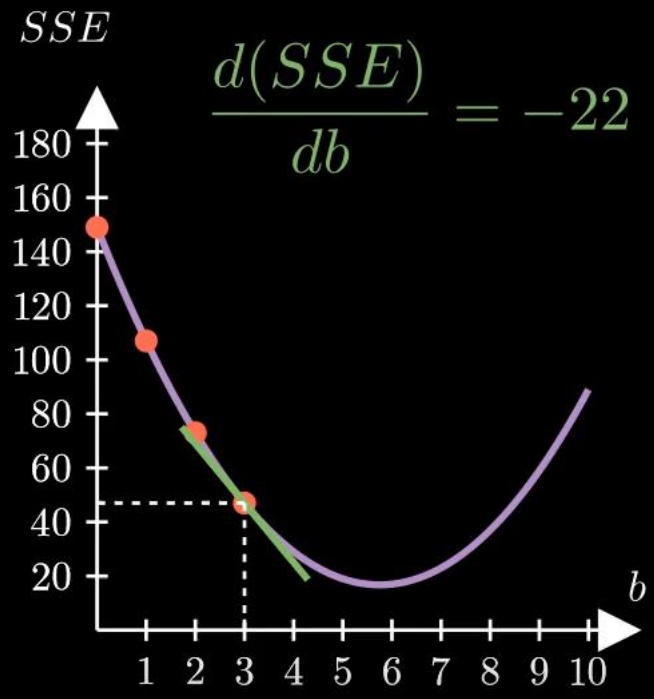
- Increasing b ***decreases*** SSE.
- We should increase b to decrease SSE.

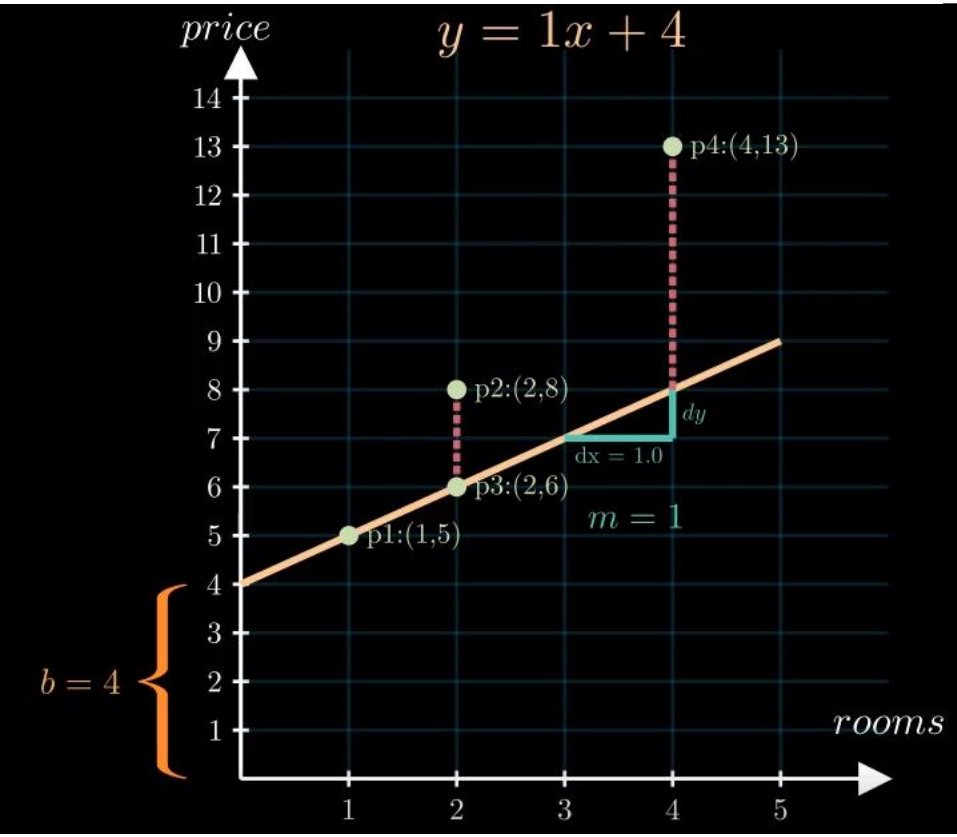




When $b = 3$, the slope of the tangent is negative.

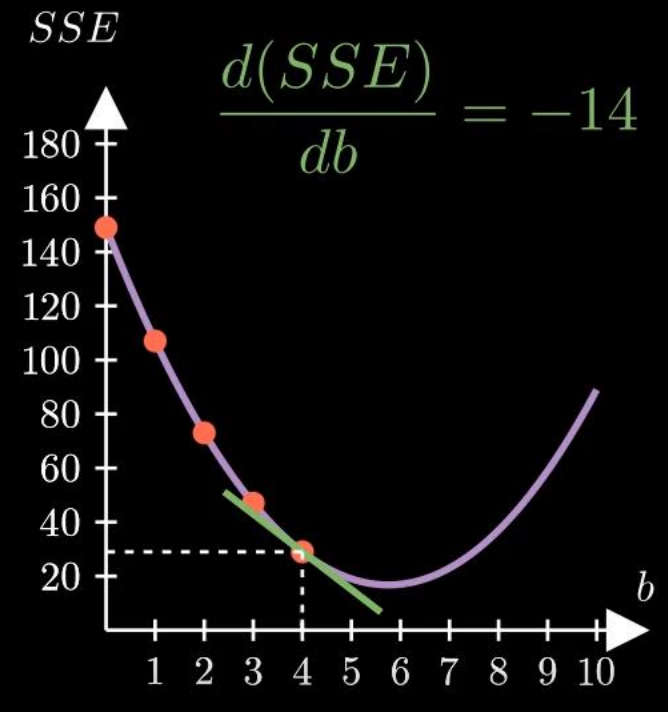
- Increasing b ***decreases*** SSE.
- We should increase b to decrease SSE.

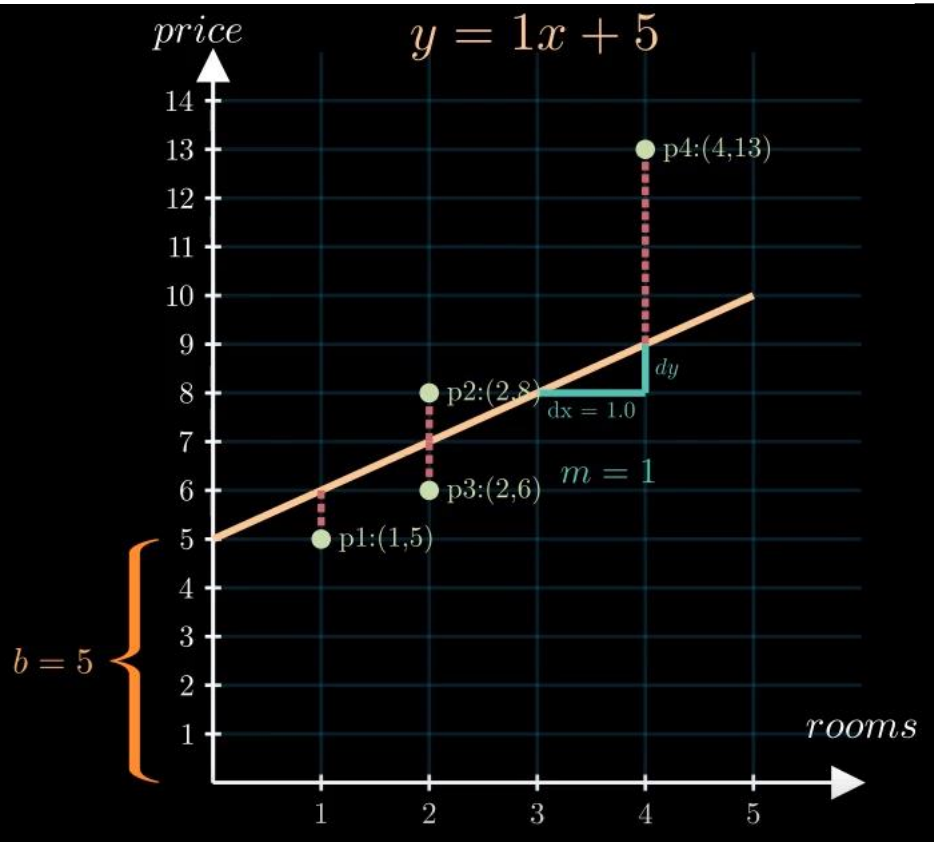




When $b = 4$, the slope of the tangent is negative.

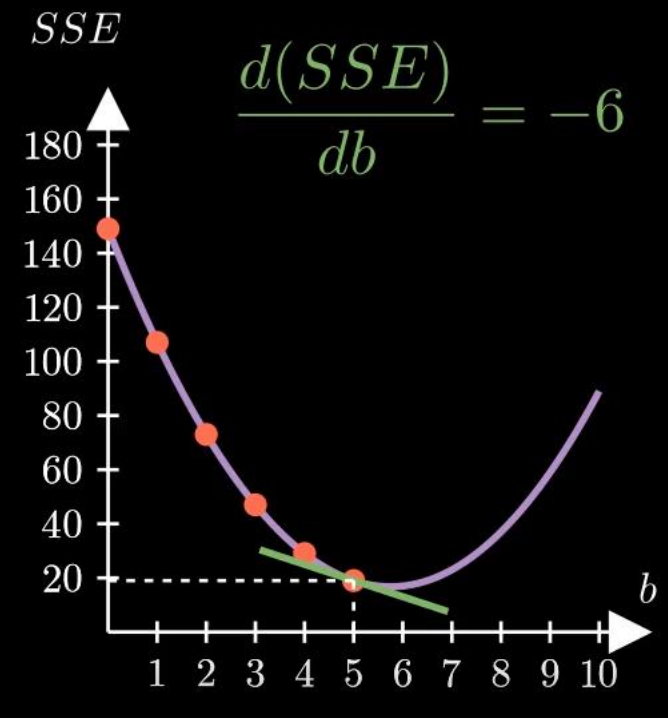
- Increasing b ***decreases*** SSE.
- We should increase b to decrease SSE.

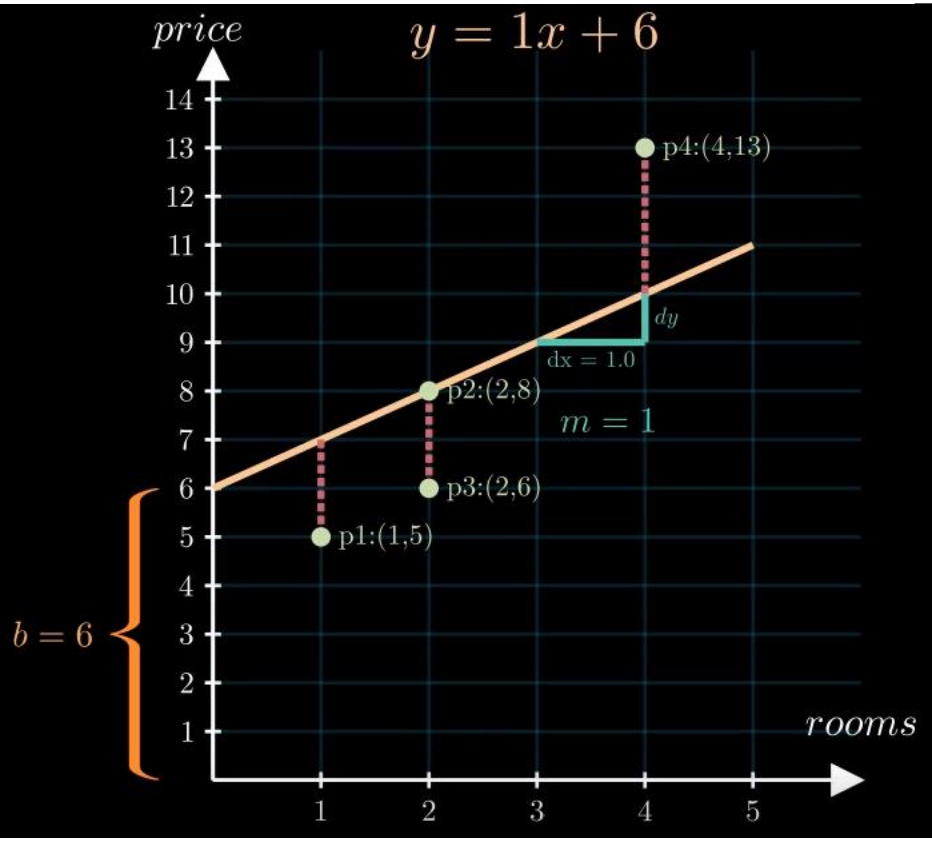




When $b = 5$, the slope of the tangent is negative.

