Linear Regression (LR)

- We are given this insurance.csv dataset.
- We want to do some LR on this dataset.
- For example. Can we predict how much a person has to pay for his/her insurance based on their BMI?

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

How to look at dataset

(input) Features, variable, dimension, attribute

target output,

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

rows
= observations/
Samples

ith observation

Revision of Linear Algebra

x , $x = 5$: scalar

$\mathbf{X}$, $\mathbf{X} = [1 \quad 2 \quad 3]$ → $\mathbf{X}$ is a 1×3 row vector

(3 col; row × col; 1 row)

$\mathbf{X} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ (column) vector

$\mathbf{X}$ is a 3×1 column vector

(3 row; row × col; 1 col)

$\mathbf{X}$ is a 3×3 matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 10 & 100 \\ 2 & 20 & 200 \\ 3 & 30 & 300 \end{bmatrix}$$

(3 col; 3 row)

Continue

$$X_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 10 \\ 20 \\ 30 \end{bmatrix} \qquad X_3 = \begin{bmatrix} 100 \\ 200 \\ 300 \end{bmatrix}$$

$$X = \begin{bmatrix} X_1 & X_2 & X_3 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 10 & 100 \\ 2 & 20 & 200 \\ 3 & 31 & 300 \end{bmatrix}$$

How when we see a dataset, lets imagine it as vectors and matrices

	X_1 age	X_2 sex	X_3 bmi	X_4 children	X_5 smoker	X_6 region	y charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

X_i : the i^{th} feature/ attributes of the dataset. X_i is a col vector

y : a col vector of the target output

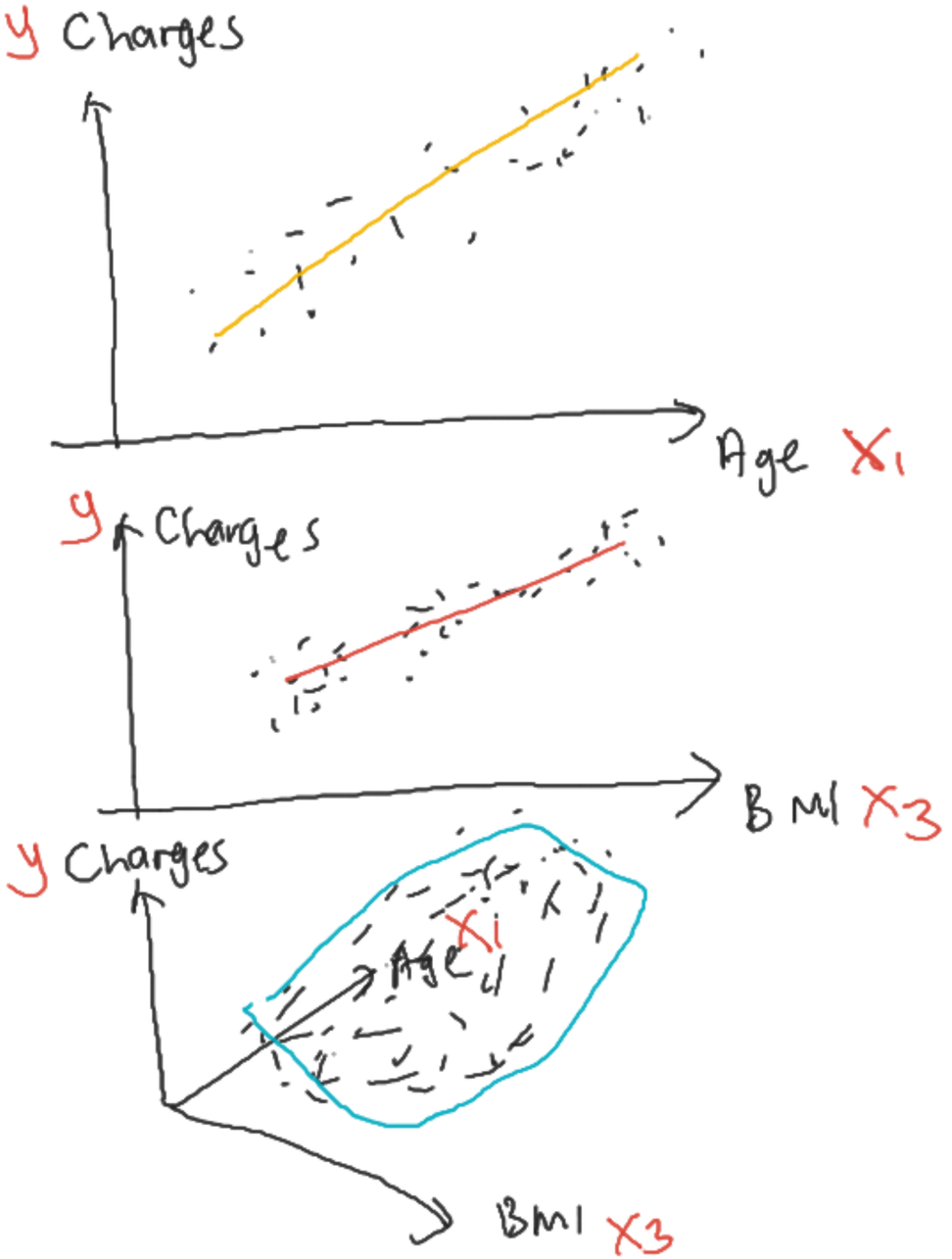
$X = \begin{bmatrix} X_1 & X_2 & \cdots & X_6 \\ 19 & \vdots & & \vdots \\ 18 & & & \\ \vdots & & & \\ \vdots & & & \\ 1337 & & & \end{bmatrix}$ — we have 6 features.

