DataStar Machine Learning

ADAX

Modules

01. Introduction
02. **Regression**
03. Classification
04. Ensemble Methods & Cross-Validation
05. Machine Learning Algorithms
06. Regularization Techniques
07. Introduction to Unsupervised ML
08. Dimensionality Reduction Techniques
09. Clustering Techniques
10. Introduction to Natural Language Processing

Session 2: Regression

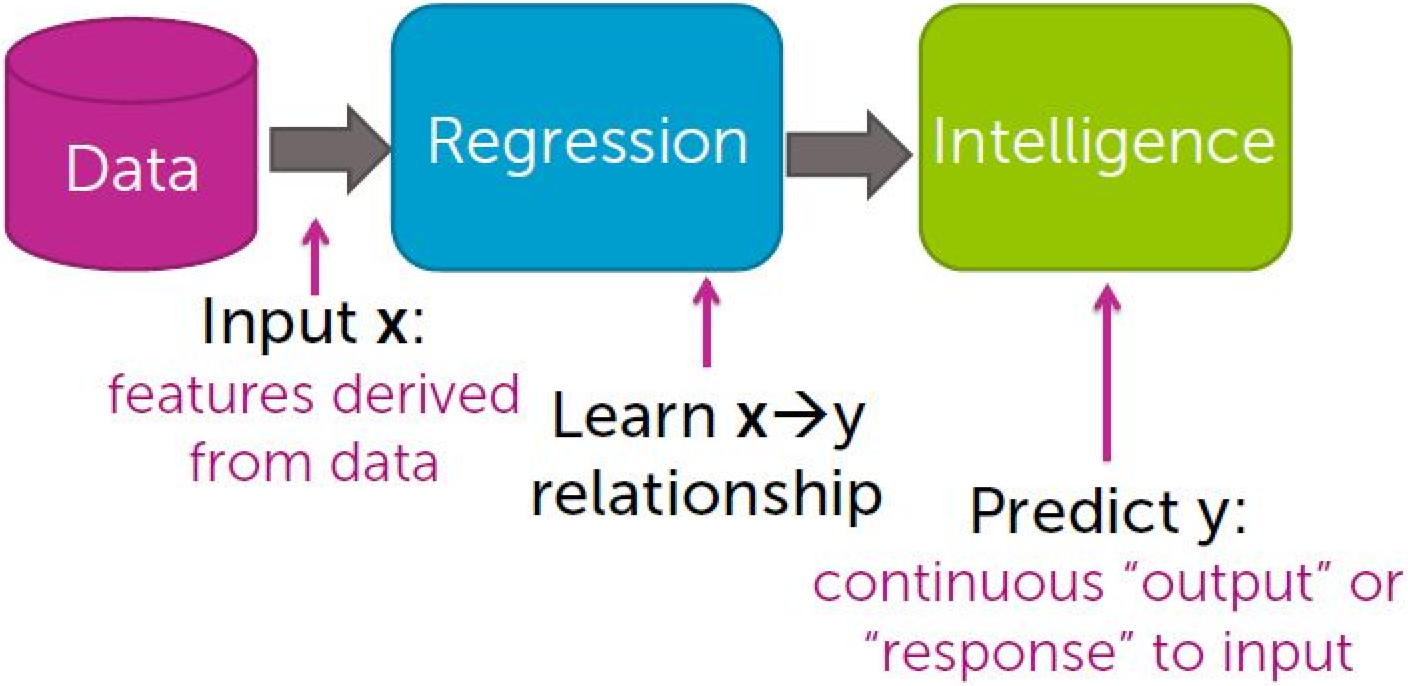
15 Sept 2017

ADAX

What is Regression?

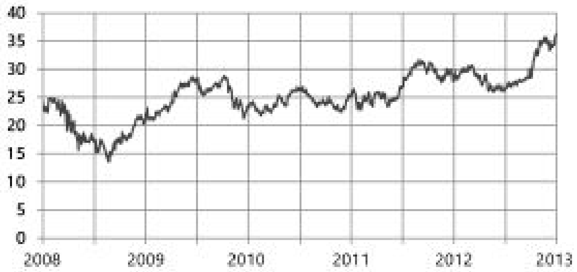
What is Regression?

From features to predictions



Stock prediction

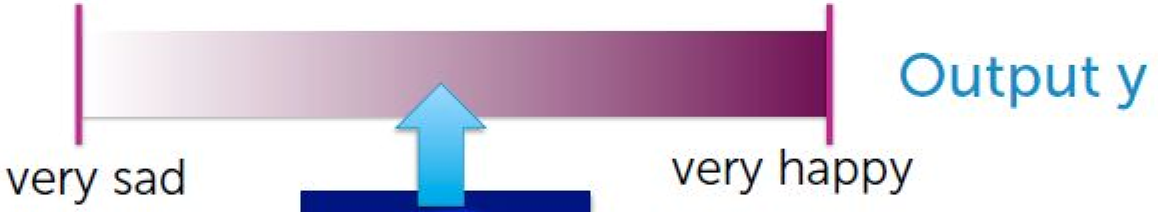
- Predict the price of a stock (y)
- Depends on **x** =
  - Recent history of stock price
  - News events
  - Related commodities

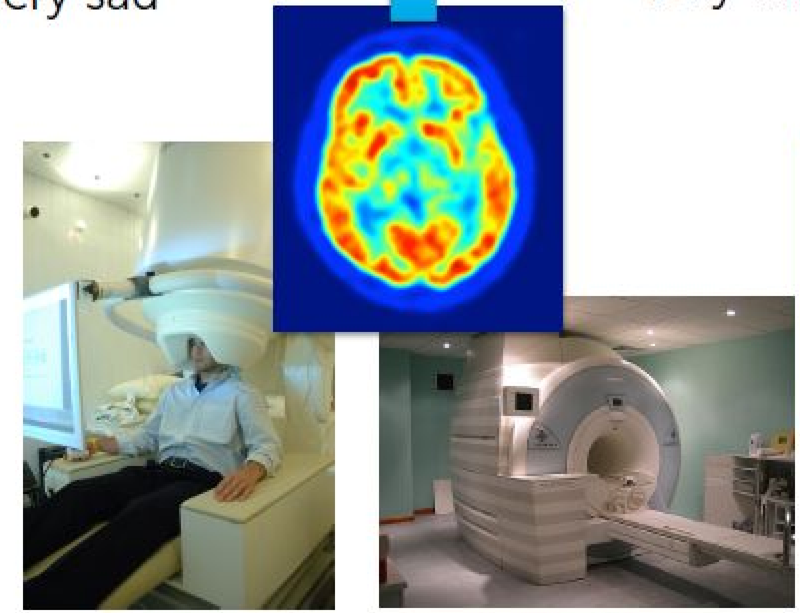


Tweet Popularity

- How many people will retweet your tweet? (y)
- Depends on **x** = # followers,
  # of followers of followers,
  features of text tweeted,
  popularity of hashtag,
  # of past retweets,...

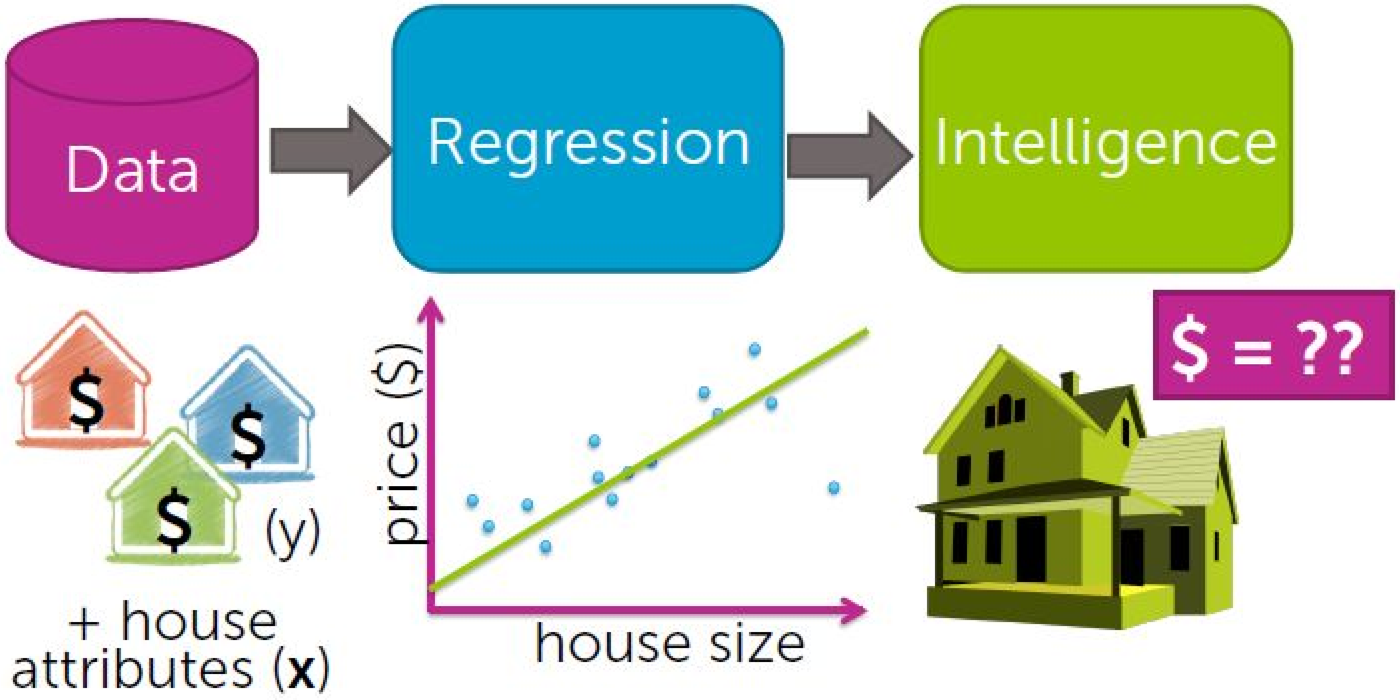
Reading Your Mind





Inputs x are brain region intensities

Case Study: Predicting House Prices



Simple Regression

Linear regression with one input

How much is my house worth?

Look at recent sales in my neighborhood

- How much did they sell for?