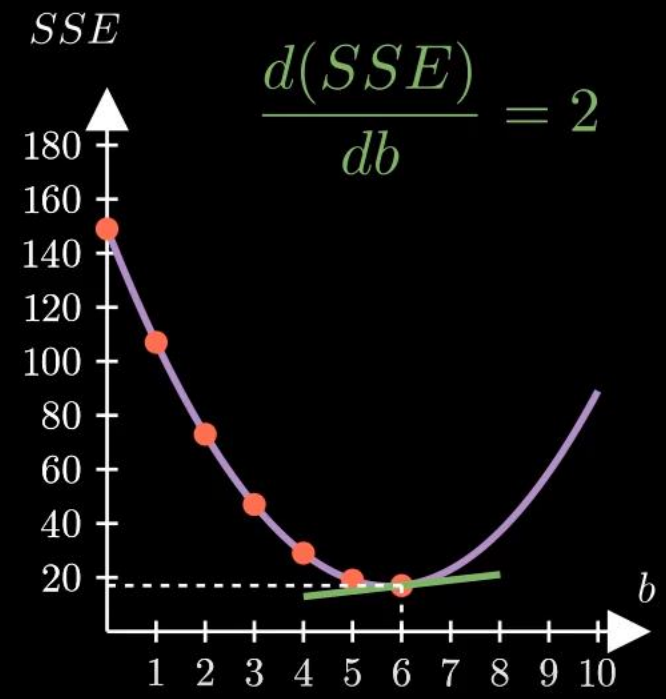
- Increasing b ***decreases*** SSE.
- We should increase b to decrease SSE.

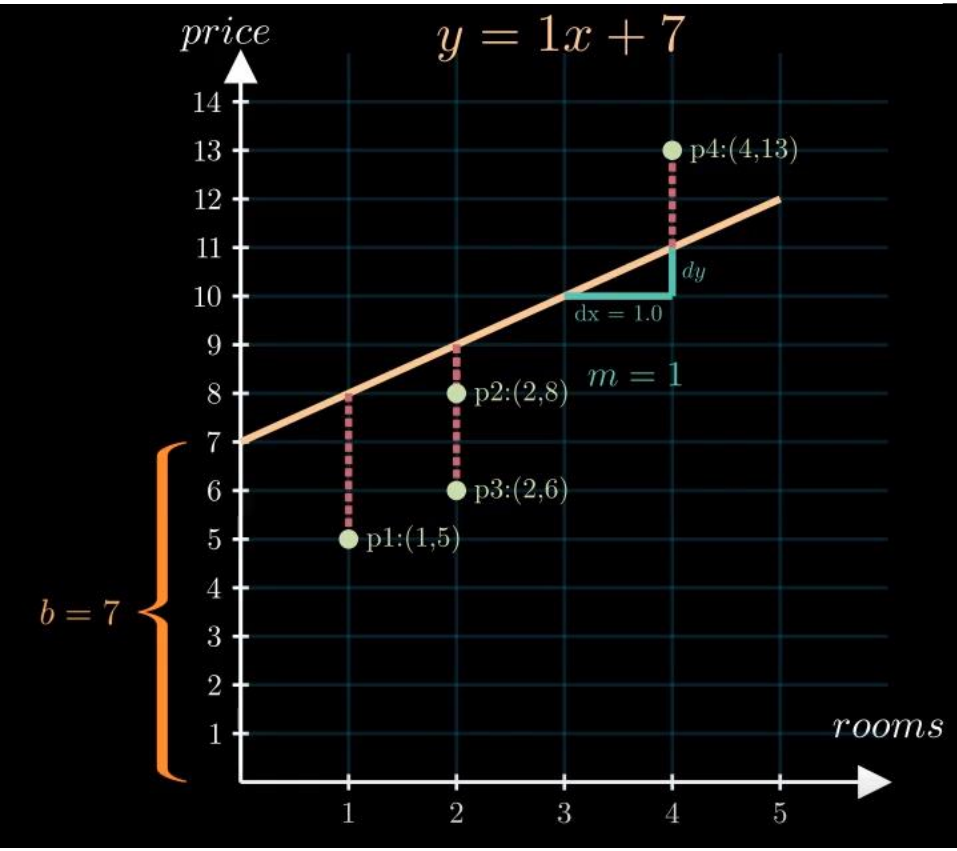




When $b = 6$, the slope of the tangent is negative.

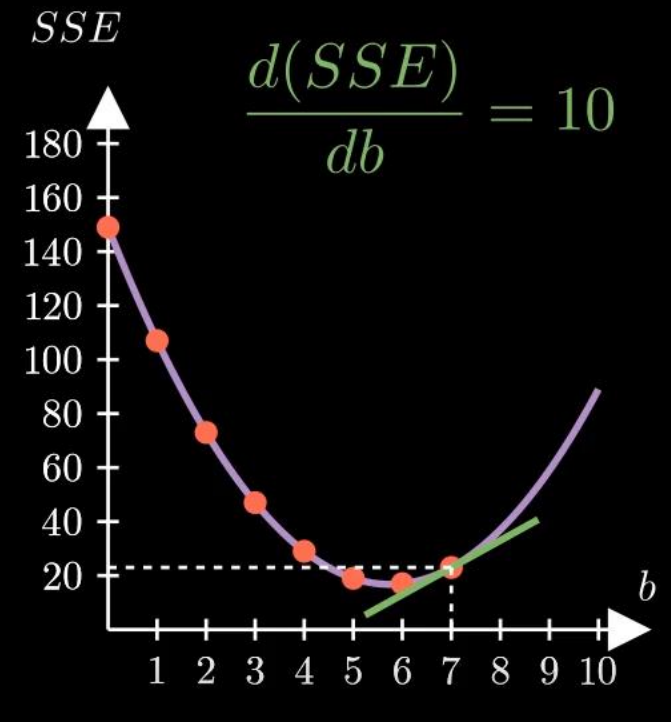
- Increasing b ***decreases*** SSE.
- We should increase b to decrease SSE.

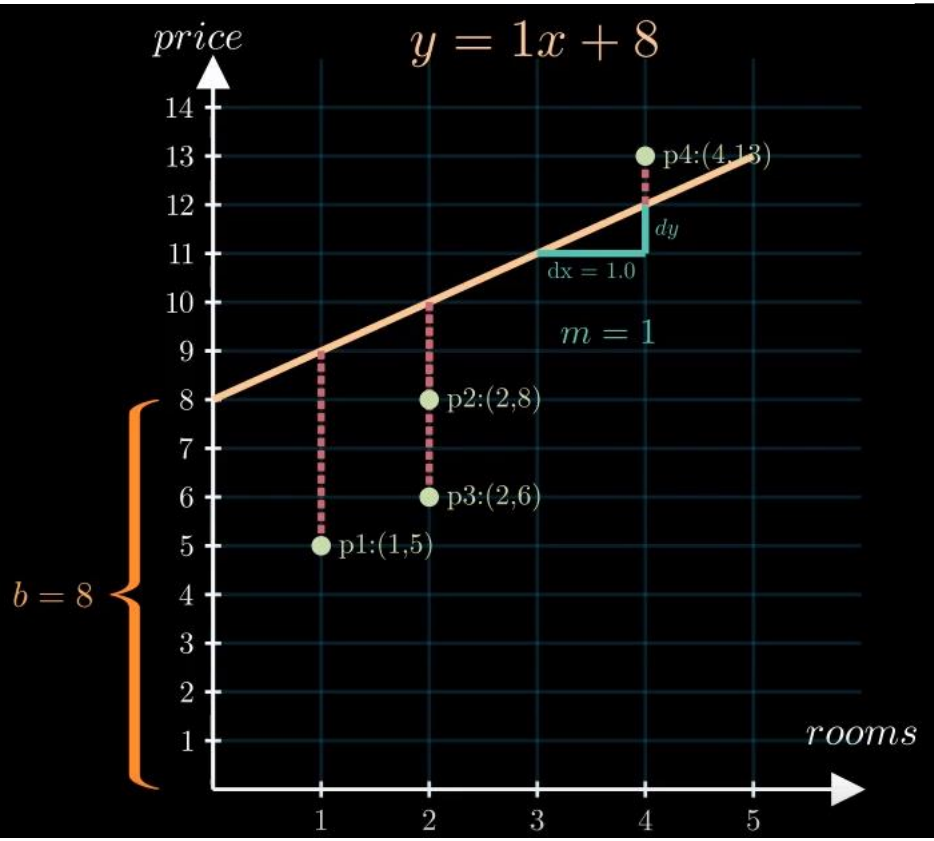




When $b = 7$, the slope of the tangent is positive.

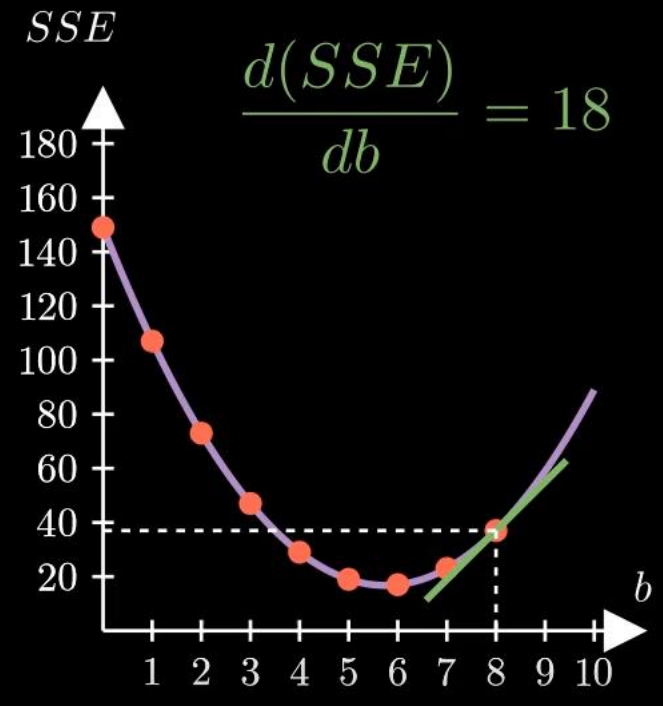
- Increasing b ***increases*** SSE.
- We should decrease b to decrease SSE.





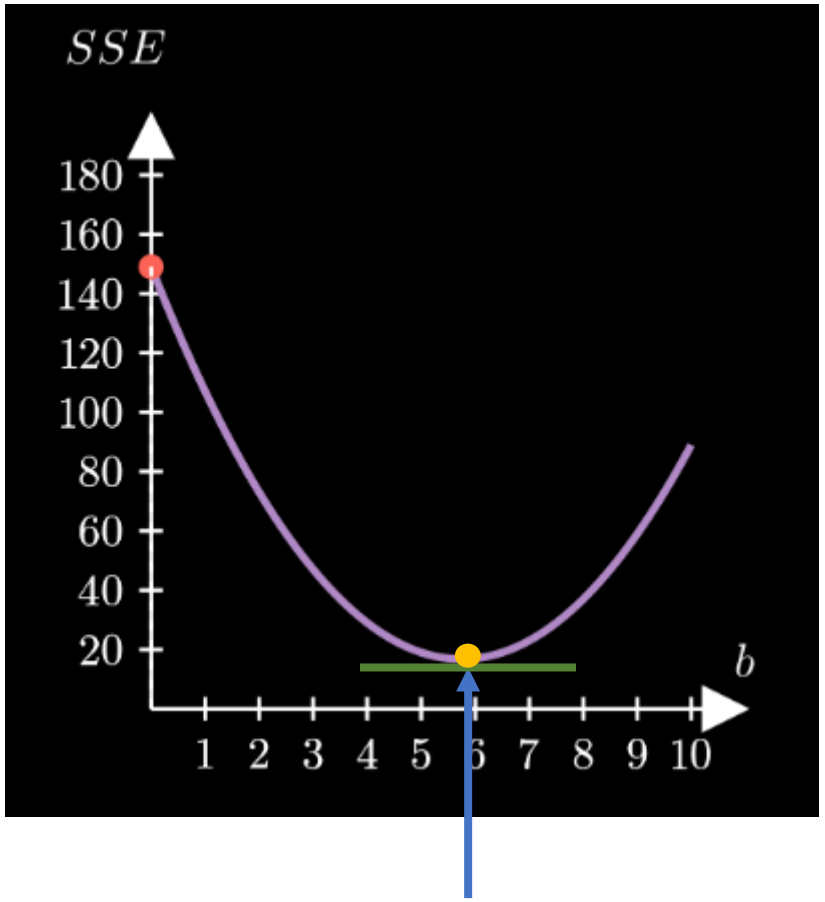
When $b = 8$, the slope of the tangent is negative.

- Increasing b ***increases*** SSE.
- We should decrease b to decrease SSE.



Finding the minimum loss

- At the minimum value, the slope of the tangent is 0
- When the magnitude of the slope is large, we are far away from the minimum SSE
- When the magnitude of the slope is small, we are near to the minimum SSE

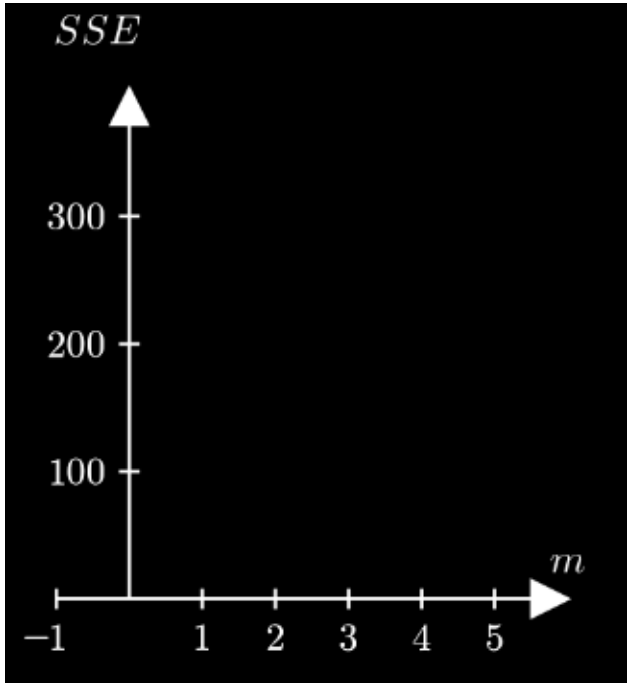


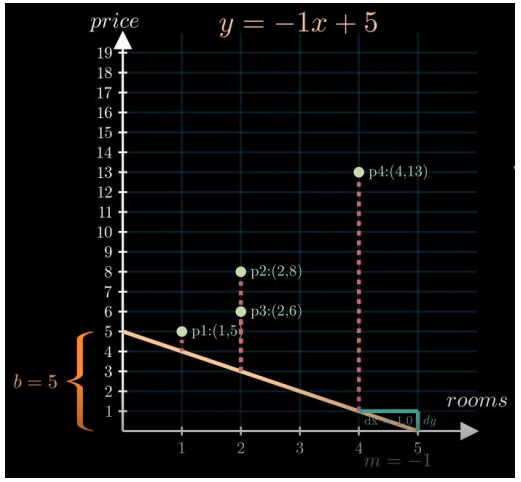
When m=2, what is the b which minimizes the SSE?