But how many rows of observation/samples do we have?

cols = 6
rows = 1338

X is a 1338×6 matrix #

Our 'intuition'



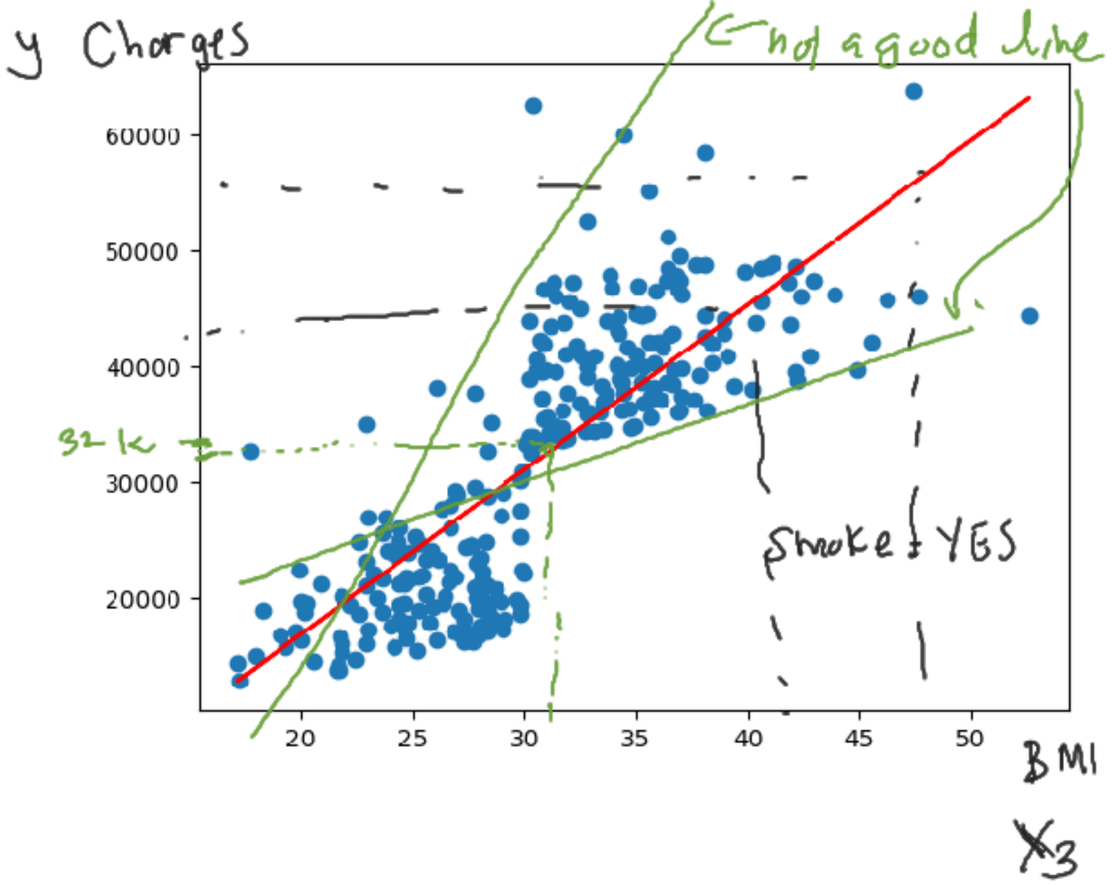
	X_1	X_2	X_3	X_4	X_5	X_6	y
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