Regression fundamentals

Data, Model, Task

Data

input *output*

(x_1 = sq.ft., y_1 = \$)

(x_2 = sq.ft., y_2 = \$)

(x_3 = sq.ft., y_3 = \$)

(x_4 = sq.ft., y_4 = \$)

(x_5 = sq.ft., y_5 = \$)

⋮

Input vs. Output:

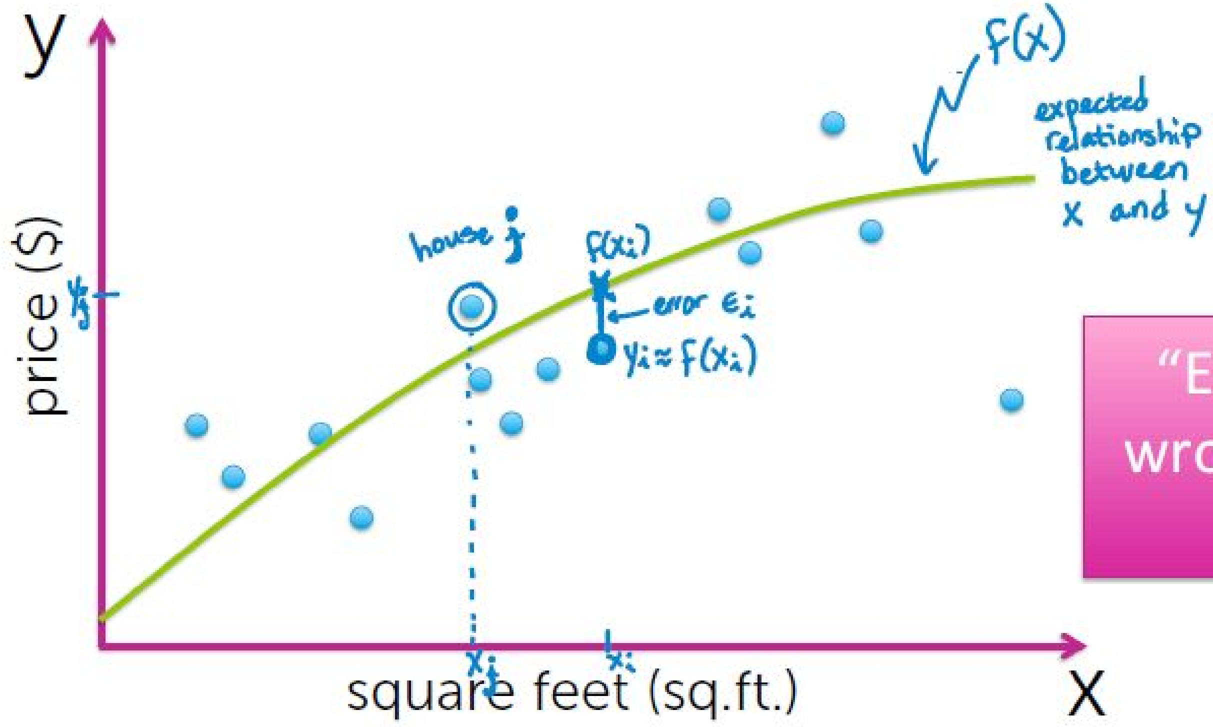
- y is the quantity of interest
- assume y can be predicted from x

Regression fundamentals

Data, Model, Task

Model

How we assume the world works



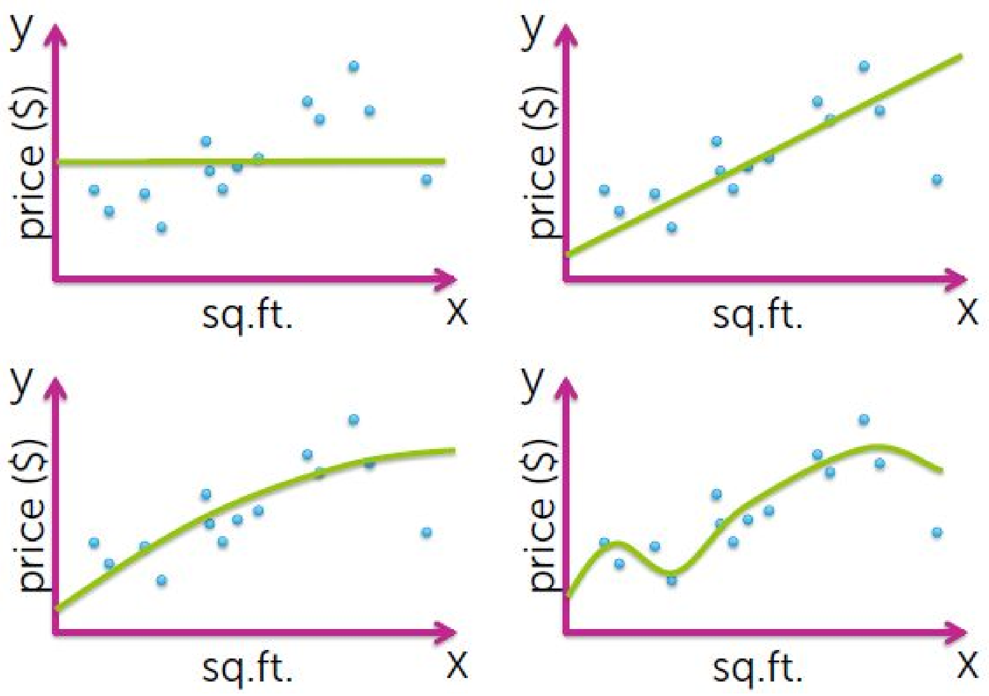
"Essentially, all models are wrong, but some are useful."
George Box, 1987.

Regression fundamentals

Data, Model, Task

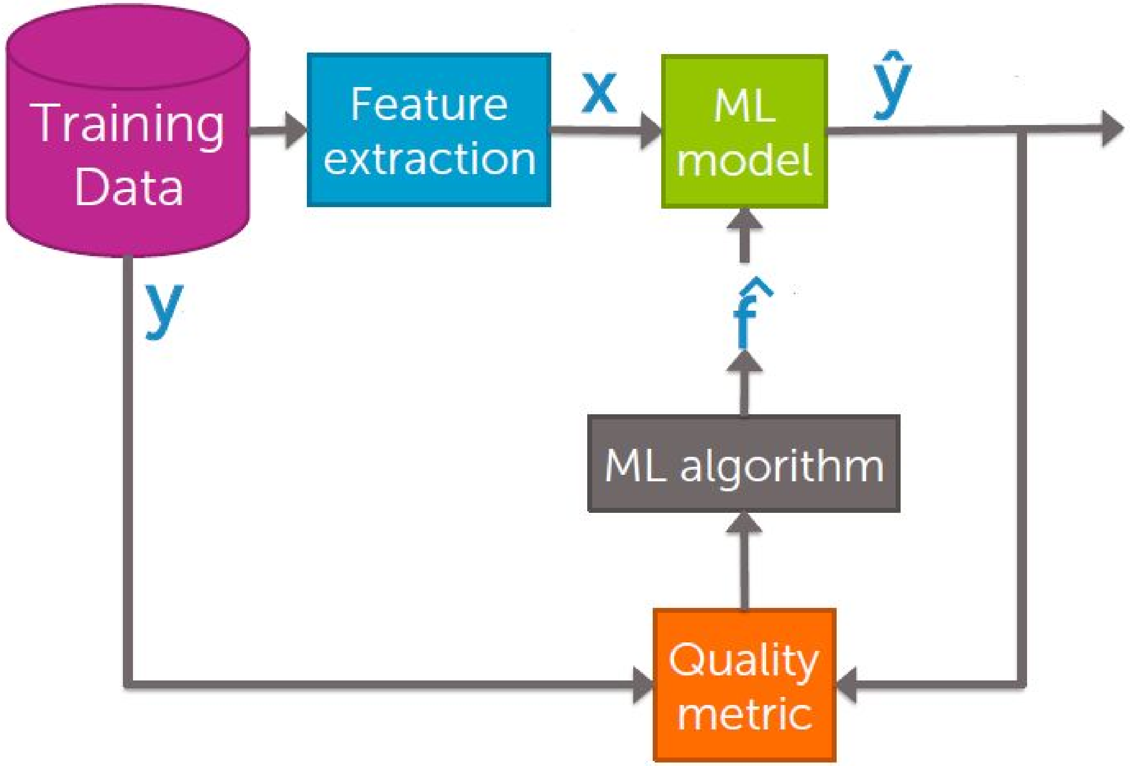
Task

Which model f(x) is the best fit?



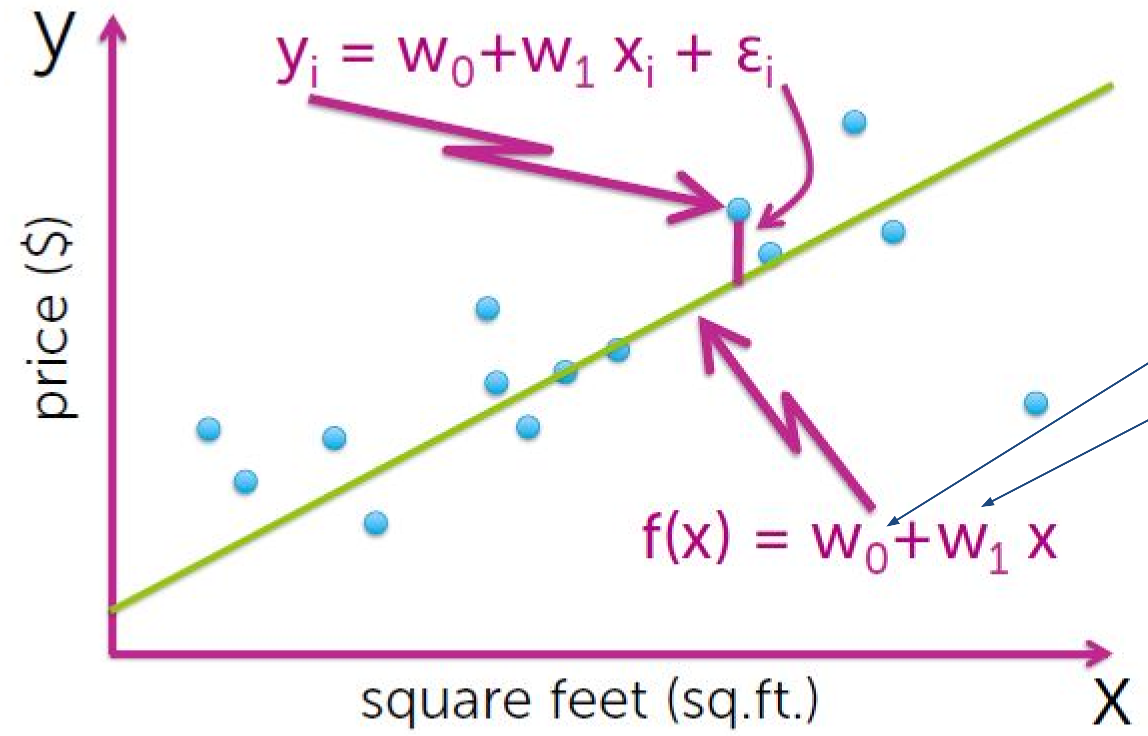
From the model that we estimate, we can do a lot more other tasks