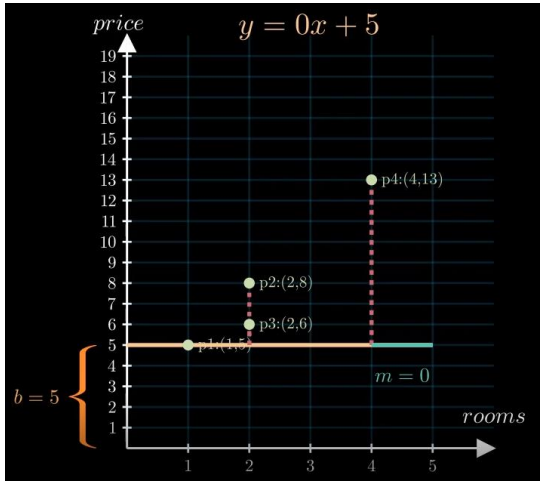
Exercise

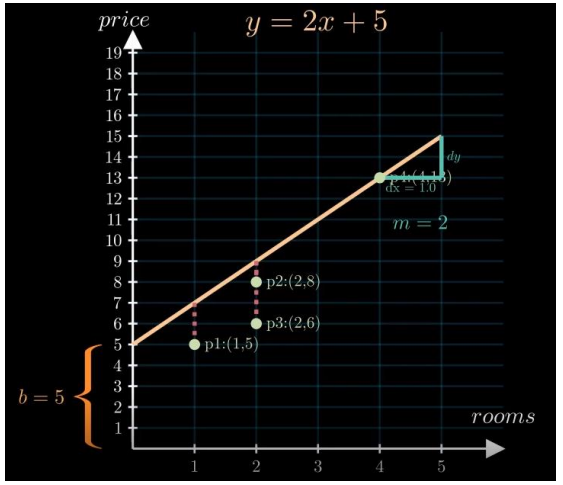
$$y = mx + 5$$

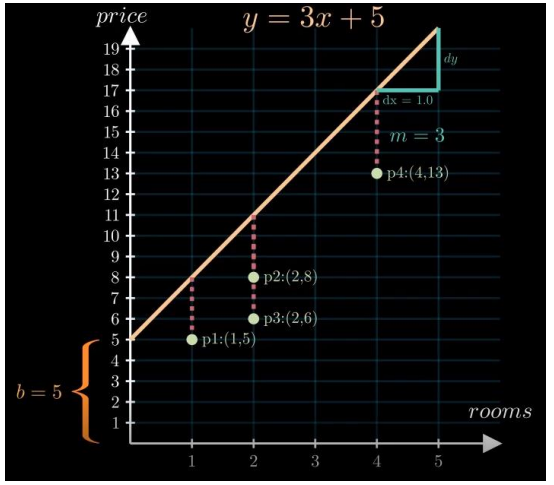
- With b=5, compute SSE with m ={-1, 0, 1, 2, 3}.
- Plot the SSE for the different m values. Find the best model.







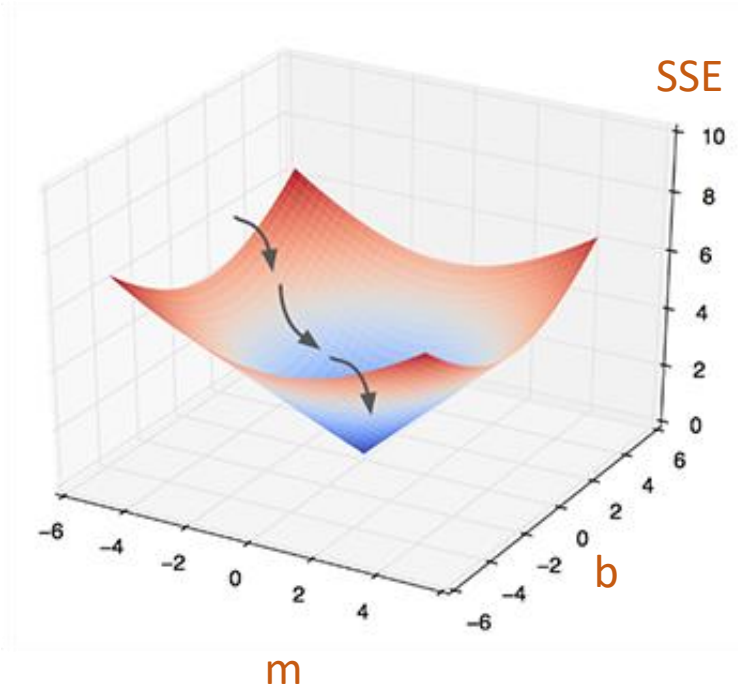




Optimizing multiple parameters at a time

$$y = mx + b$$

Objective: Find **m** of and **b** which minimize the SSE



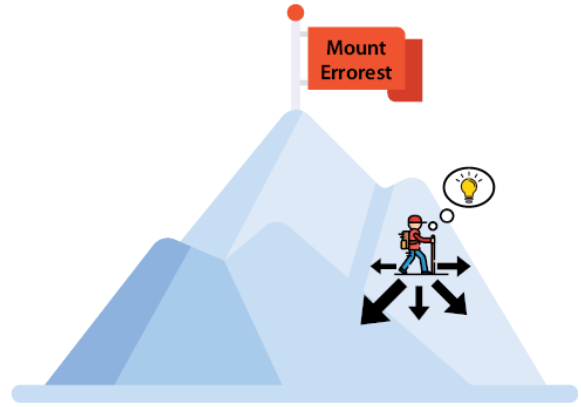
b

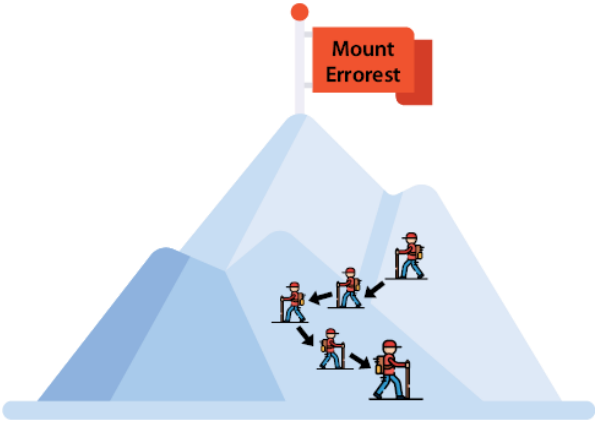
	-5	-4	-3	-2	-1	0	1	2	3	4	5
5	50	41	34	29	26	25	26	29	34	41	50
4	41	32	25	20	17	16	17	20	25	32	41
3	34	25	18	13	10	9	10	13	18	25	34
2	29	20	13	8	5	4	5	8	13	20	29
1	26	17	10	5	2	1	2	5	10	17	26
0	25	16	9		1	0	1	4	9	16	25
-1	26	17	10	5	2	1	2	5	10	17	26
-2	29	20	13	8	5	4	5	8	13	20	29
-3	34	25	18	13	10	9	10	13	18	25	34
-4	41	32	25	20	17	16	17	20	25	32	41
-5	50	41	34	29	26	25	26	29	34	41	50

m

Gradient Descent

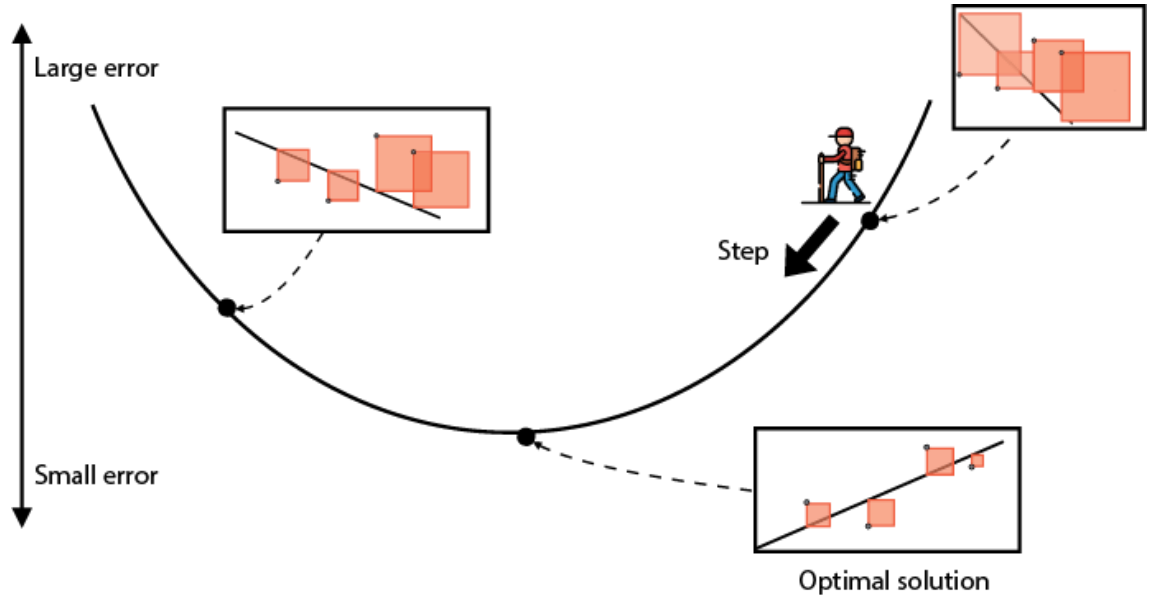
- A major part of machine learning is to optimize the model parameters to fit the data
  - m and b in linear regression models
- A gradient measures the change in the error function with respect to the change in model parameters
- Gradient descent is an important algorithm in machine learning which estimates where a function outputs its lowest values.
  1. We start somewhere on the mountain.
  2. Find the best direction to take one small step, in the direction of greatest descent. Take this small step.
  3. Repeat the step 2 many times





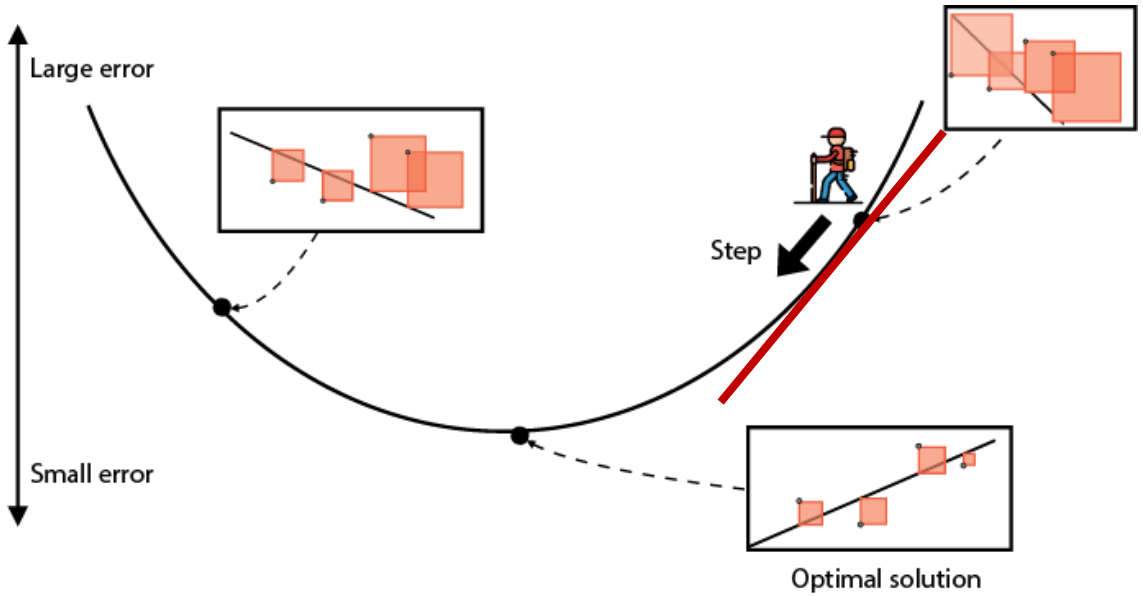
Gradient Descent for linear regression

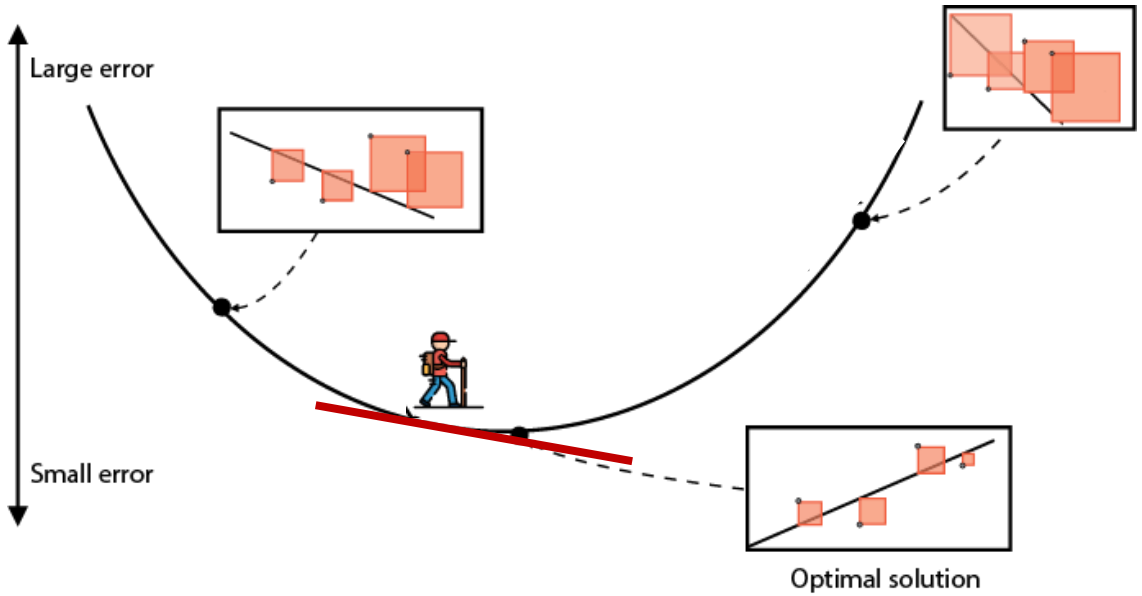
1. Start with any line (random m and b)
2. Find the best direction to move our line a little bit. Move the line a little bit in this direction by adjusting the values of m and b.
3. Repeat steps 2 many times