X: features / attributes, independent variable

y: target output, dependant variable

Linear Regression with one variable

	X_3 BMI	y Charges
1	29	50000
2	30	60000
⋮		
1338	25	70000
1337	31	?



- Linear regression is to find a linear line that fits the dataset.
- We can make predictions for unseen observation. For example
  what is the insurance charges for a person with BMI = 31?

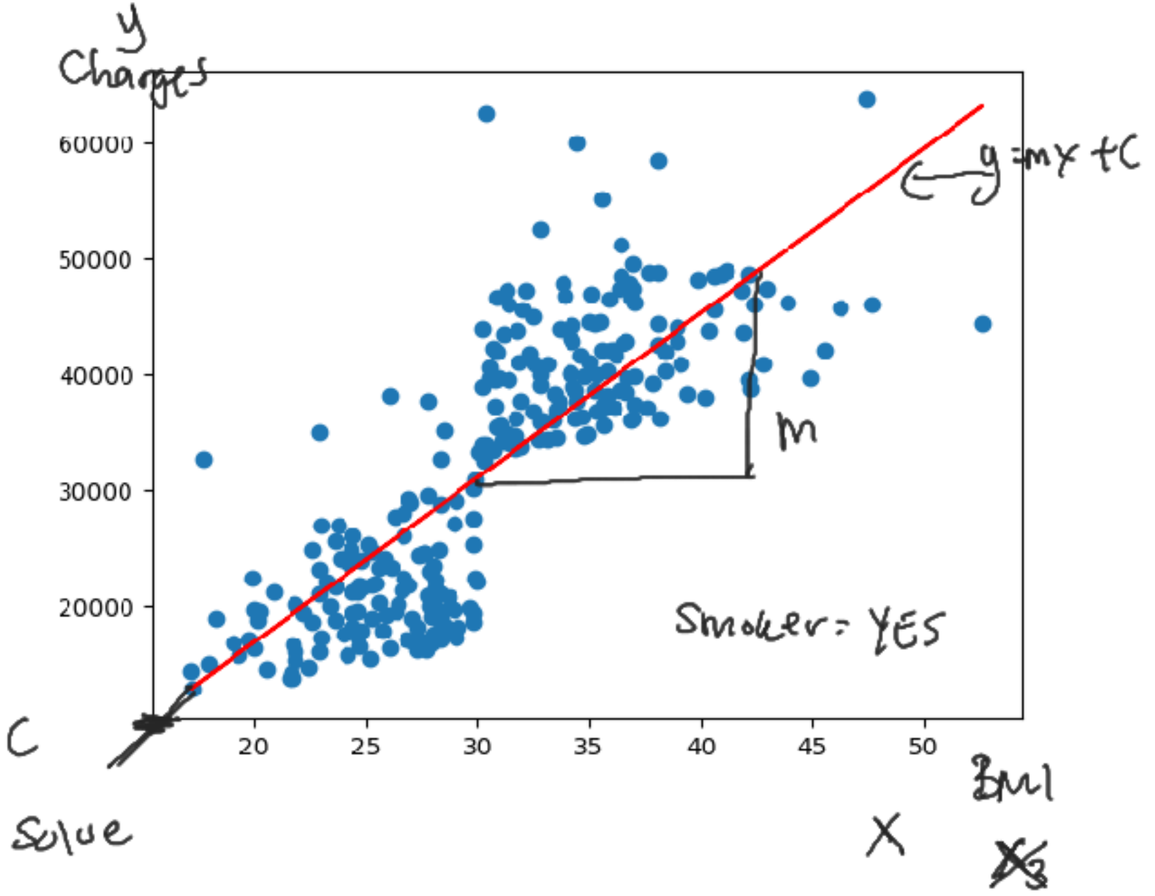
How do we find the best line (red) that fits the data?