How this works

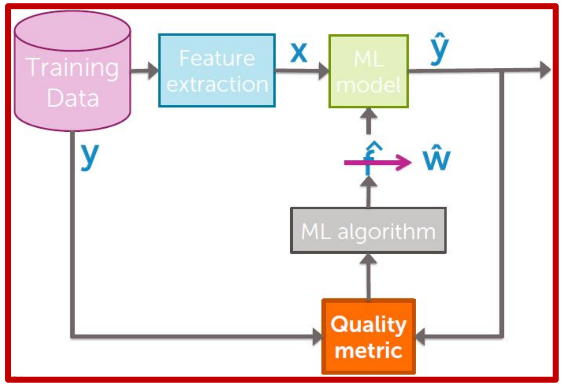


Simple linear regression

What's the equation of a line?



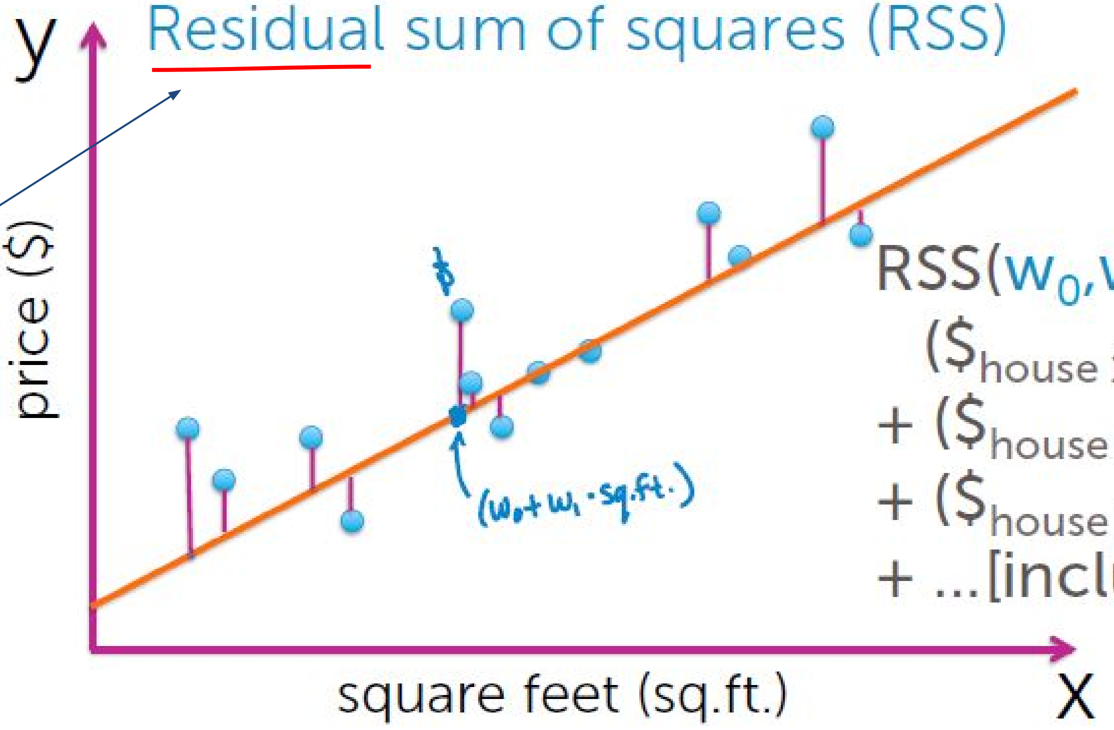
Regression coefficients



"Cost" of using a given line

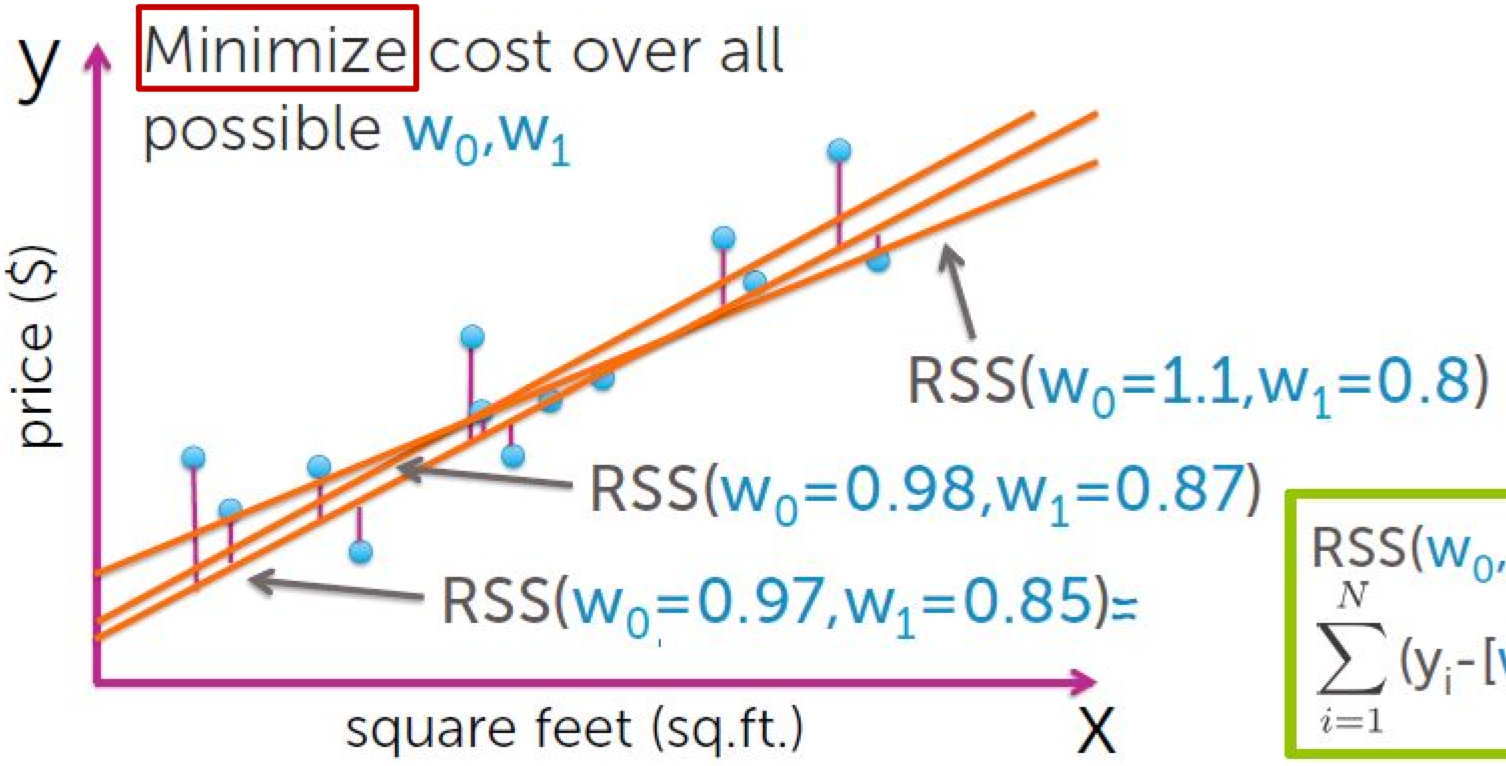
Residual sum of squares (RSS)

What's this residual?



$$
\begin{aligned}
RSS(w_0,w_1) = \; & (\$_{\text{house 1}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 1}}])^2 \\
& + (\$_{\text{house 2}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 2}}])^2 \\
& + (\$_{\text{house 3}} - [w_0 + w_1 \text{sq.ft.}_{\text{house 3}}])^2 \\
& + \ldots\text{[include all training houses]}
\end{aligned}
$$

Find the "best" line



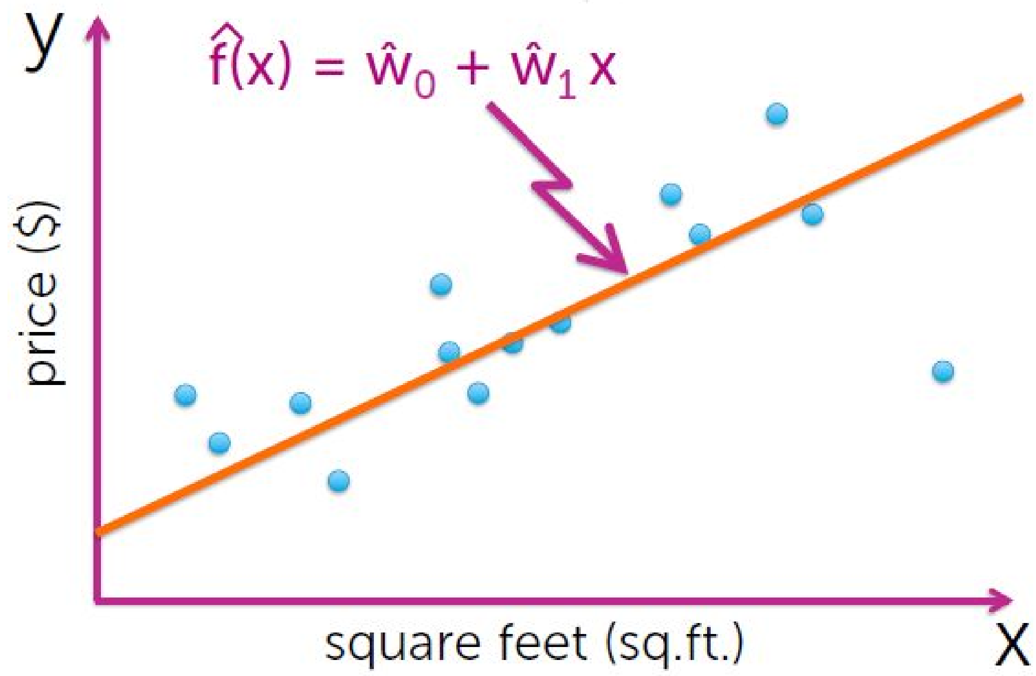
$$\text{RSS}(w_0,w_1) = \sum_{i=1}^{N} (y_i-[w_0+w_1x_i])^2$$