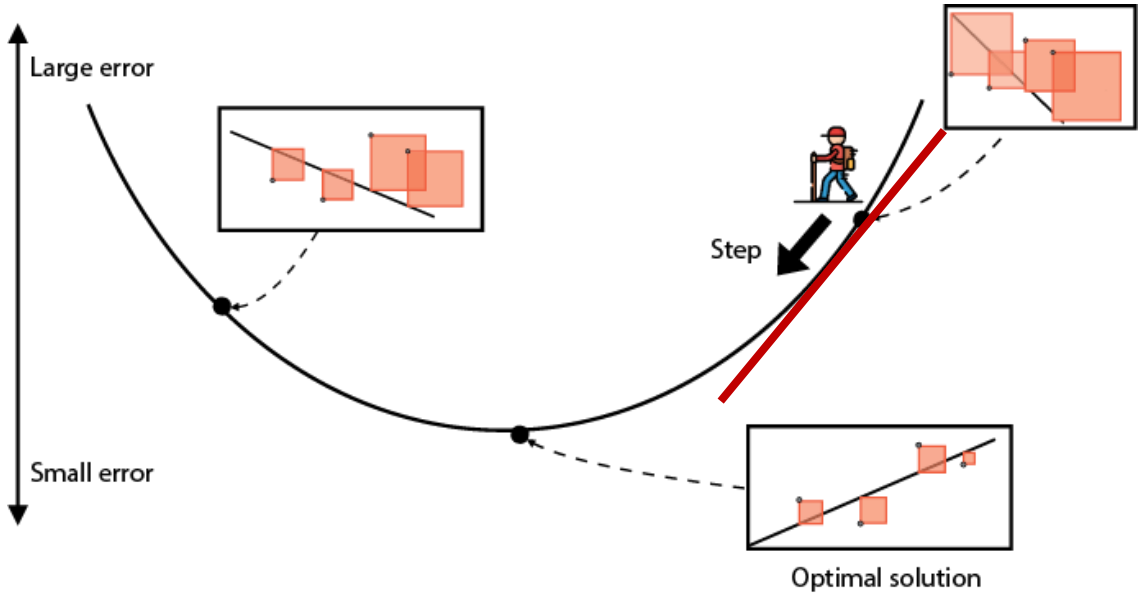
Gradient Descent (More Details)

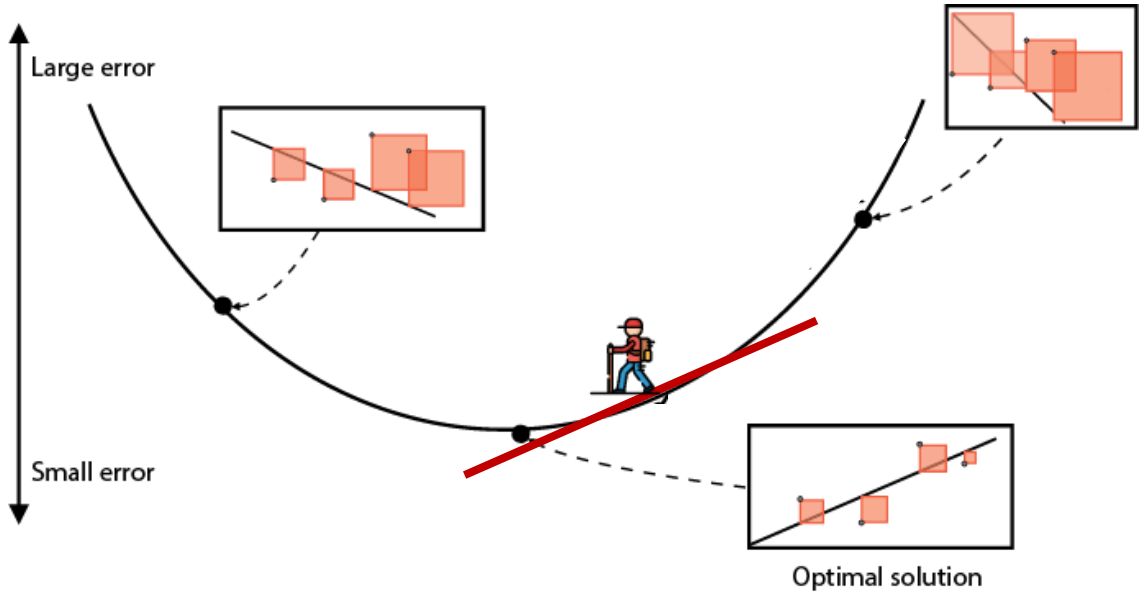
- A positive slope tells us that we should take a step to the left to get to the lowest SSE
- A negative slope tells us that we should take a step to the right to get to the lowest SSE





- A relatively large value for the derivative corresponds to a steep slope for the tangent line
  - We are relatively far from the bottom of the curve
  - We should take a relatively large step
- A relatively small value for the derivative suggest we are relatively close to the bottom of the curve
  - We should take a relatively small step

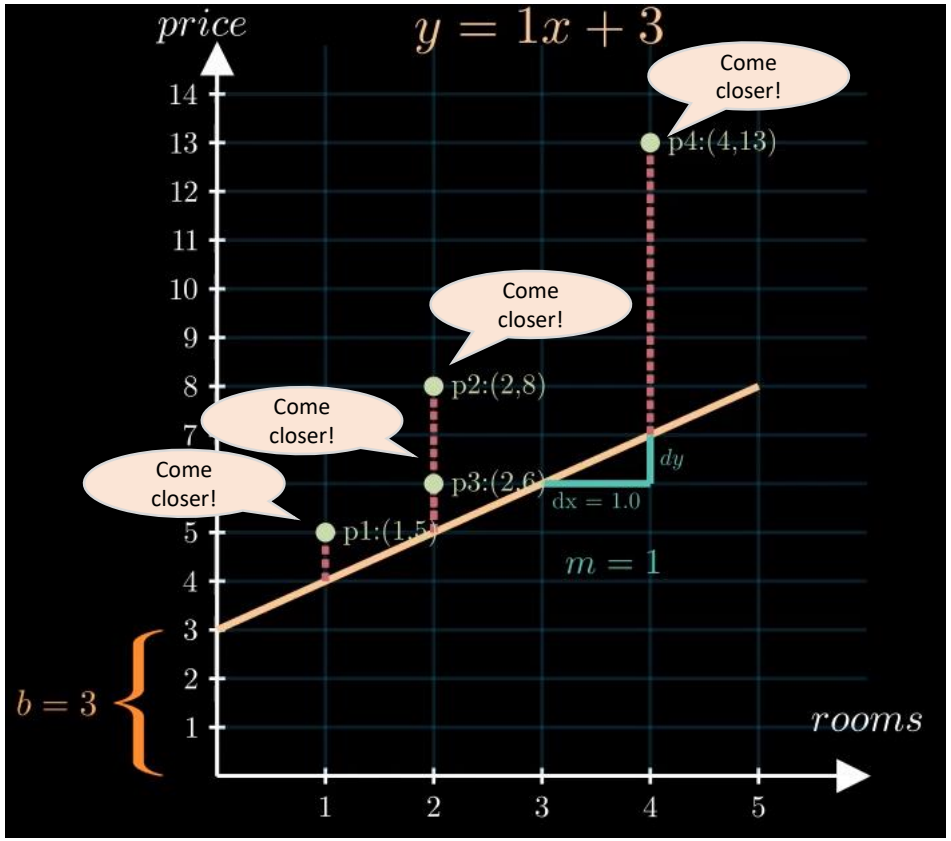




How to reduce the prediction error? (I)

$y = 1x + 3$

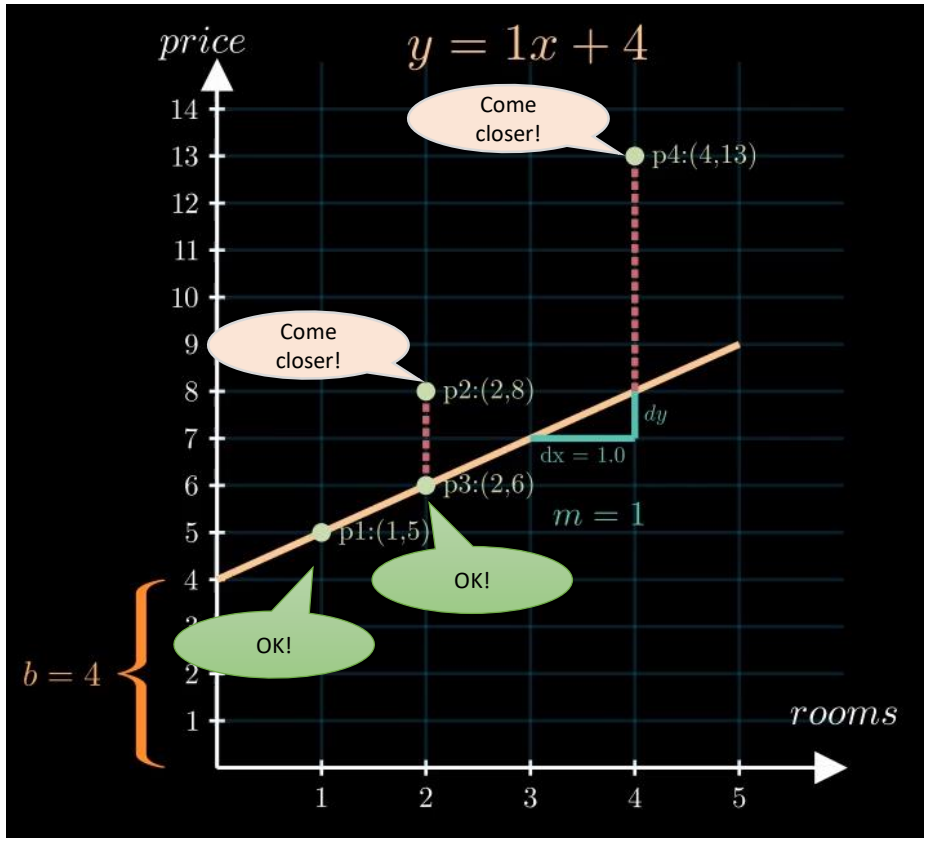
$SSE = 47$



A better model

$y = 1x + 4$

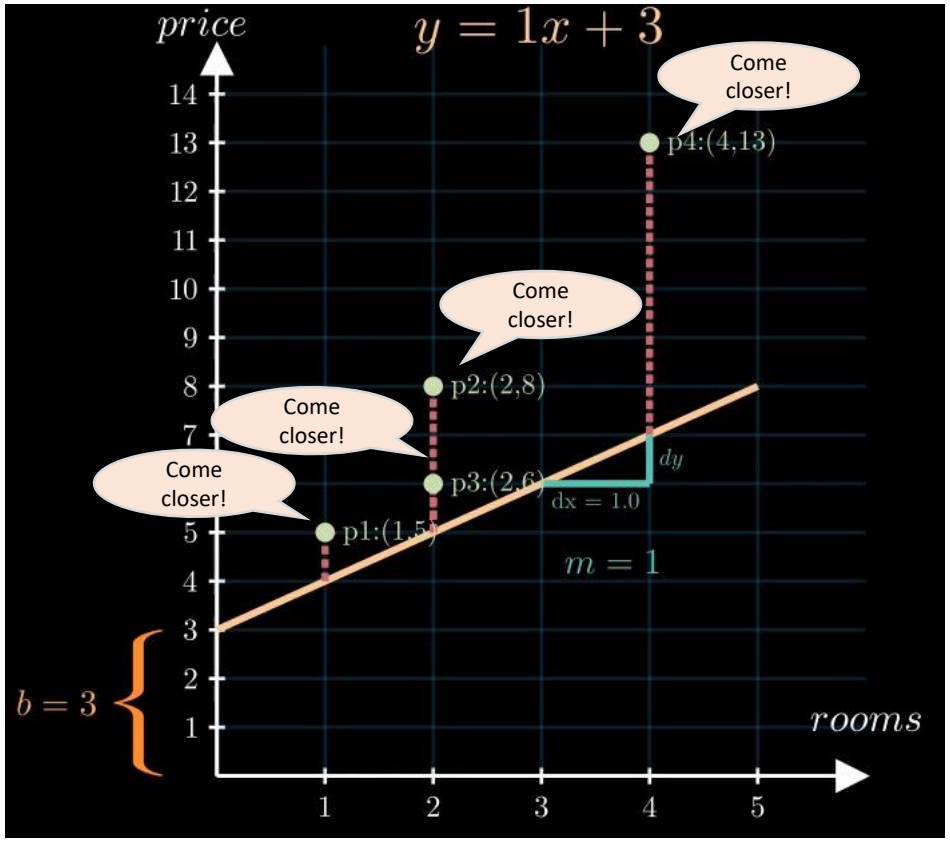
$SSE = 19$



How to reduce the prediction error? (II)