$y = mx + C$

$h(x) = \theta_0 + \theta_1 x$

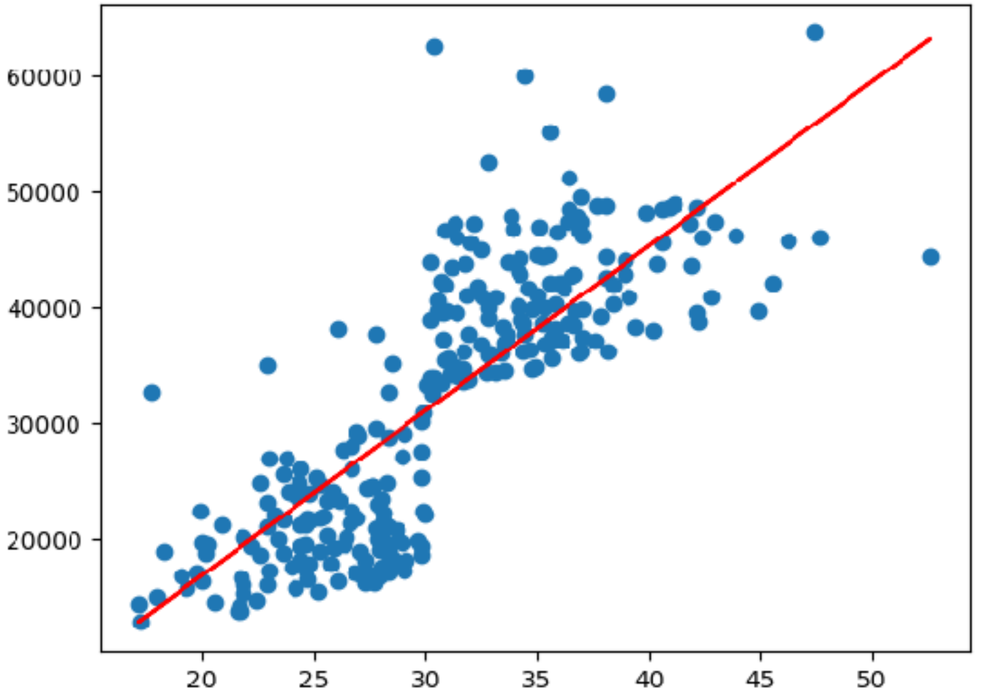
C: intercept m: slope

- So our job (for one variable) is to solve for m and C.



How do we find the 'best' m and c ?

$$y = mx + c$$



Two methods

A) Solve using equation

B) Use gradient descent

A) Solve using equation

Ordinary Least Square method (from Stats)

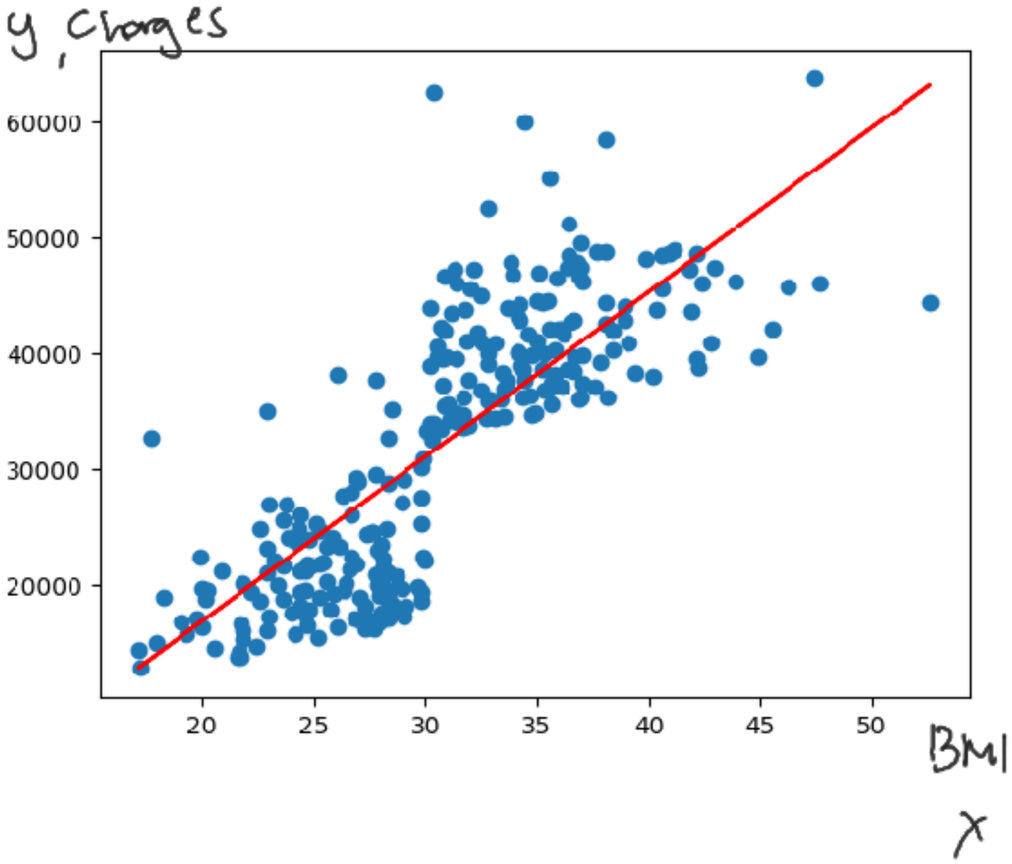
c, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x};$