Model vs. Fitted line

Let's say

w_0=-44850

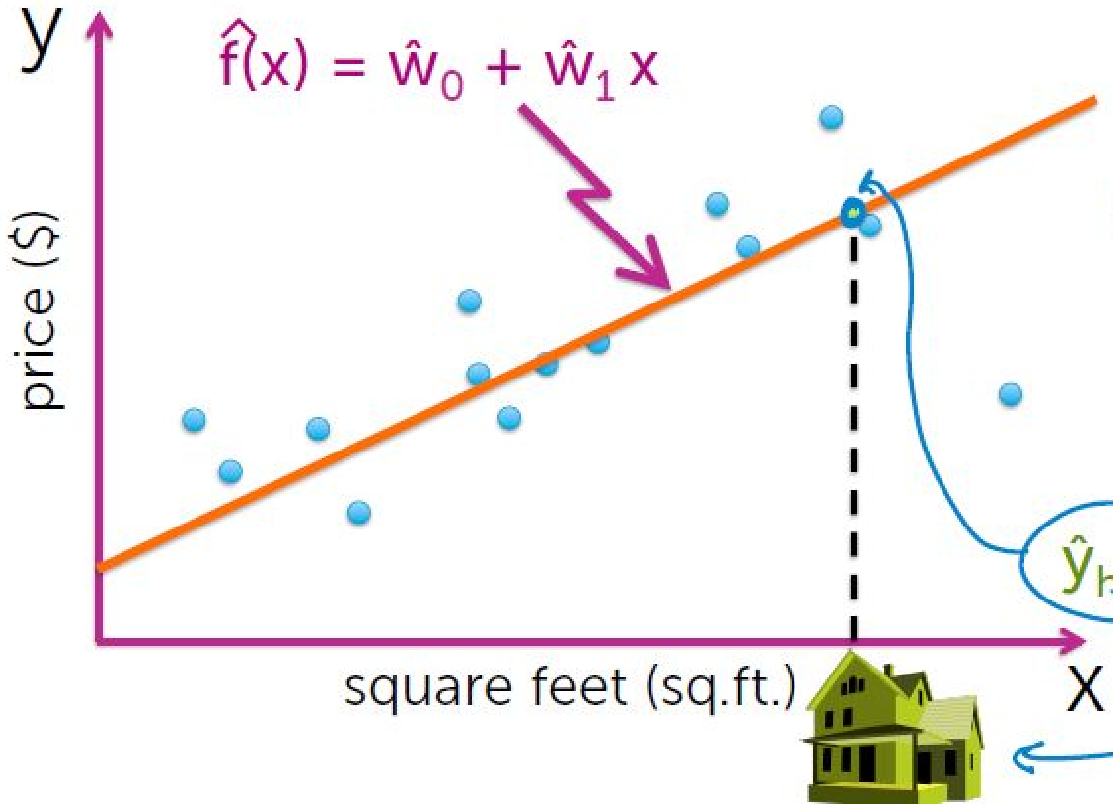
w_1=280.76



Regression model:

$y_i = w_0 + w_1 x_i + \varepsilon_i$

Seller: Predicting house price



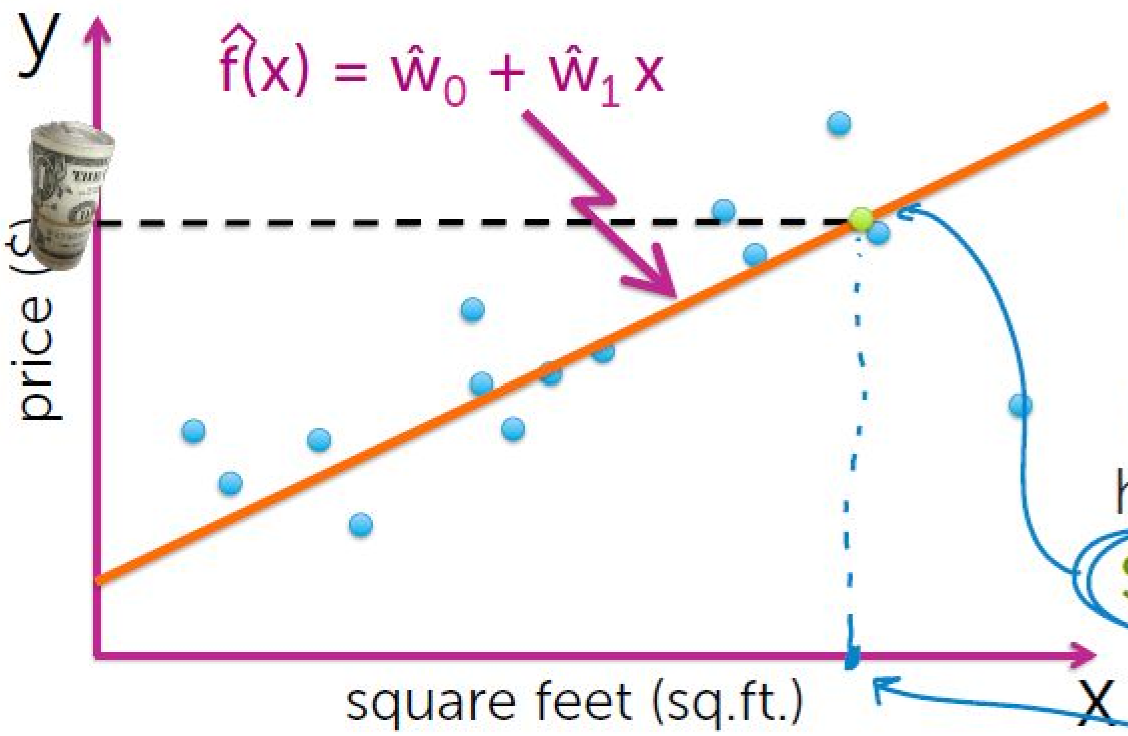
Regression model:

$y_i = w_0 + w_1 x_i + \varepsilon_i$

Best guess of your house price:

$\hat{y}_{house} = \hat{w}_0 + \hat{w}_1 \, sq.ft._{house}$

Buyer: Predicting size of house



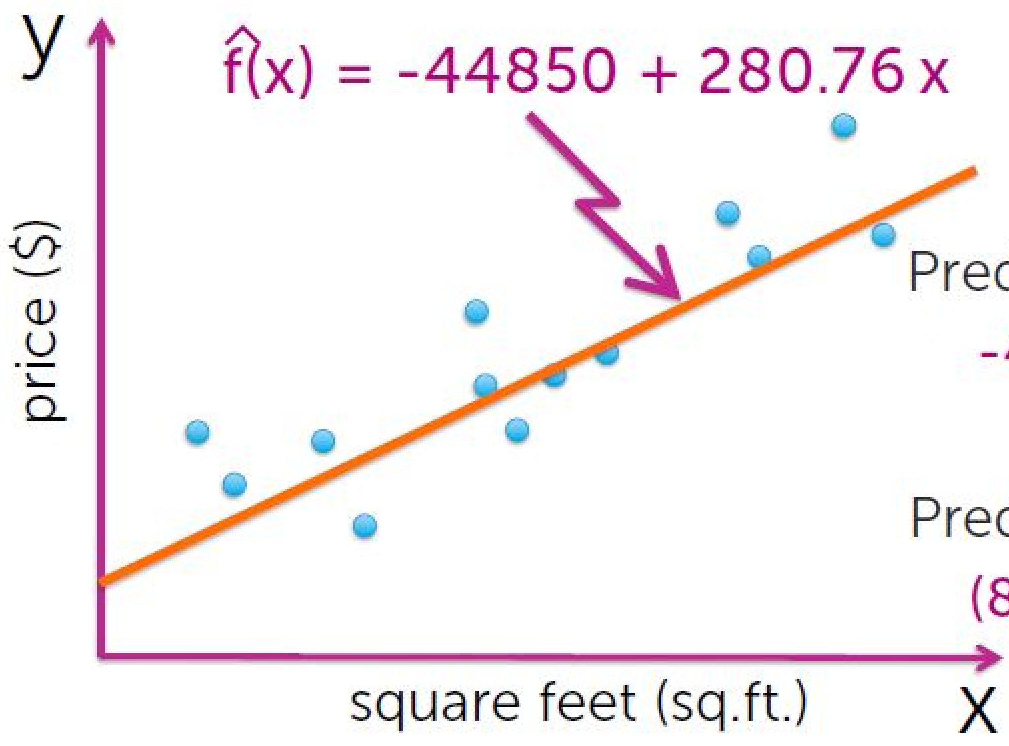
Regression model:

$y_i = w_0 + w_1 x_i + \varepsilon_i$

Best guess of size of house you can afford:

$\$_{\text{in bank}} = \hat{w}_0 + \hat{w}_1 \widehat{\text{sq.ft.}}$

A concrete example



Predict $ of 2,640 sq.ft. house:

$-44850 + 280.76 * 2{,}640$

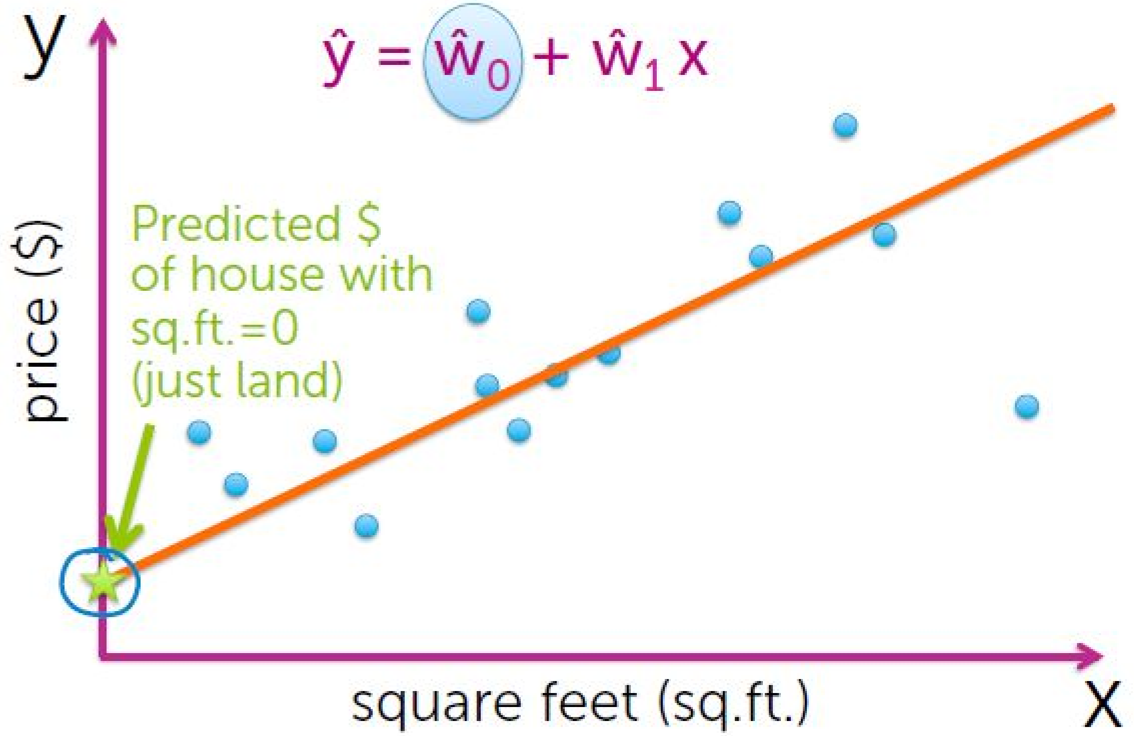
$= \$696{,}356$

Predict sq.ft. of $859.000 sale:

$(859000+44850)/\ 280.76$

$= 3{,}219$ sq.ft.

Making sense of the coefficients

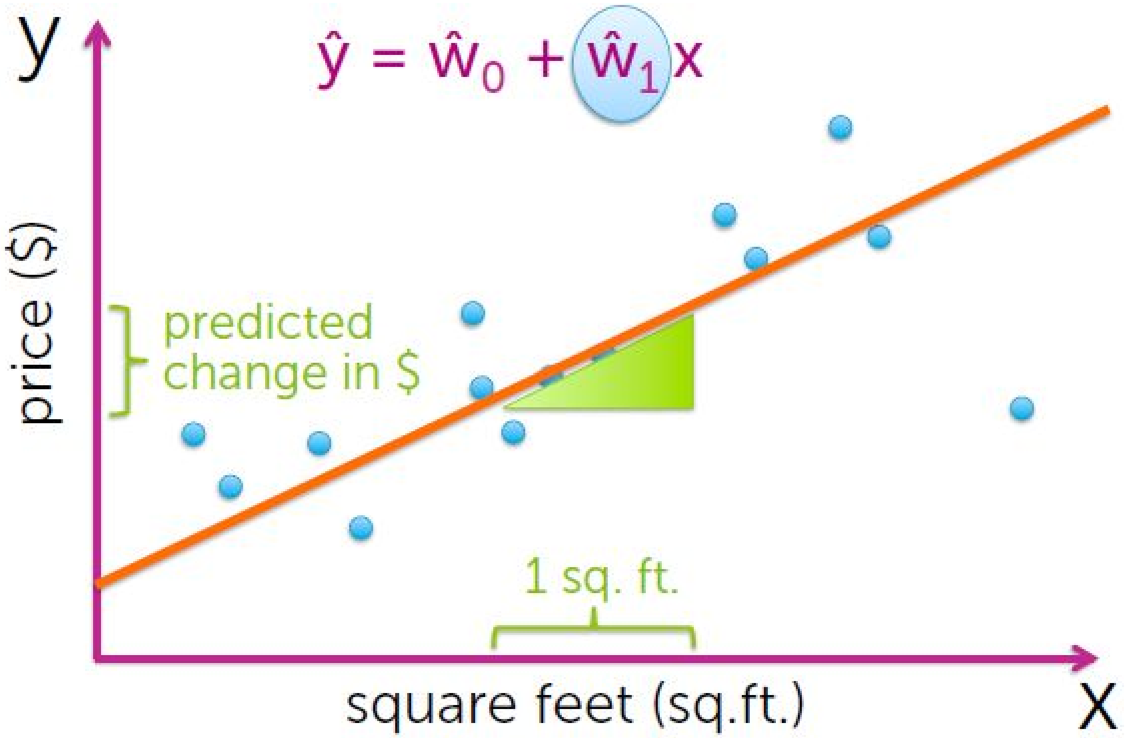


Note:

y = w0 when x=0

Not very meaningful!