$y = 1x + 3$

$SSE = 47$



A better model

$y = 2x + 3$

$SSE = 6$

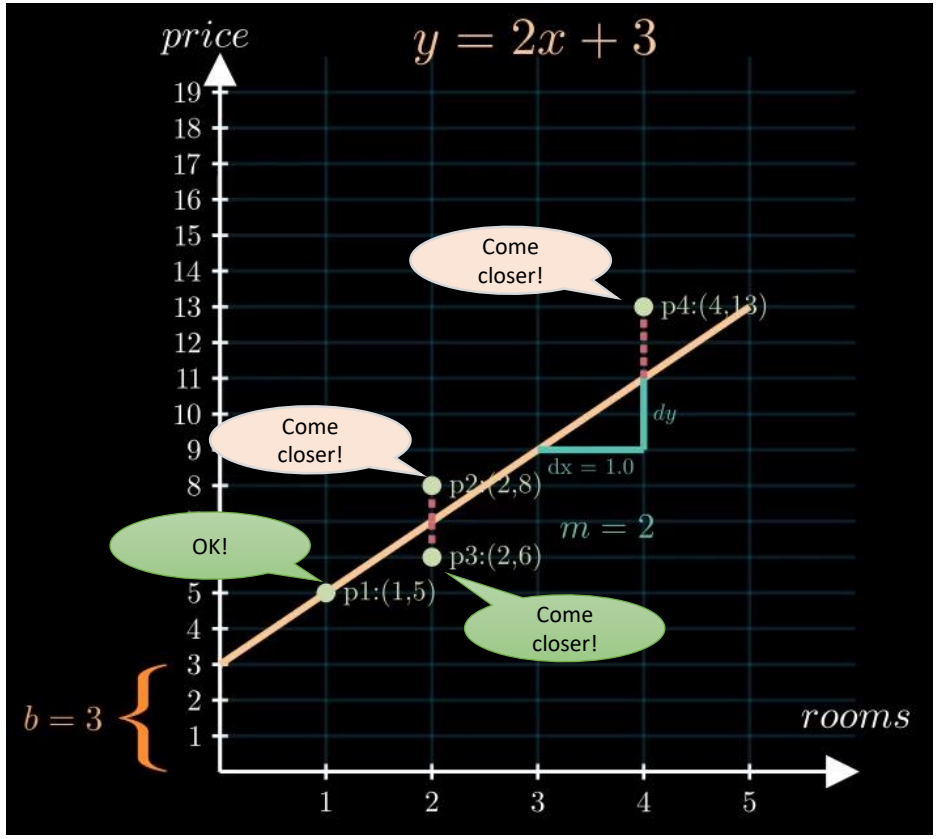
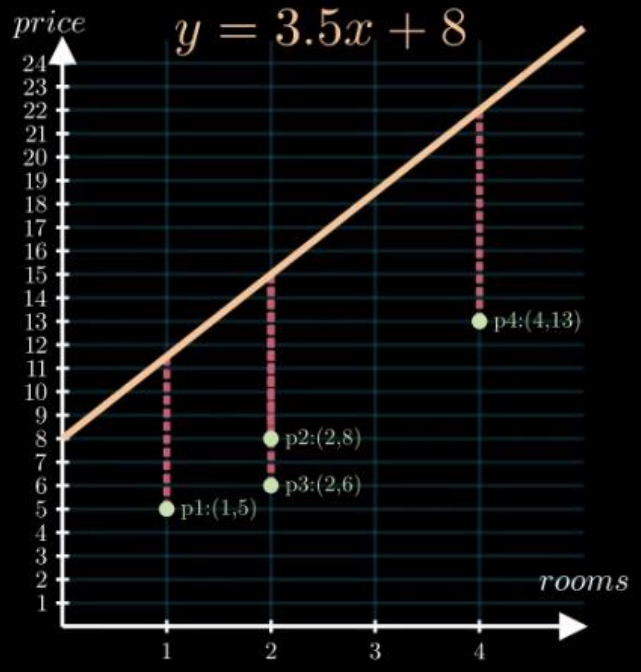
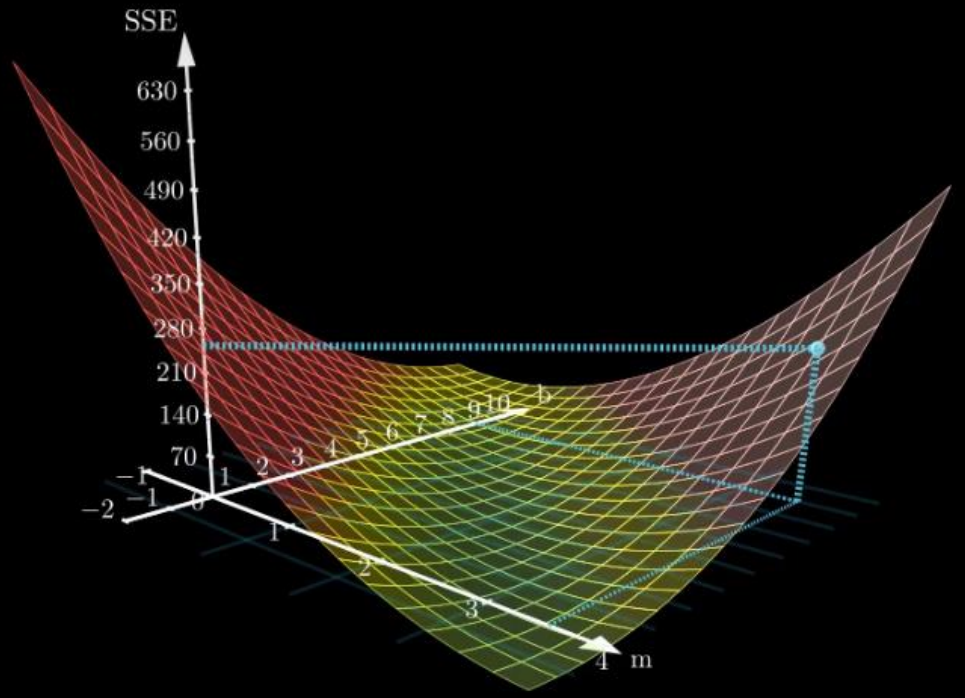
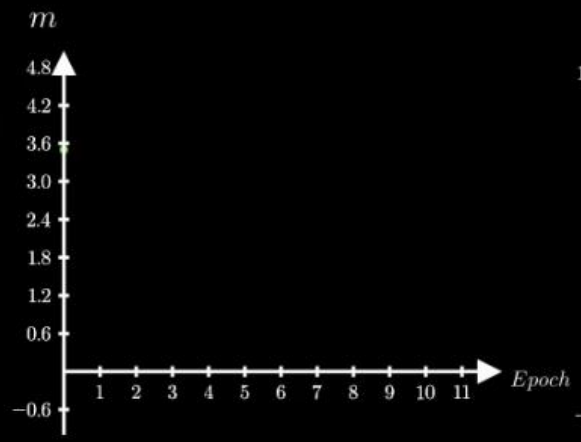
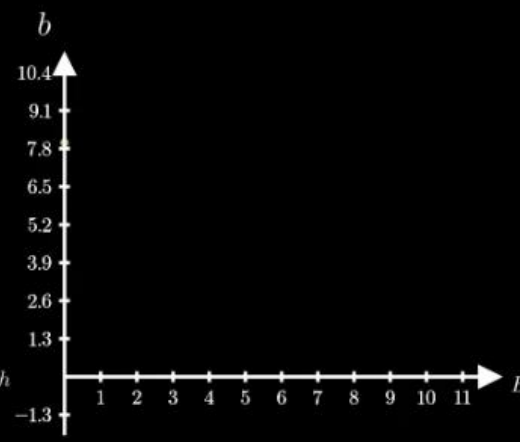