$y = mx + c$

$y = \beta_0 + \beta_1 x$

m, $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$

← what is this?



From our dataset

index	X_3 BMI	y Charges
1	19	
2	23	
⋮		
j	$x_i = 25$	$y_i = 20000$

n = 1338

$\bar{x}$: average of x = $\frac{1}{N}\sum_{i=1}^{N} x_i$

$\bar{y}$ = " of y = $\frac{1}{N}\sum_{i=1}^{N} y_i$

3.) Use gradient Descent

(gradient of the cost function)

We need to define a cost function

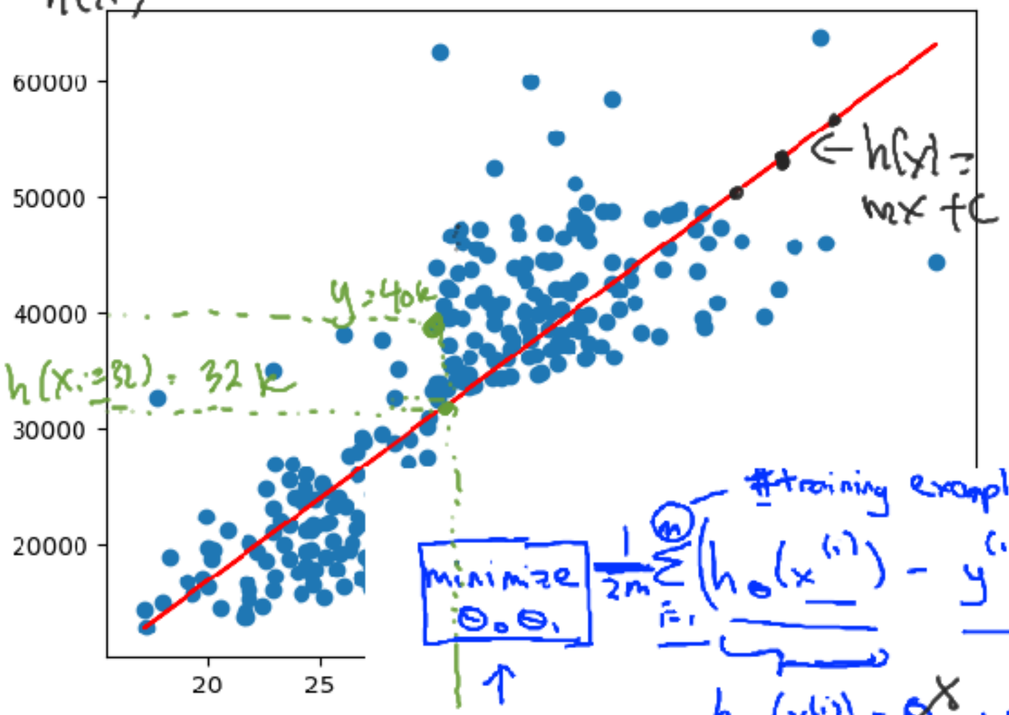
$h(x) = mx + c$ ← our red line

● : our data ● : our prediction (line)