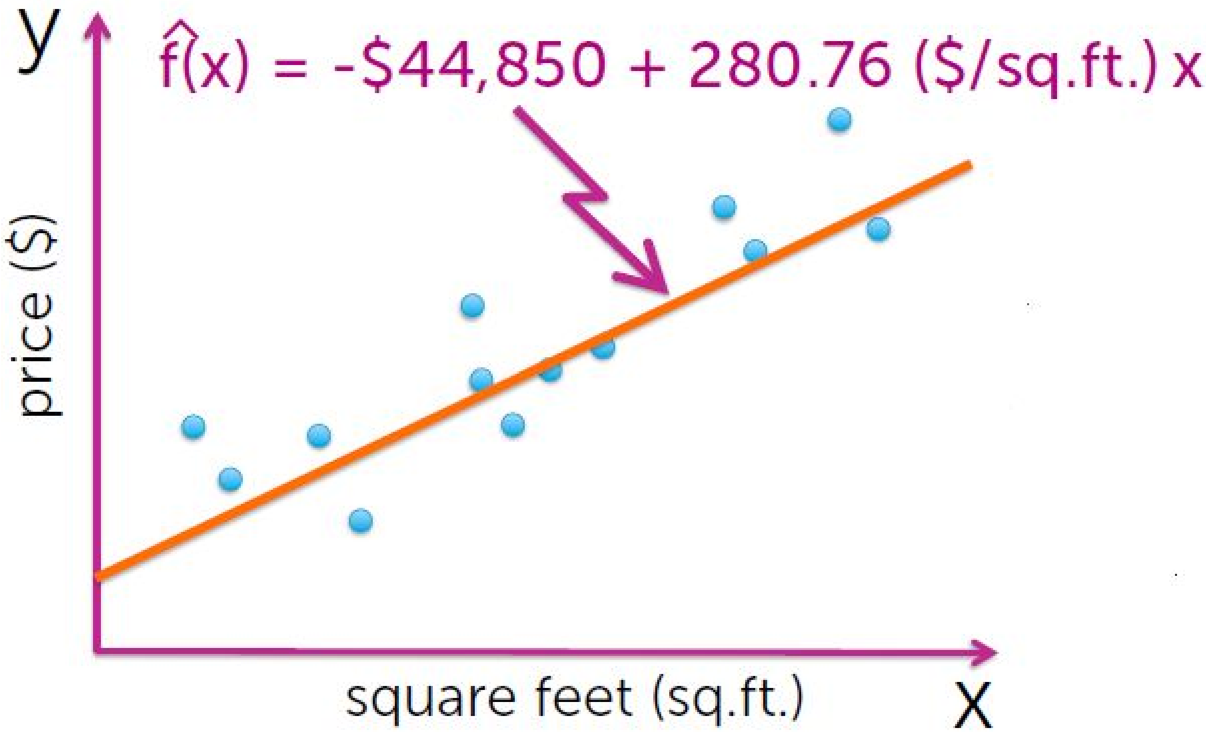
Making sense of the coefficients



w_1

Predicted change in the output per unit change in input

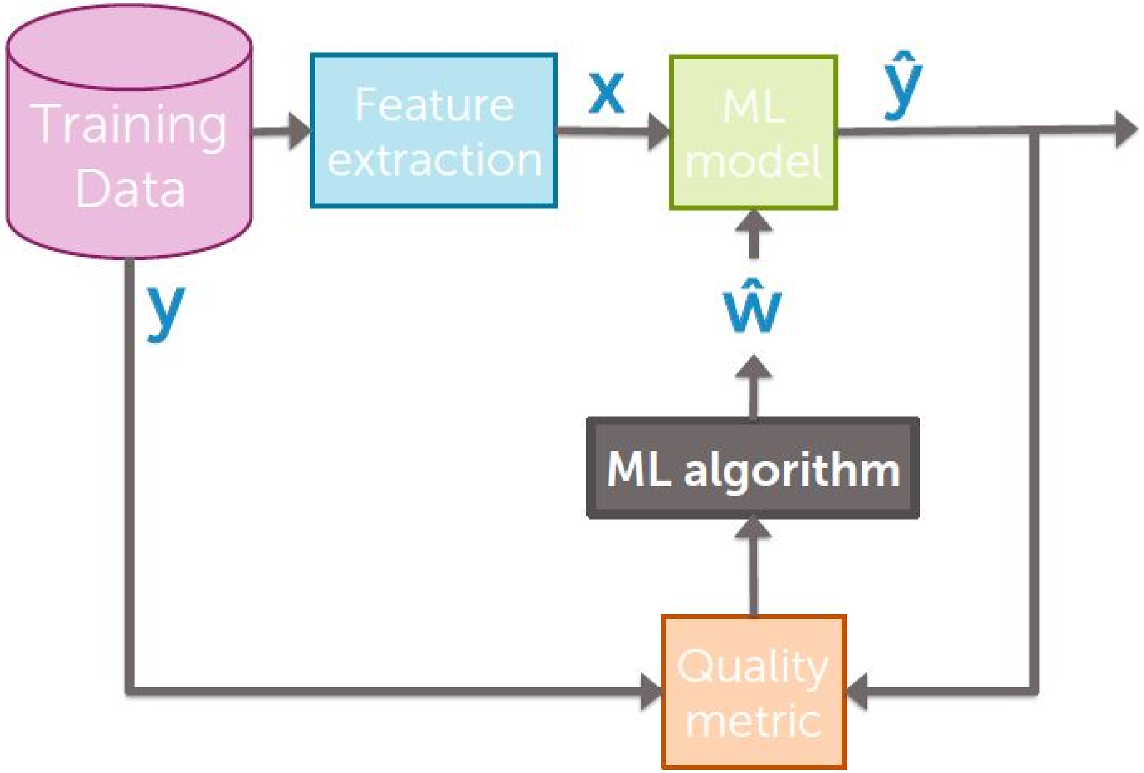
What-ifs



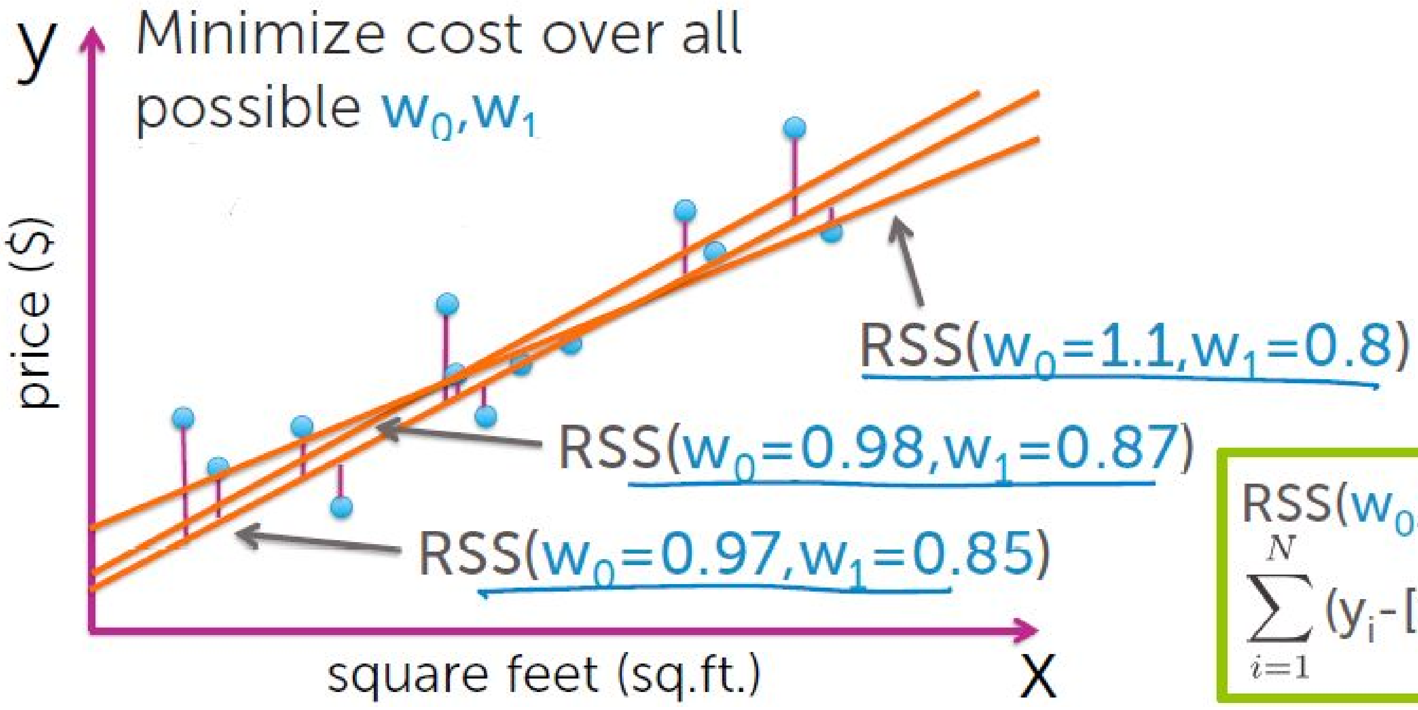
What if…

1. What if house was measured in square meters?
2. Price was measured in RM? Euros?

The Algorithm



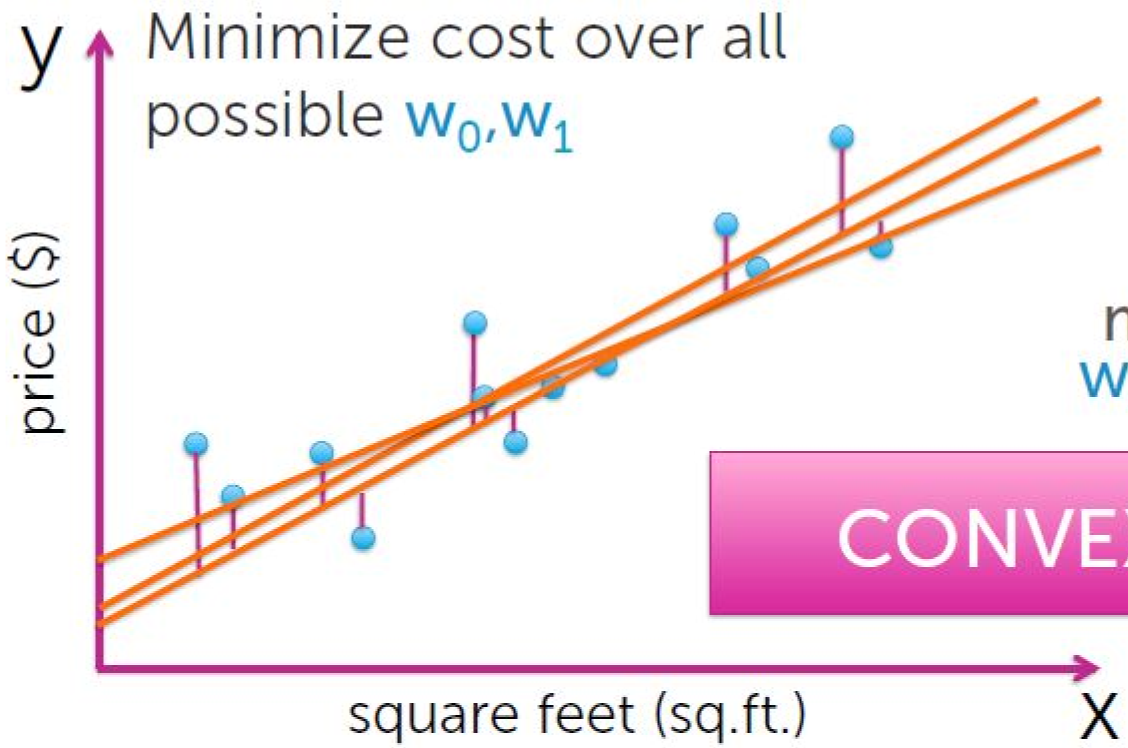
Find "best" line



Recall!

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

Find "best" line

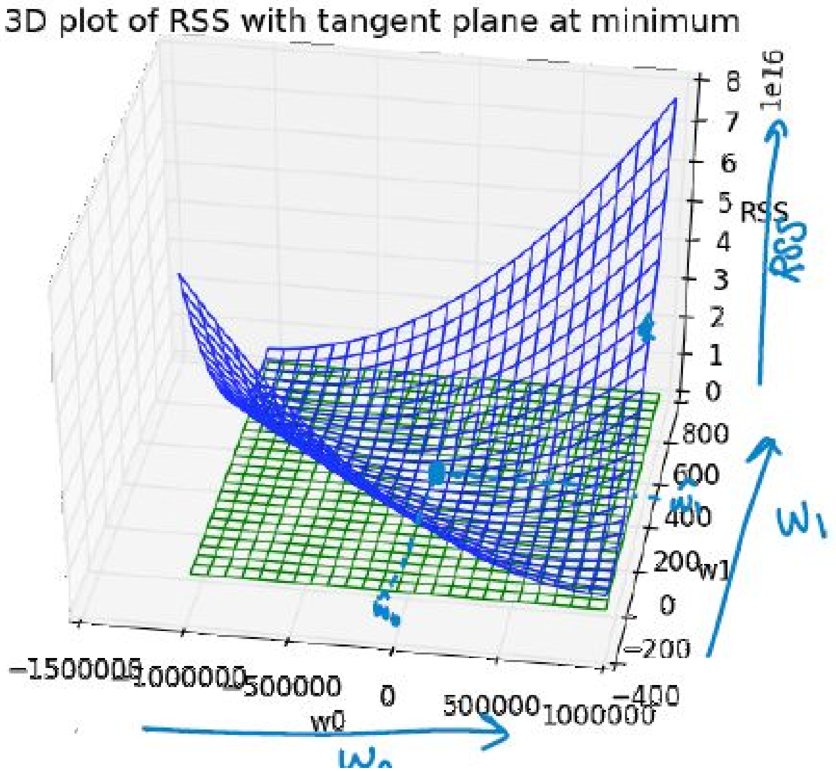


$$\min_{w_0,w_1} \sum_{i=1}^{N} (y_i - [w_0 + w_1 x_i])^2$$

CONVEX

The RSS is convex. What does it mean?