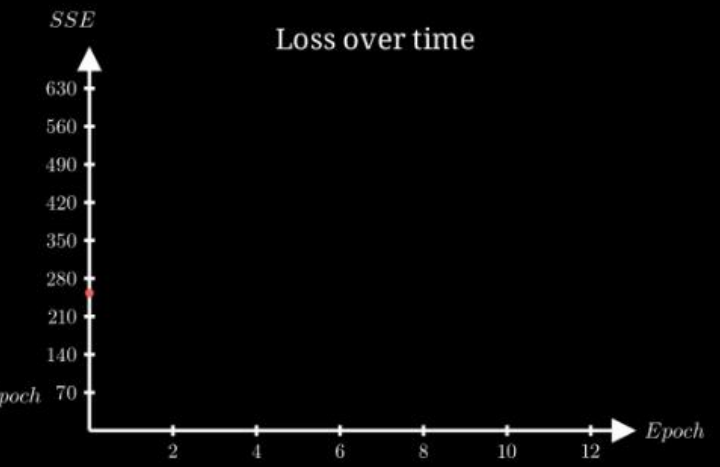
Illustration: Gradient Descent



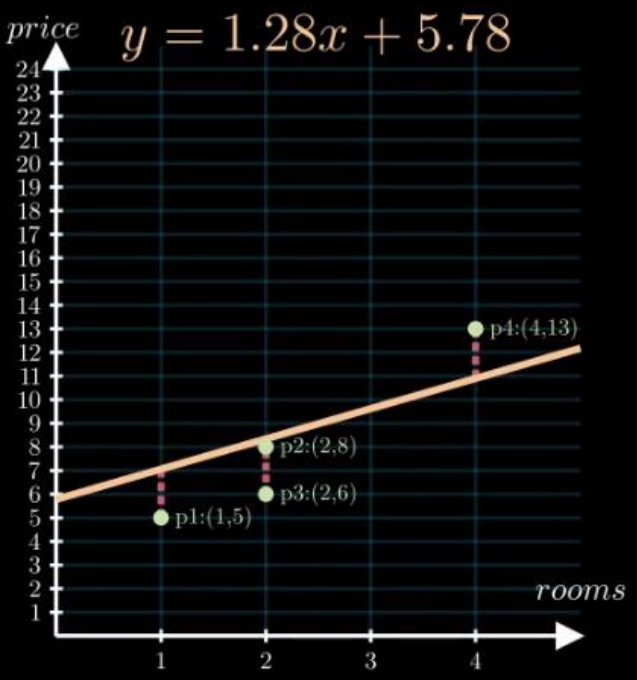


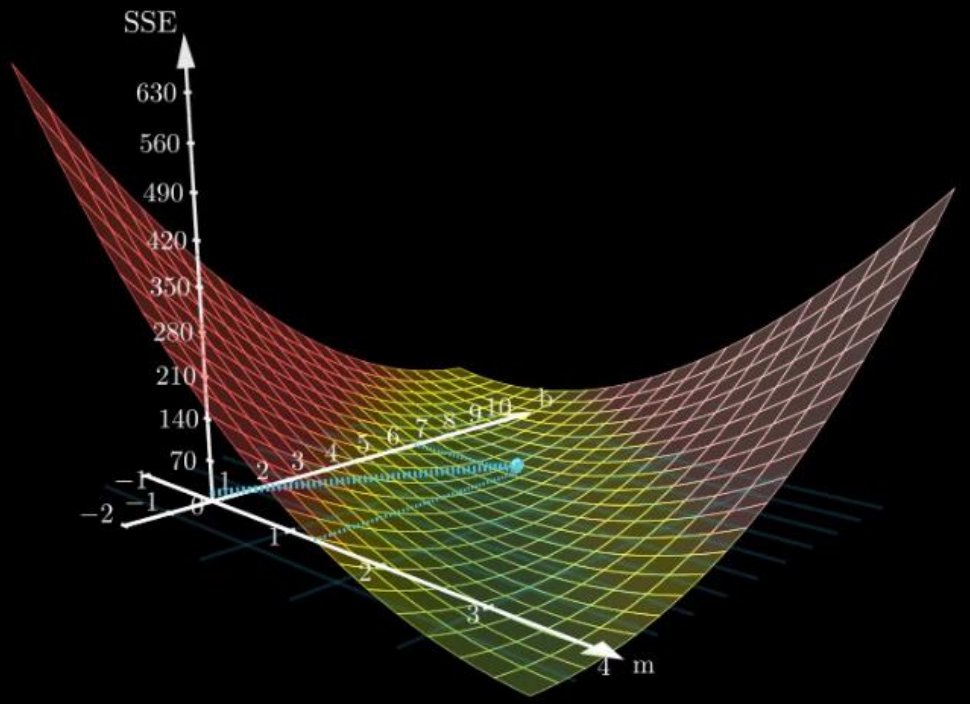


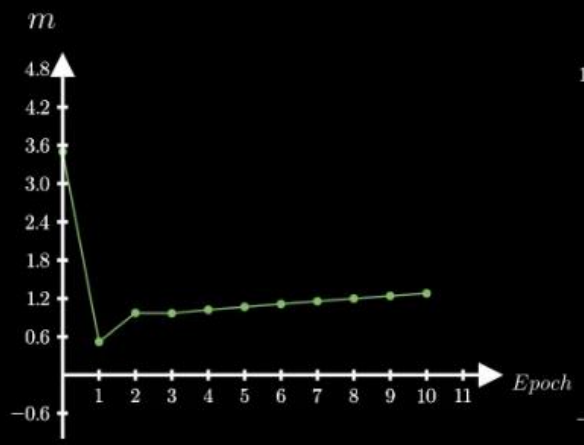


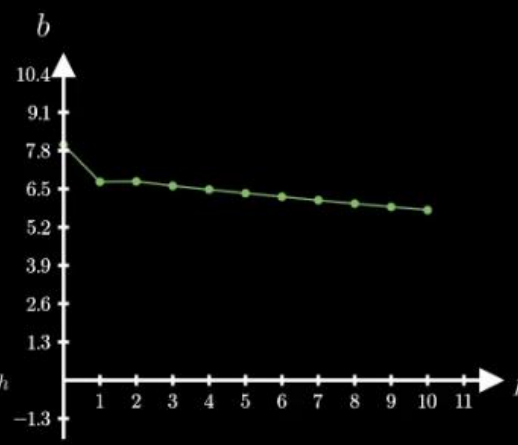


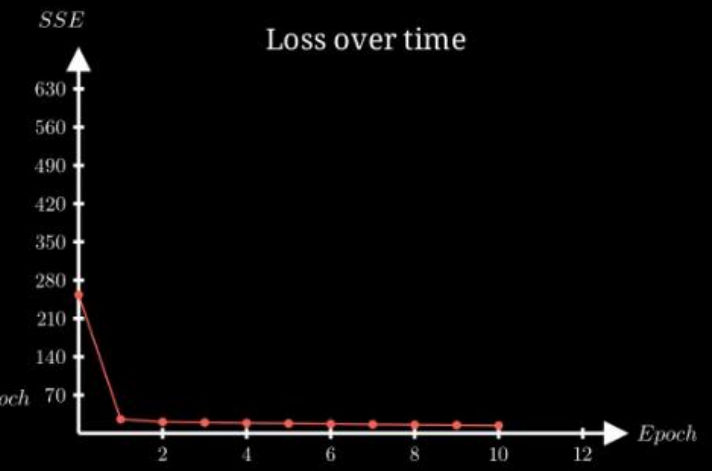
Epoch: 10 SSE: 14.24
m: 1.28 b: 5.78
Learning Rate: 0.02





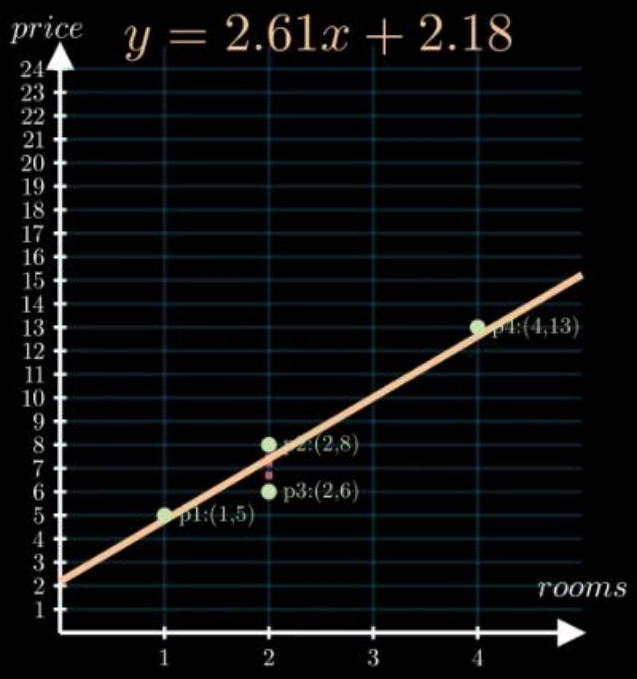


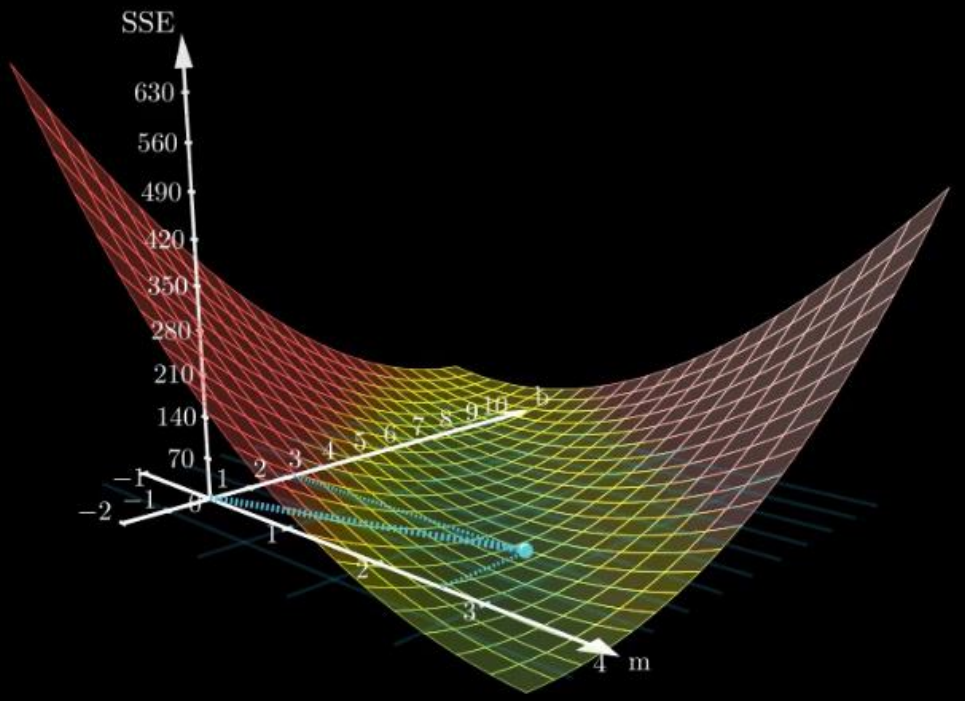


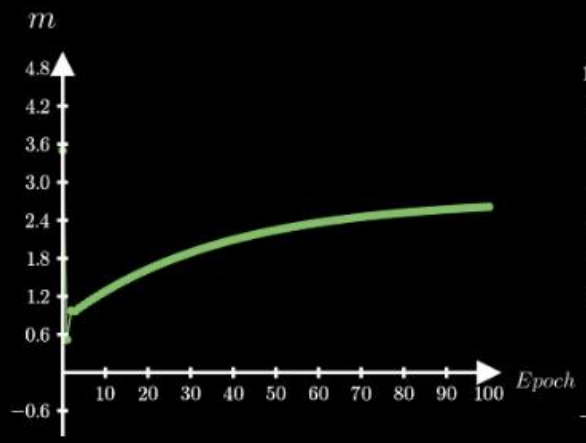


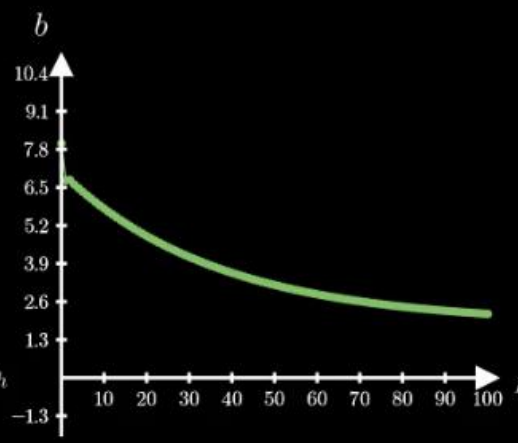
Loss over time

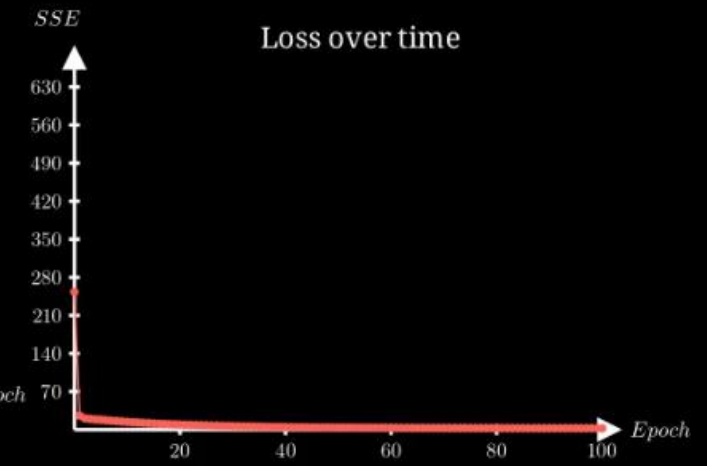
Epoch: 100 SSE: 2.51
m: 2.61 b: 2.18
Learning Rate: 0.02







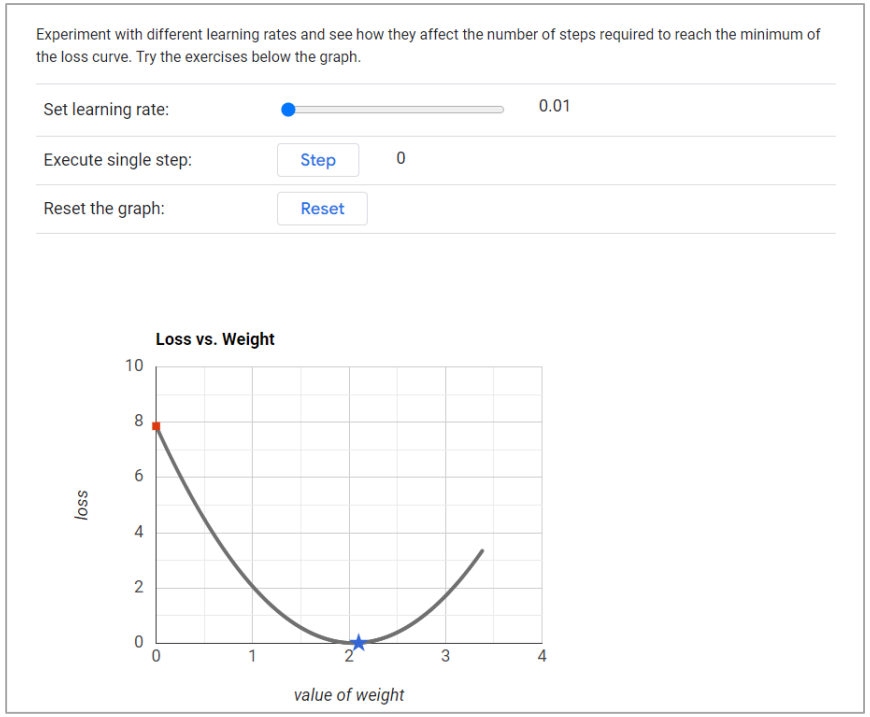




Loss over time

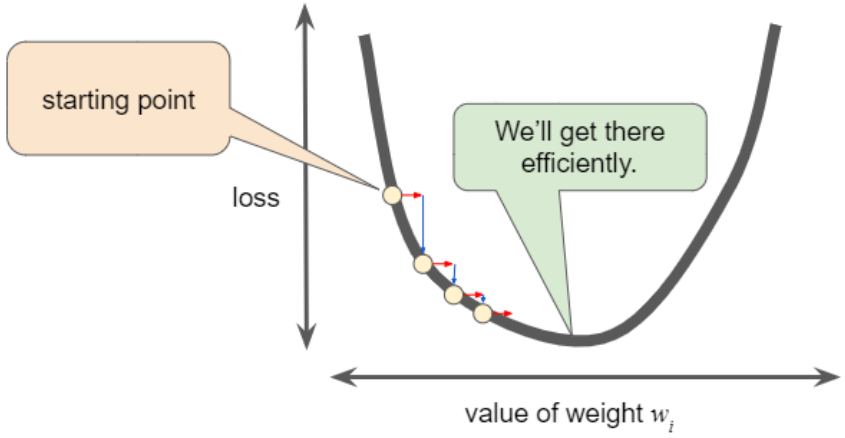
Learning Rate

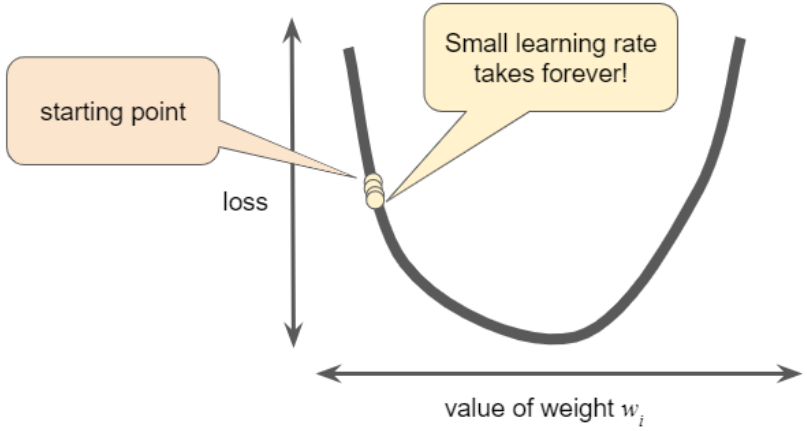
- Select the learning rate λ before the training, which determines the size of the step using the *learning rate*