$(h(x^i) - y^i)^2$

∴ So our cost function is

$$J(c,m) = \frac{1}{2N}\sum_{i=1}^{N}(h(x^i) - y^i)^2$$



indx	X BMI	y Charges	h(x)
1	27	40000	
2	31	45,000	
⋮			
i	$x^i = 32$	$y_i = 40,000$	32,000
⋮			
N = 1338	25	38 000	

$J(\Theta_0,\Theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$

Minimize $J(\Theta_0,\Theta_1)$ ← Cost function
Θ_0,Θ_1

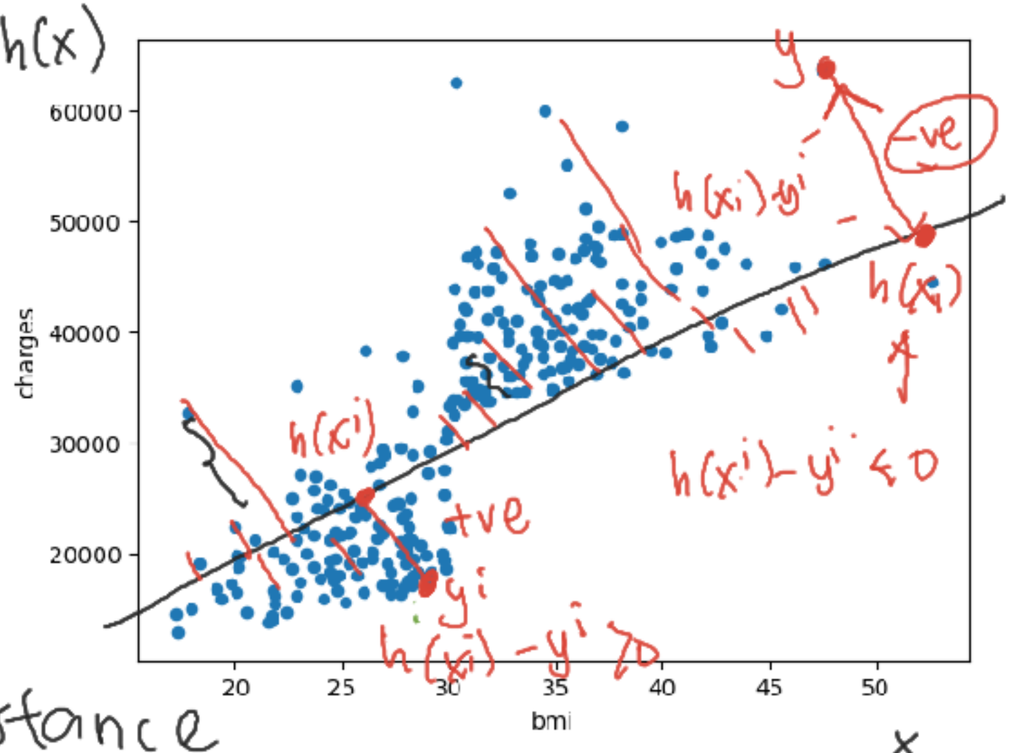
Squared error function

3.) Use gradient Descent

(gradient of the cost function)