Minimizing a cost function



Minimize function over all possible w_0, w_1

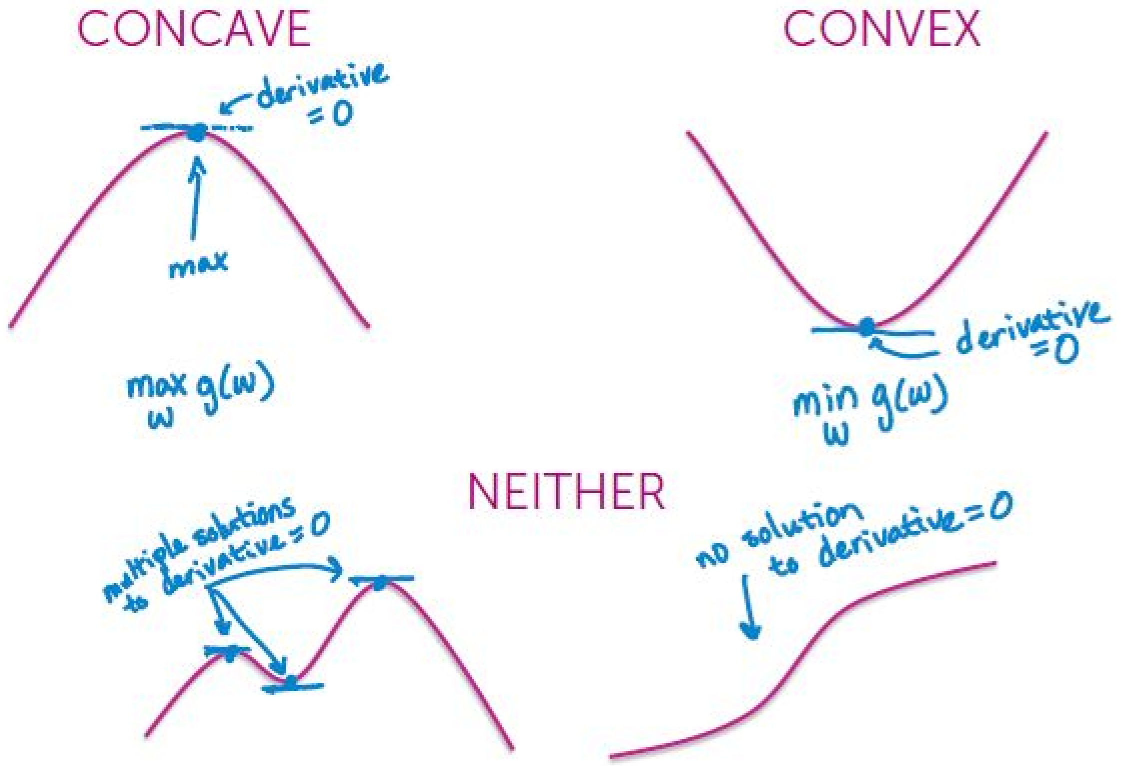
$$\boxed{\min_{w_0, w_1}} \underbrace{\sum_{i=1}^{N} (y_i - [w_0 + w_1 x_i])^2}$$

$RSS(w_0, w_1)$ is a function of 2 variables

Question: Why do we need to MINIMIZE the RSS function?

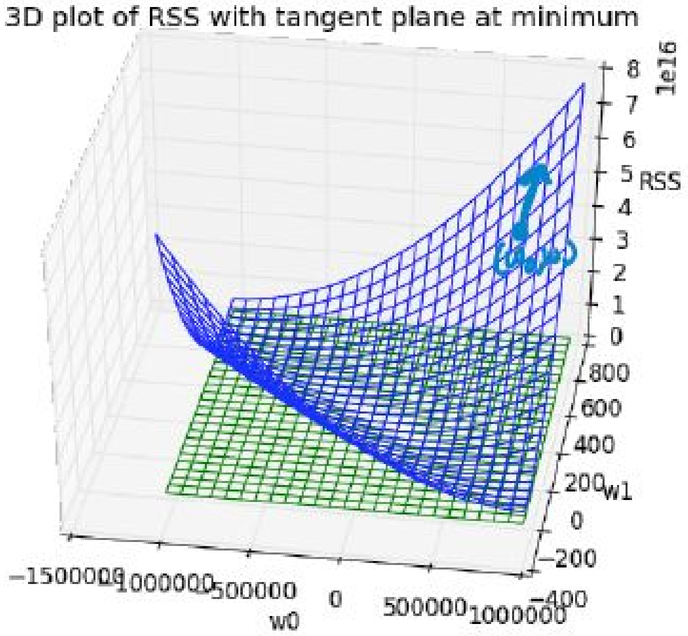
The idea of gradients

Finding the max and min analytically



Work out how to find that minimum point!

Gradient example

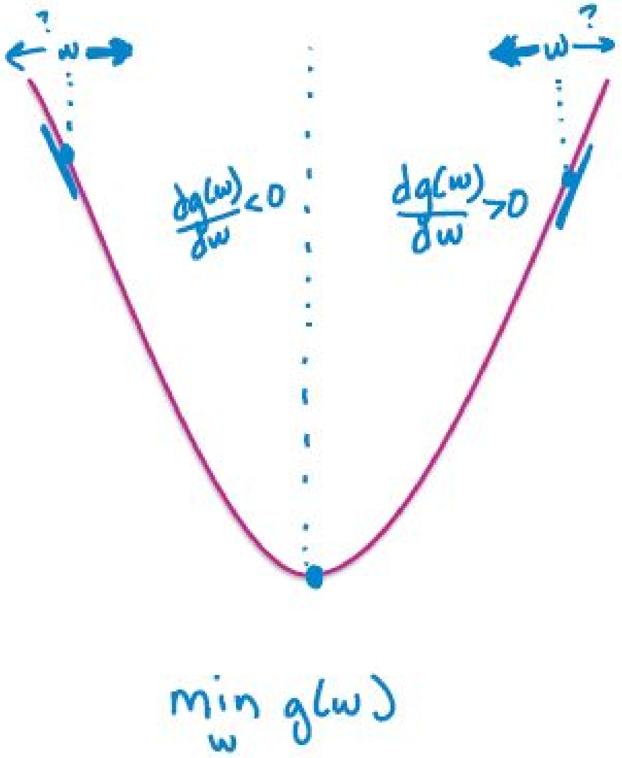


$$g(w) = 5w_0 + 10w_0w_1 + 2w_1^2$$

$$\nabla g(w) =$$

Descending the hill

What we want is to get w which minimizes the cost function g(w)



Step size

While not converged

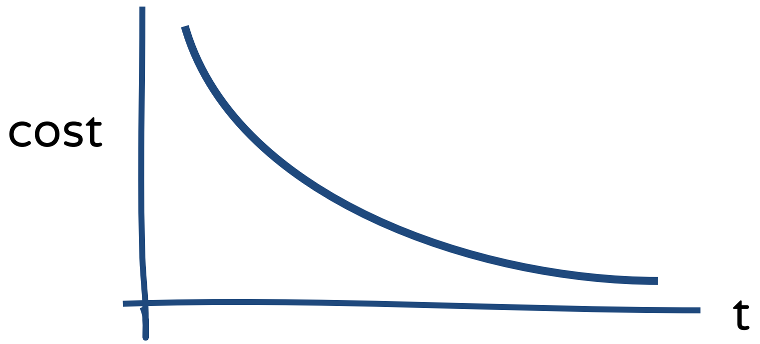
$$w^{(t+1)} \leftarrow w^{(t)} - \eta \left.\frac{dg}{dw}\right|_{w^{(t)}}$$

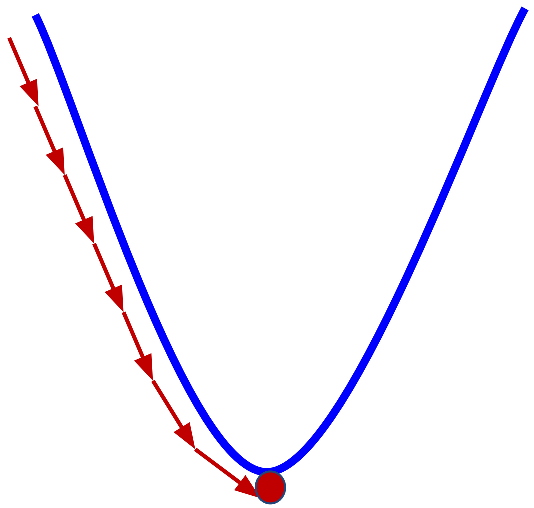
The value of w needs to be iteratively updated by "descending" the gradient ⇒ Gradient Descent

Will it ever converge at 0?

For convex functions, optimum occurs when $\left|\frac{dg}{dw}\right| = 0$

In practice, stop when $\left|\frac{dg}{dw}\right| < \varepsilon$





Is a fixed step size desirable?

Compute the gradient of RSS

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

Taking the derivative w.r.t. w_1 and w_2

Putting it together:

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

Method 1: Set gradient = 0

$$\nabla RSS(w_0,w_1) = \begin{bmatrix} -2\sum_{i=1}^{N}[y_i - (w_0+w_1x_i)] \\ -2\sum_{i=1}^{N}[y_i - (w_0+w_1x_i)]x_i \end{bmatrix} = 0$$

Can you solve these 2 equations to get a closed-form solution?

Method 2: Gradient Descent

Interpreting the gradient:

$$\nabla RSS(w_0,w_1) = \begin{bmatrix} -2\sum_{i=1}^{N}[y_i - (w_0+w_1x_i)] \\ -2\sum_{i=1}^{N}[y_i - (w_0+w_1x_i)]x_i \end{bmatrix} = \begin{bmatrix} -2\sum_{i=1}^{N}[y_i - \hat{y}_i(w_0,w_1)] \\ -2\sum_{i=1}^{N}[y_i - \hat{y}_i(w_0,w_1)]x_i \end{bmatrix}$$