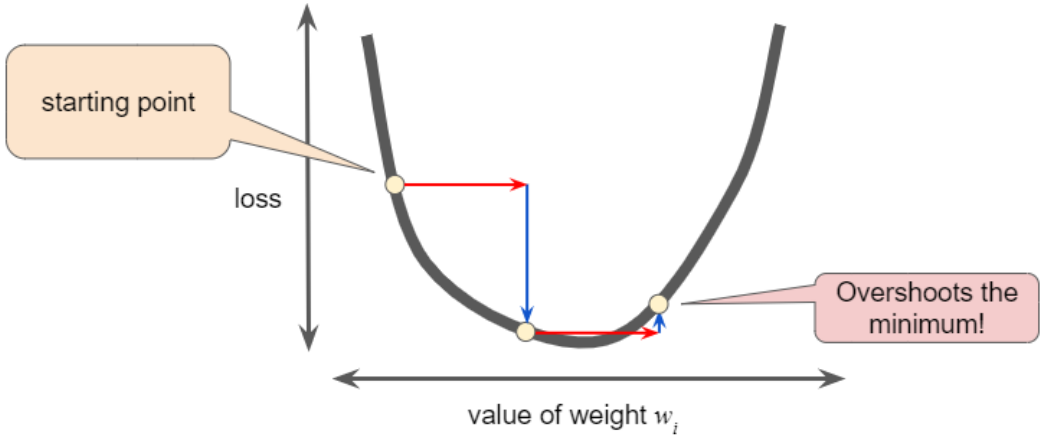


https://developers.google.com/machine-learning/crash-course/fitter/graph

Effects of learning rate on model training

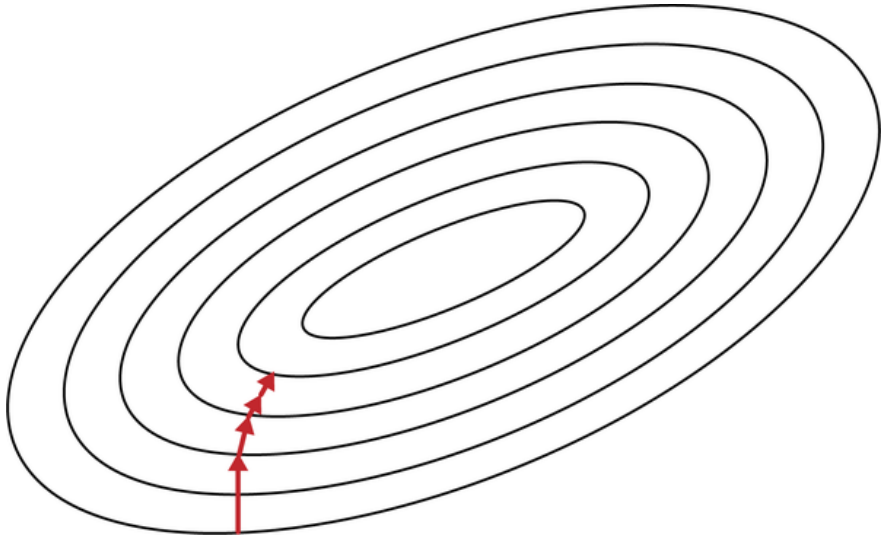


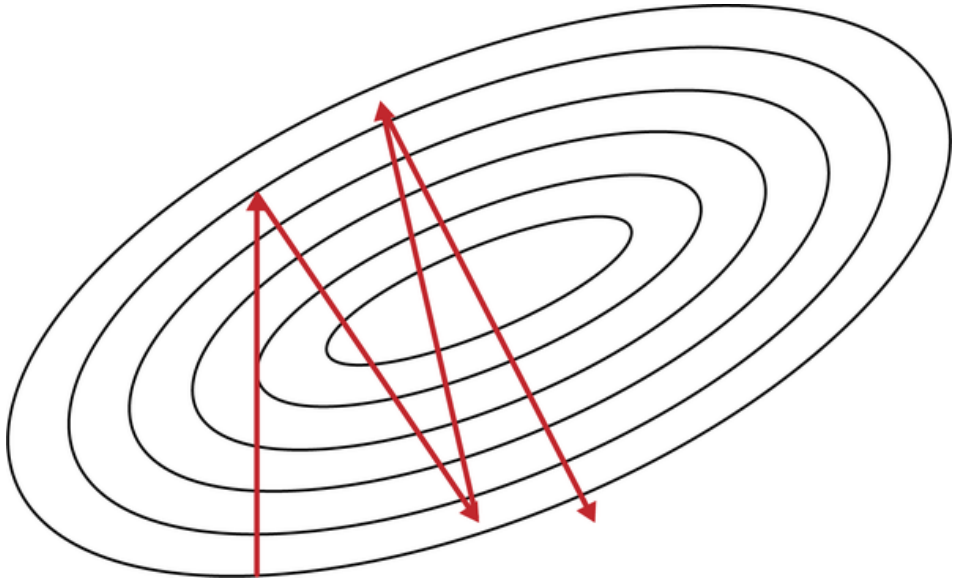




Visualizing the error surface as a set of contours

- Convergence is difficult when our learning rate is too large





Multiple regression

bias

weights

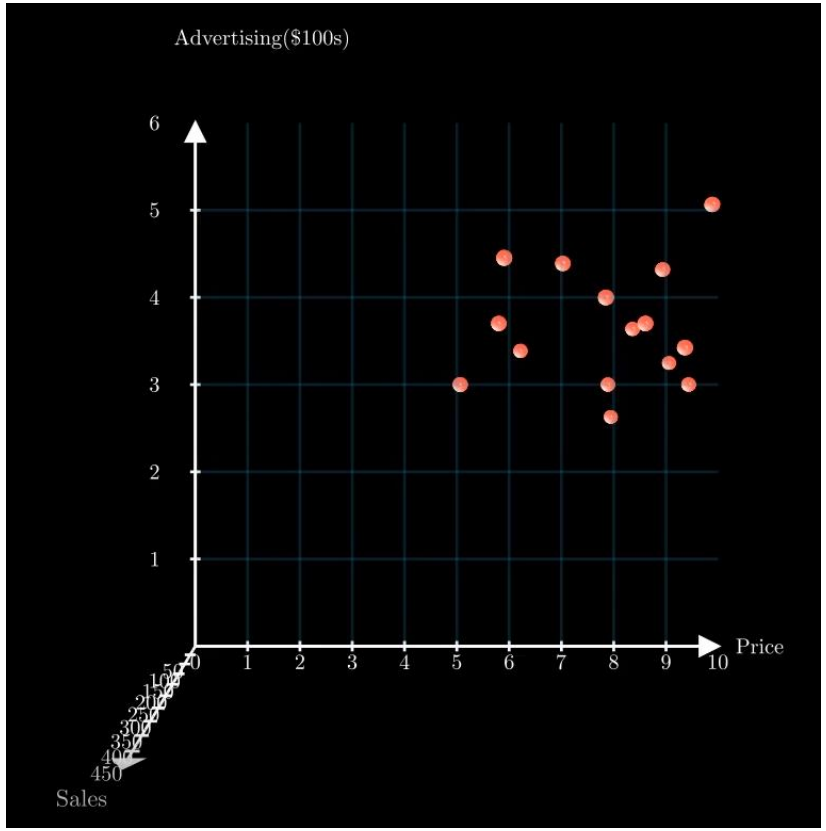
$$y = b + w_1 x_1 + w_2 x_2 + \cdots$$

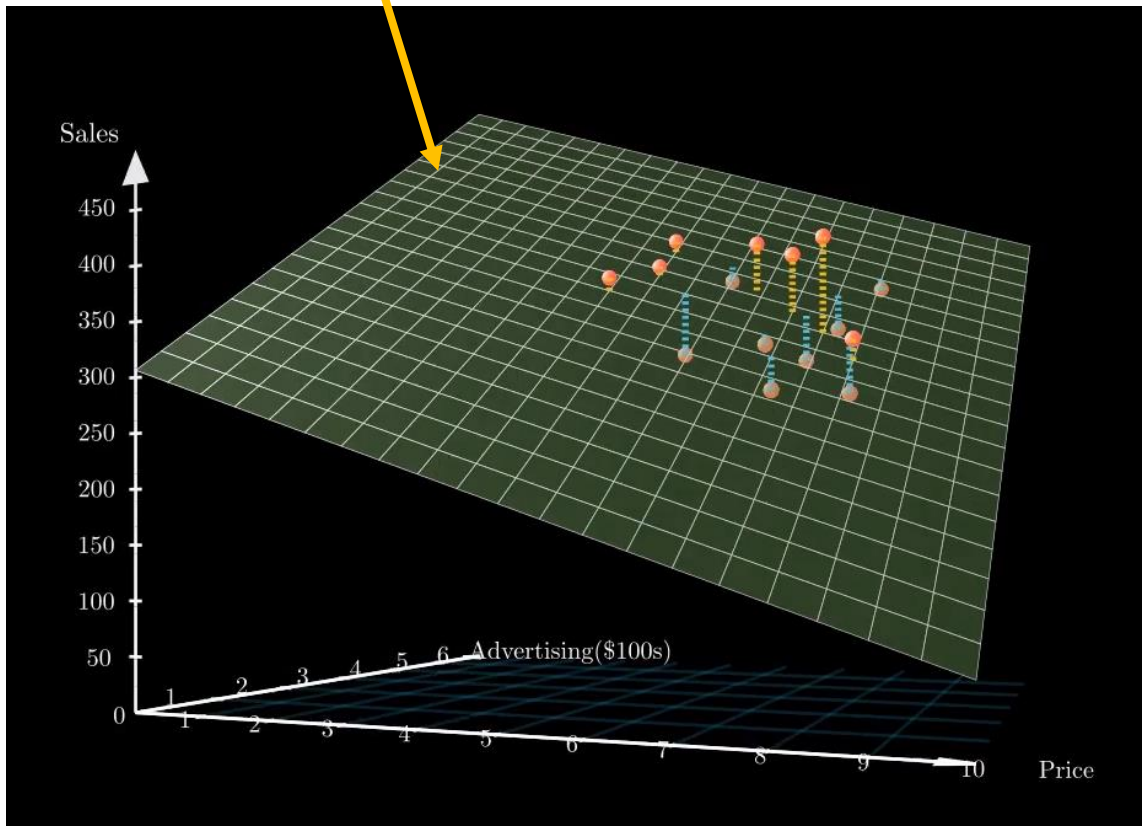
- Multiple regressions are similar to simple regression except that we are dealing with more than one independent variable.
- Example
  - Dependent variable:
    - **Pie sales** (units per week)
  - Independent variables:
    - **Price** (in $)
    - **Advertising Budget** ($100's)

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.5	3.3
2	460	7.5	3.3
3	350	8	3
4	430	8	4.5
5	350	6.8	3
6	380	7.5	4
7	430	4.5	3
8	470	6.4	3.7
9	450	7	3.5
10	490	5	4
11	340	7.2	3.5
12	300	7.9	3.2
13	440	5.9	4
14	450	5	3.5
15	300	7	2.7

A multiple regression model

$$\boldsymbol{Sales} = \mathbf{306.53} + (\mathbf{-24.98})Price + (\mathbf{74.13})Advertising$$



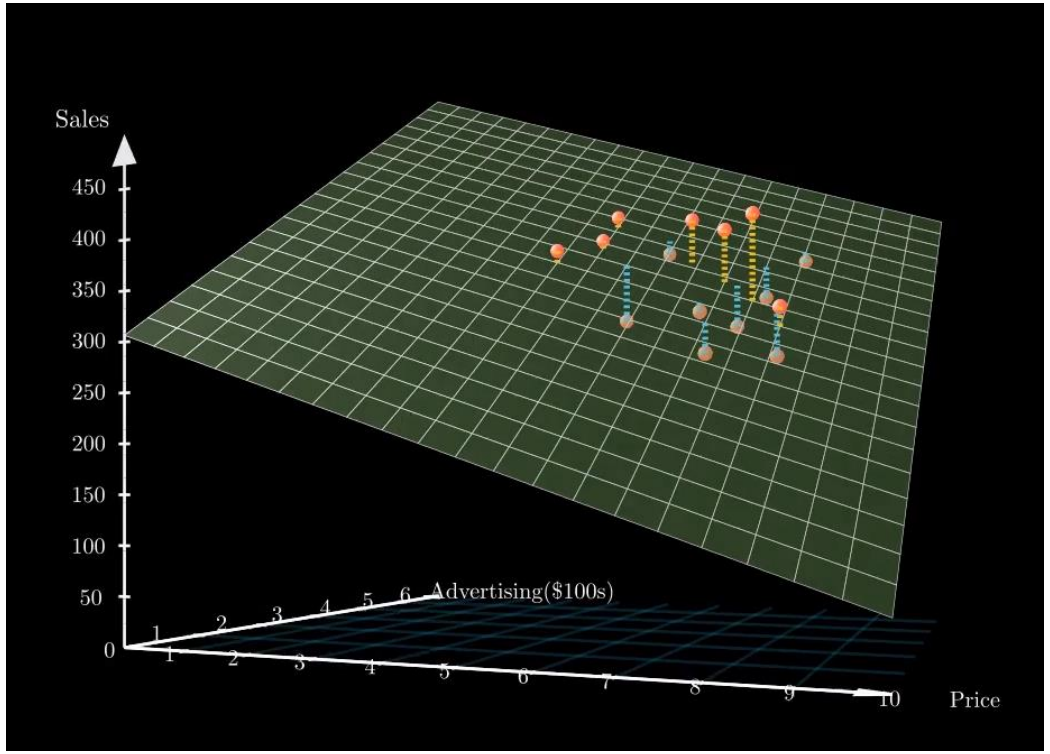


Prediction

$$\boldsymbol{Sales} = \mathbf{306.53} + (\mathbf{-24.98})Price + (\mathbf{74.13})Advertising$$