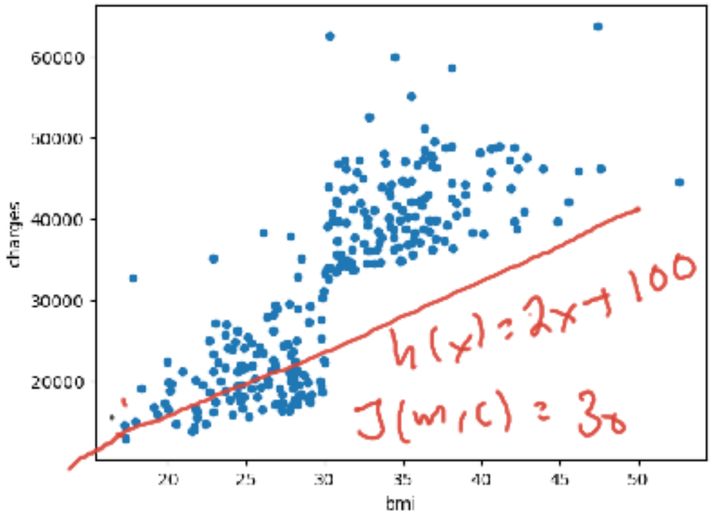
Cost function $h(x) = mx + c$

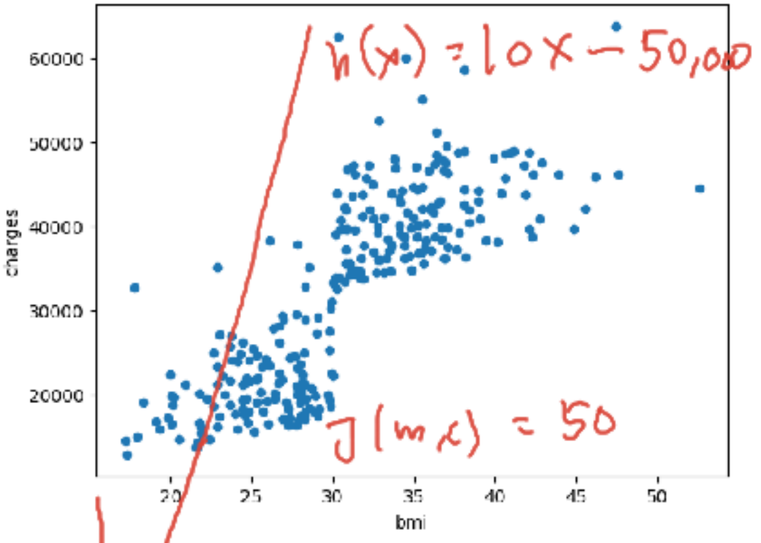
$$J(m,c) = \frac{1}{2N} \sum_{i=1}^{N=1338} (h(x^i) - y^i)^2$$

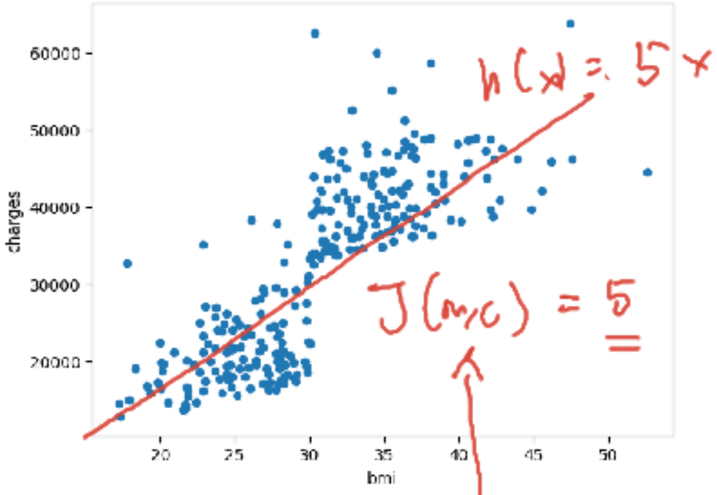


- The cost function calculates the distance between our predictions $h(x)$ and the original data.

- So then we need to find the values of m and c that minimizes the cost function J







m = 5 , c = 0
yields the lowest
error / cost function

How do we find the best m and c that minimizes the cost function?

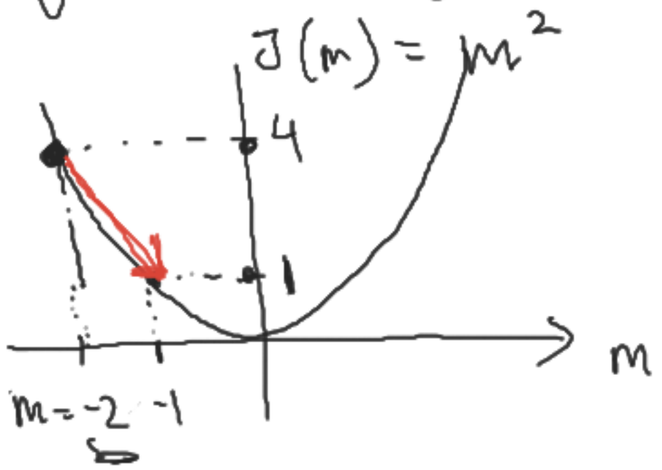
$$J(m,c) = \frac{1}{2N} \sum_{i=1}^{N=1338} \left(h(x^i) - y^i\right)^2$$

What does this cost function looks like?

(I don't know)

How do we find the best m and c that minimizes the cost function?

Let's just imagine our cost function looks like this



α: learning rate/step = 1/4

Update using gradient descent

Derivative/gradient

$$J(m) = m^2$$

$$\frac{dJ(m)}{dm} = 2m$$

THIS IS HOW WE GO DOWN THE GRADIENT OF THE COST FUNCTION

Iteration

$$m_{new} = m_{old} - \alpha \frac{dJ(m)}{dm}$$

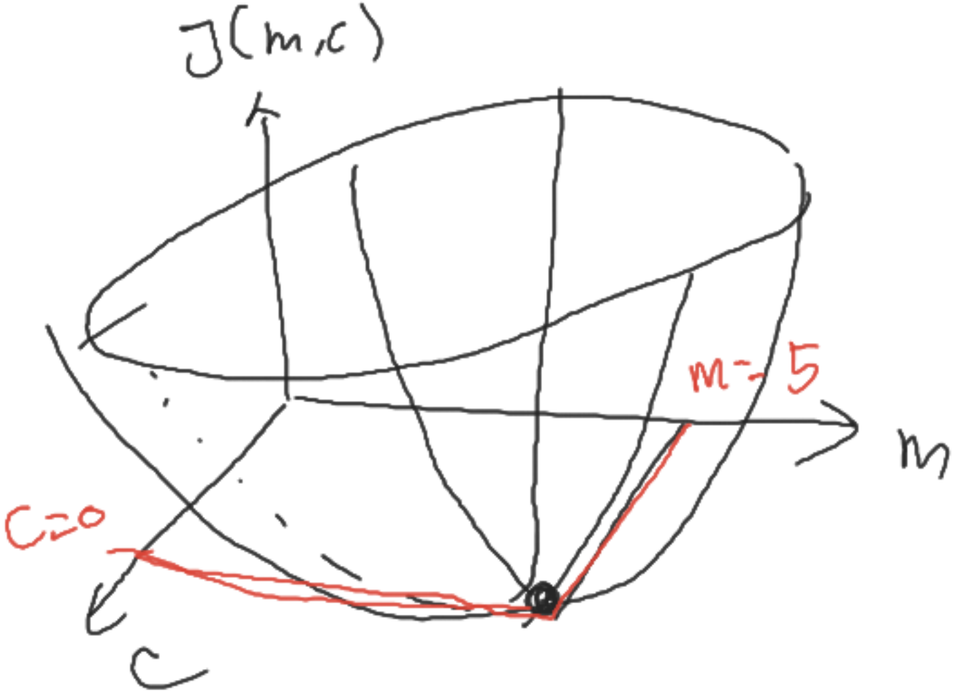
$$= -2 - \frac{1}{4} \cdot 2(-2)$$

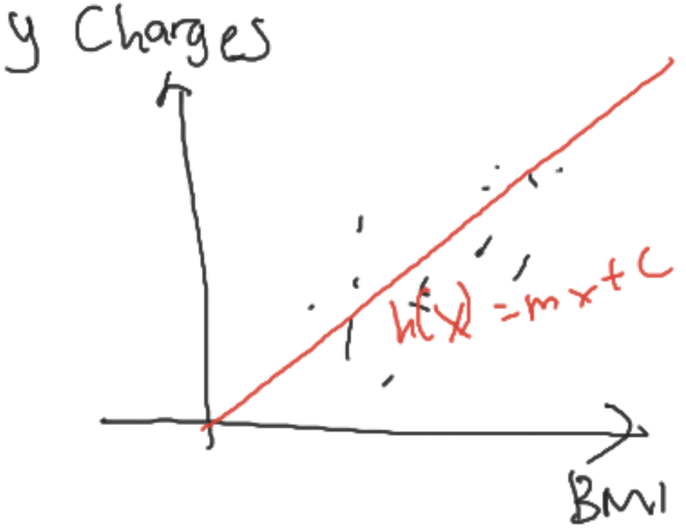
$$= -1$$

$$y_{new} = m_{new}^2 = 1$$

We define the cost function

① $J(m,c) = \frac{1}{2N} \sum_{j=1}^{N=133} (h(x^j) - y^j)^2$





② we find the best value of m and c which minimizes cost fuction J.

③ This is done by going down the gradient/derivative of J

$m_{new} = m_{old} - \alpha \frac{\partial J(m,c)}{\partial m}$

$c_{new} =$