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.5	3.3
2	460	7.5	3.3
3	350	8	3
4	430	8	4.5
5	350	6.8	3
6	380	7.5	4
7	430	4.5	3
8	470	6.4	3.7
9	450	7	3.5
10	490	5	4
11	340	7.2	3.5
12	300	7.9	3.2
13	440	5.9	4
14	450	5	3.5
15	300	7	2.7

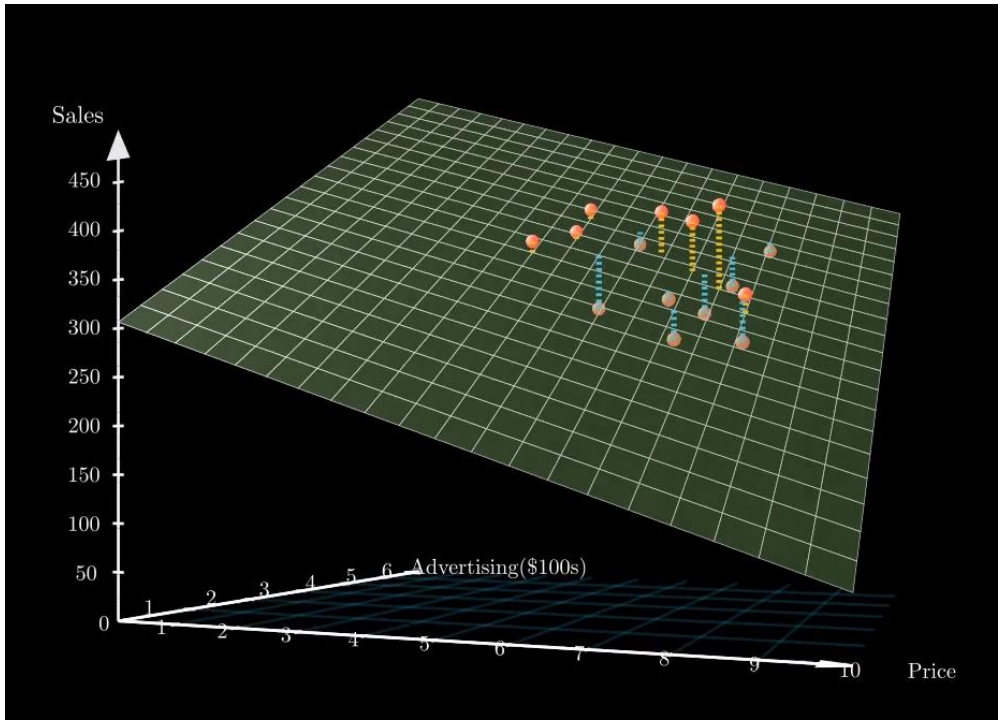
What is the predicted sales when price is $7 and the advertising budget is $270?



Which model is better?

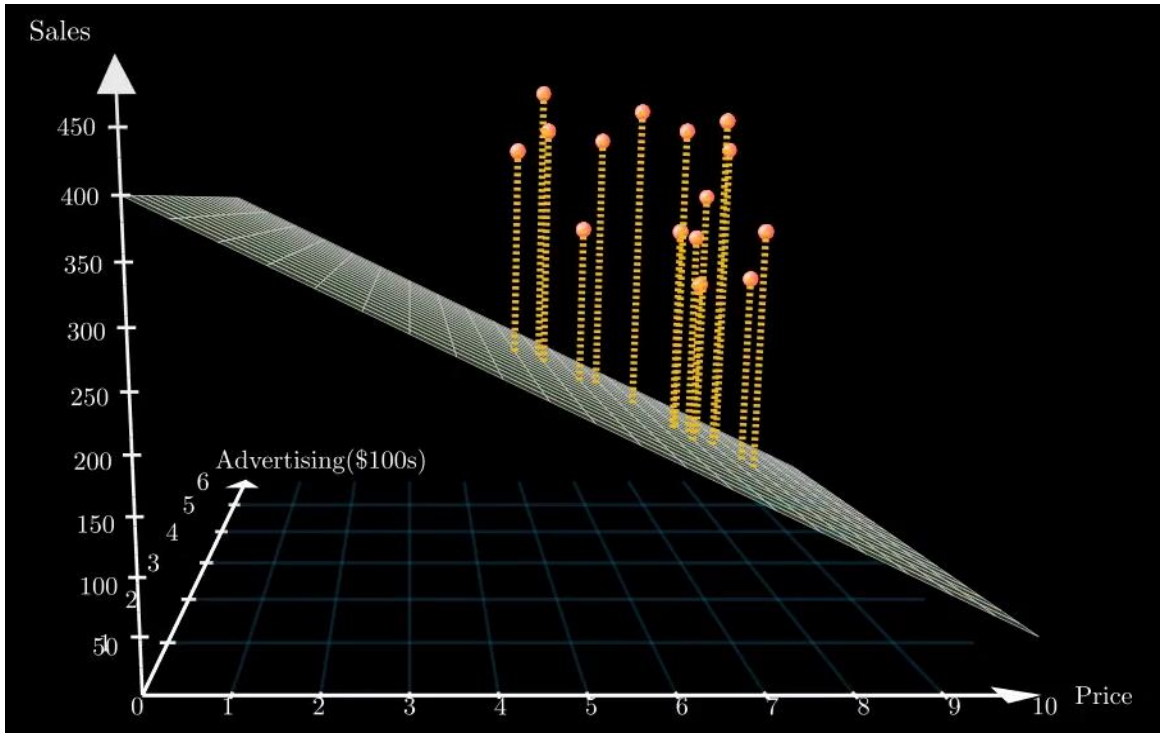
Model 1

$$\boldsymbol{Sales} = \mathbf{306.53} + (\mathbf{-24.98})Price + (\mathbf{74.13})Advertising$$



Model 2

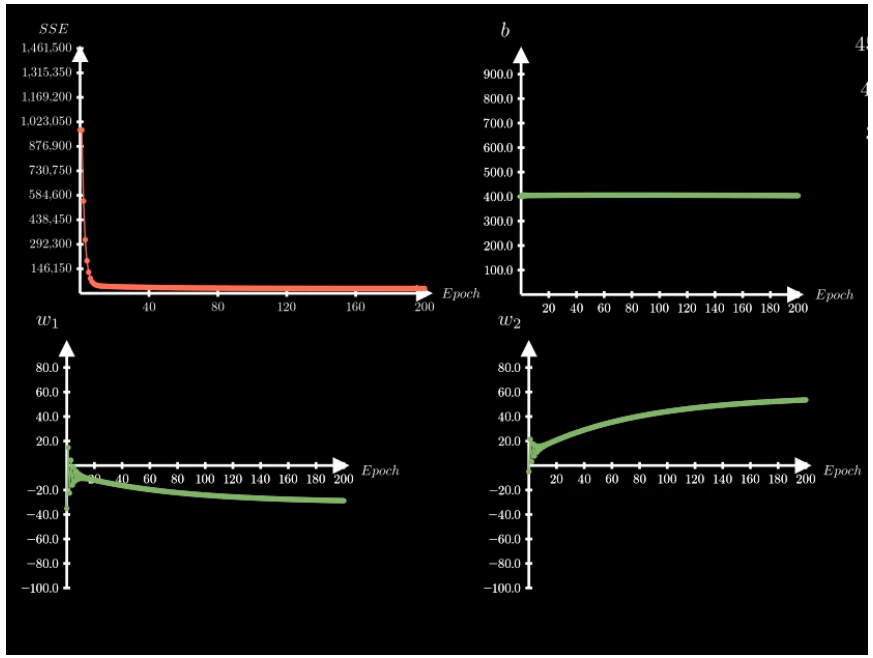
$$\boldsymbol{Sales} = \mathbf{400} + (\mathbf{-35})Price + (\mathbf{-5})Advertising$$



Gradient Descent for Multiple Regression Models

- Similar to simple linear regression, we can use Gradient Descent to find the best model with minimum SSE.
- Optimize three parameters at the same time
  - b, w_1, w_2

$$\boldsymbol{Sales} = \boldsymbol{b} + \boldsymbol{w_1} Price + \boldsymbol{w_2} Advertising$$



Interpreting the coefficients in linear regression

$$\boldsymbol{Sales} = \mathbf{306.53} + (\mathbf{-24.98})Price + (\mathbf{74.13})Advertising$$

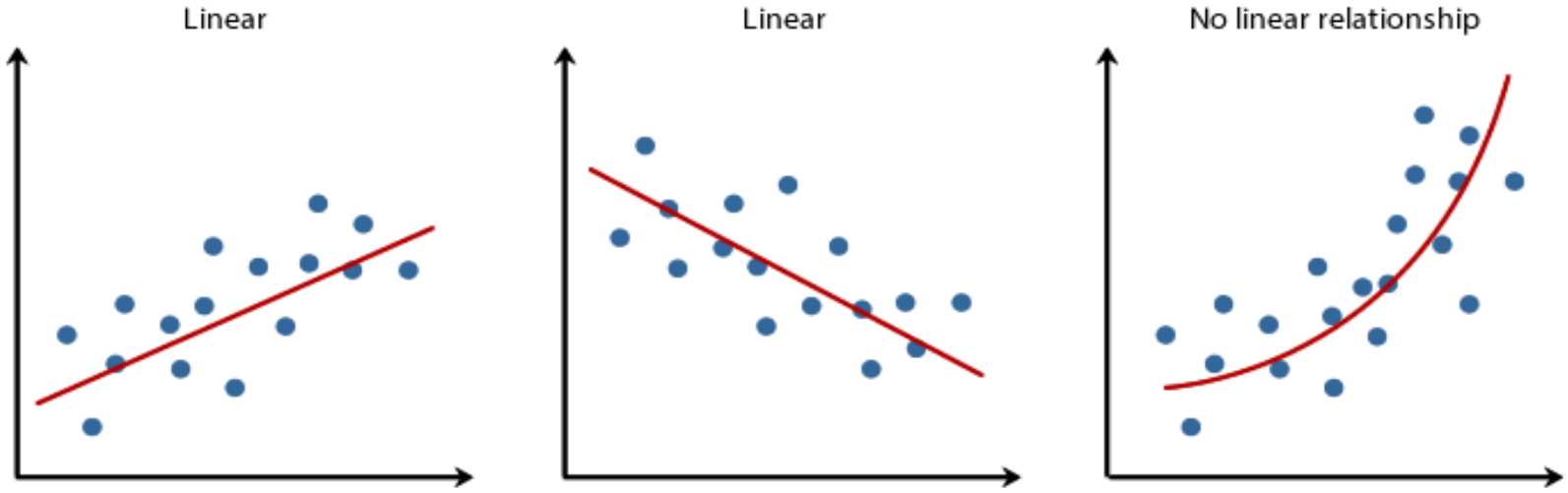
- The sign of the slope of each independent variable indicates whether it has positive/negative correlation with the dependent variable
  - An increase in advertising is associated with an increase in sales.
  - The increase in price may decrease the sale.
- The magnitude of each independent variable indicates its effect on the dependent variable, when we keep the other independent variable constant ("controlling for" the other variables).

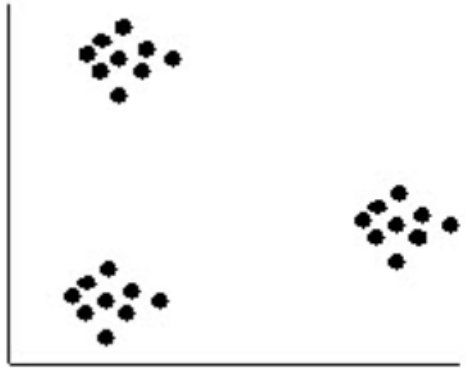
Parameters and hyperparameters in machine learning

$$y = w_1 x_1 + w_2 x_2 + b$$

- Regression models are defined by their weights and bias—the parameters of the model.
  - Any quantity that the model creates or modifies *during* the training process is a parameter.
- We can twist many other knobs before training a model
  - E.g. learning rate, the number of epochs
  - Any quantity that you set *before* the training process is a *hyperparameter*.

Linear regression may not be the always appropriate





There are three distinct groups. Using a line to summarize the data will be misleading!

You should plot the data to visualize the data before using regression.

References

- Linear Regression: A friendly introduction
  - https://www.youtube.com/watch?v=wYPUhge9w5c
- Luis Serrano, Grokking Machine Learning, Manning Publications.
  - Ch. 3, Ch.4