So plug this back to the Gradient Descent formula

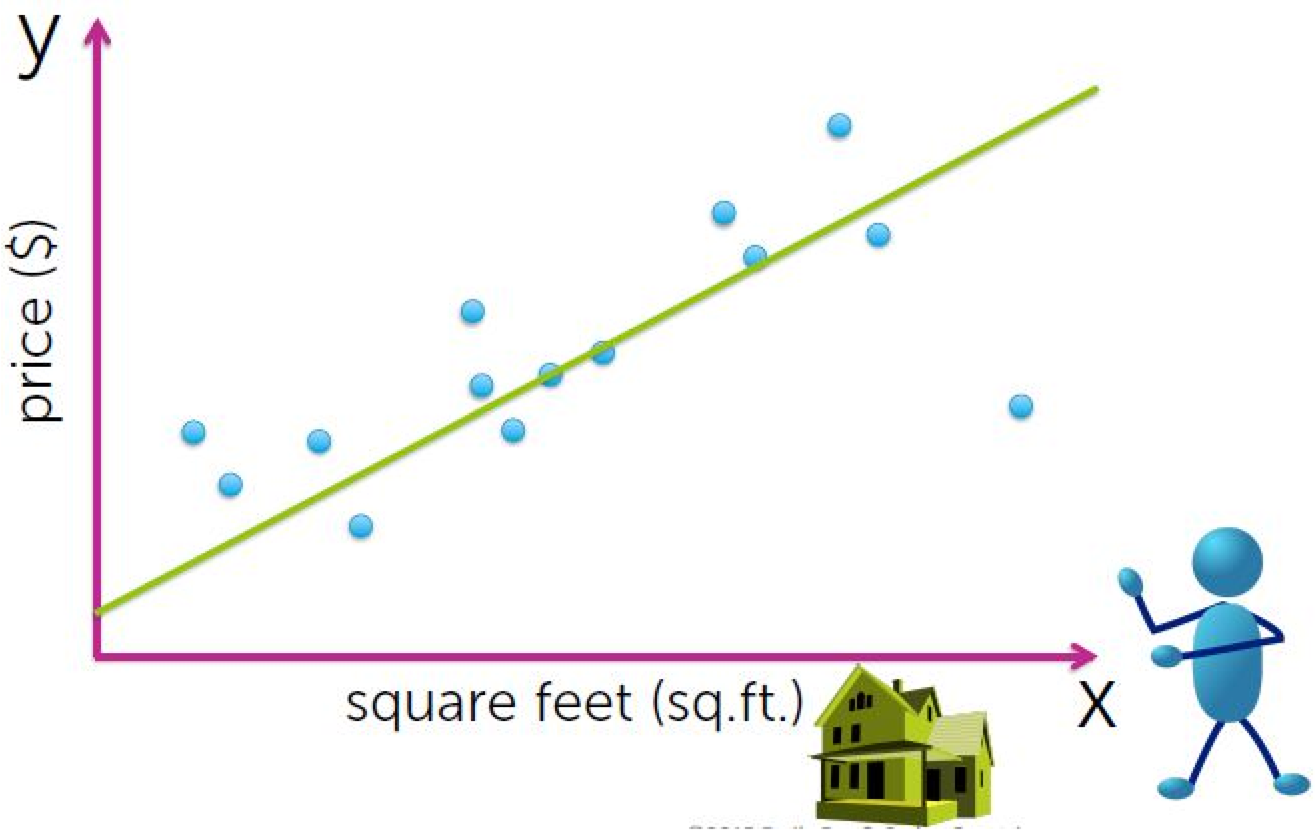
$$w^{(t+1)} \leftarrow w^{(t)} - \eta \left.\frac{dg}{dw}\right|_{w^{(t)}}$$

Gradient descent relies on choosing step size and convergence criteria

Multiple Regression

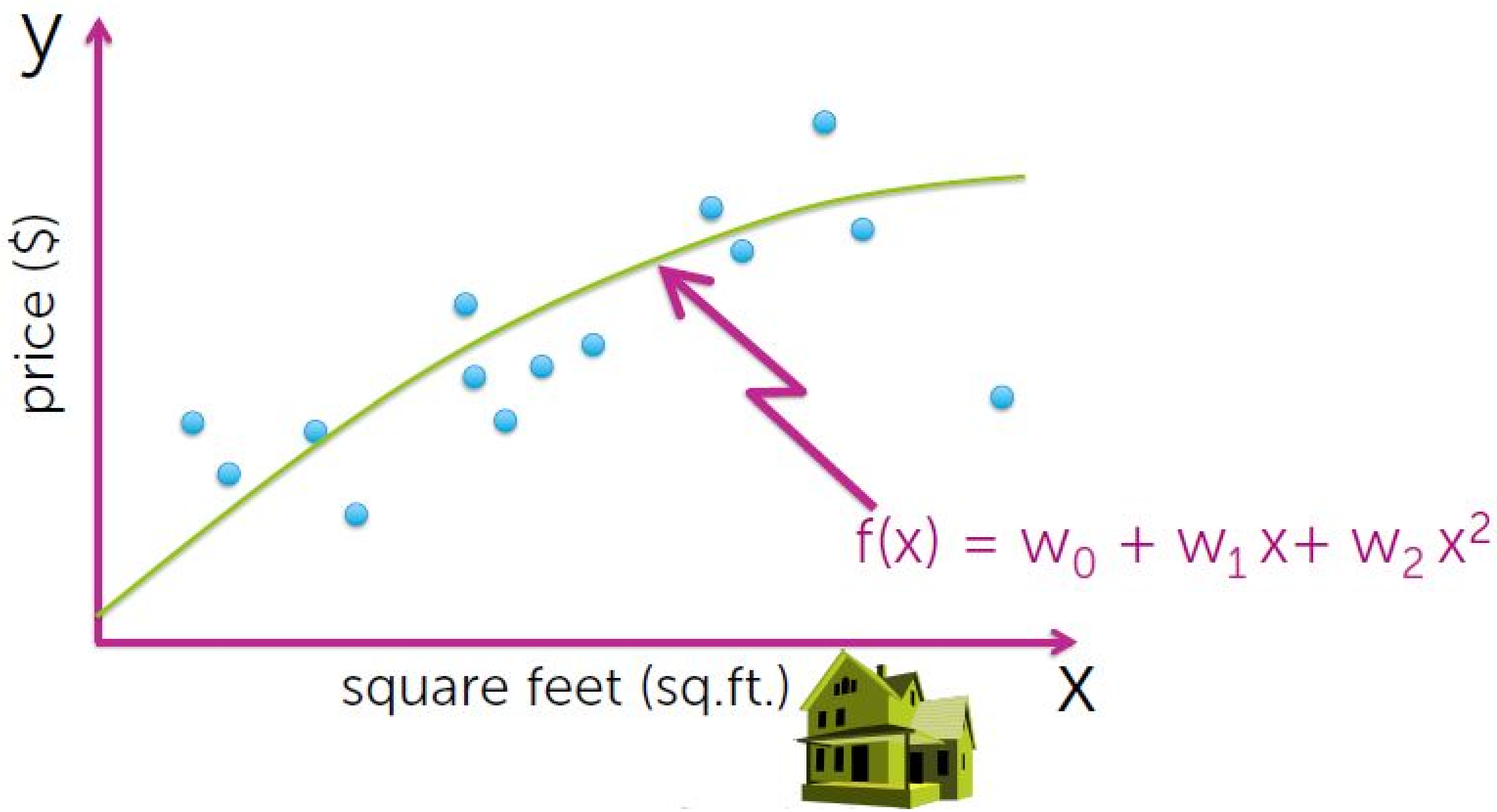
Linear regression with multiple features

Fit with a line or…?

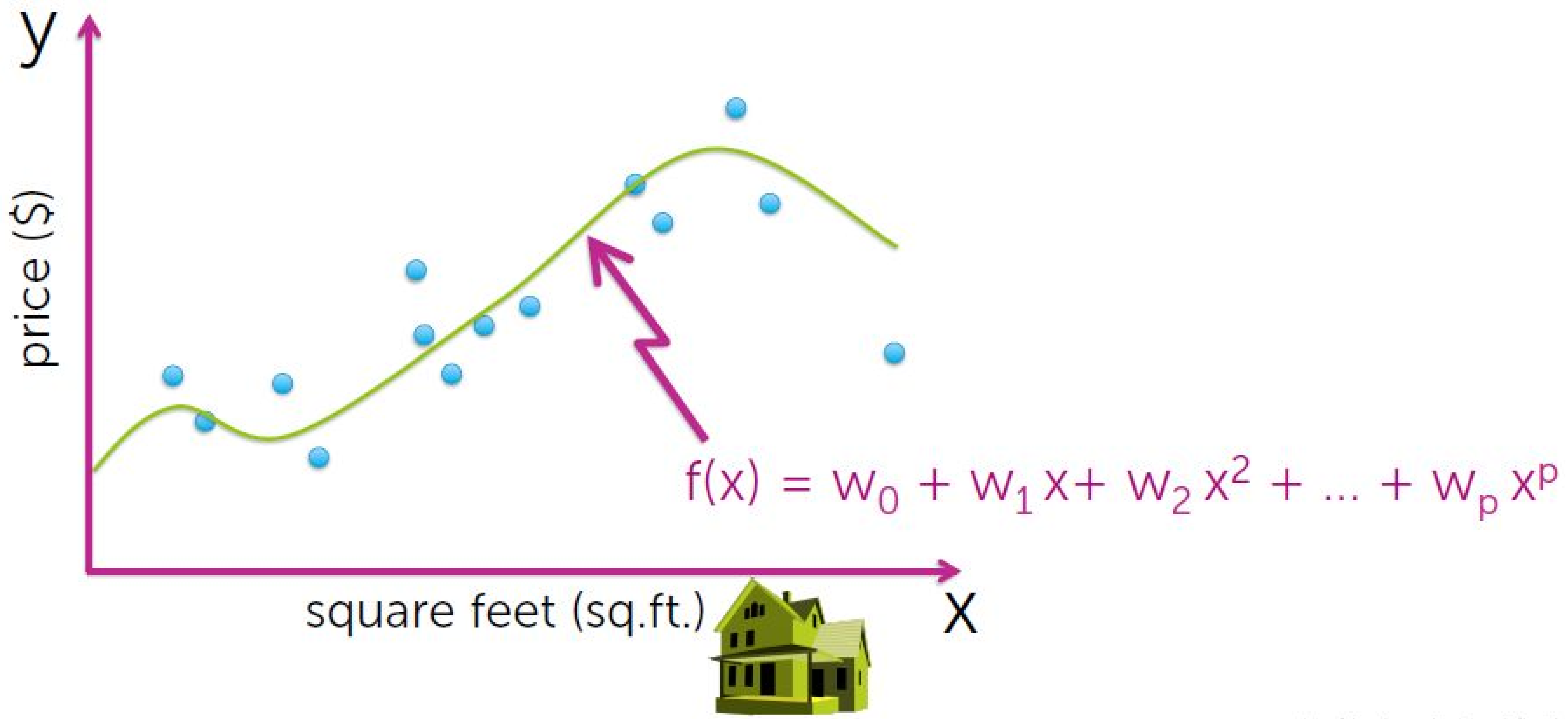


You show your friend your analysis

How about a quadratic function?



Or even higher order polynomials?



Polynomial Regression

Model:

$y_i = w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p + \varepsilon_i$

treat as different **features**

feature 1 = 1 (constant)	*parameter* 1 = w_0
feature 2 = x	*parameter* 2 = w_1
feature 3 = x^2	*parameter* 3 = w_2
…	…
feature p+1 = x^p	*parameter* p+1 = w_p

Generic basis expansion

Model:

$$y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + \ldots + w_D h_D(x_i) + \varepsilon_i$$

$$= \sum_{j=0}^{D} w_j h_j(x_i) + \varepsilon_i$$

j^{th} *feature*

j^{th} *regression coefficient*
or *weight*

feature 1 = $h_0(x)$...often 1 (constant)

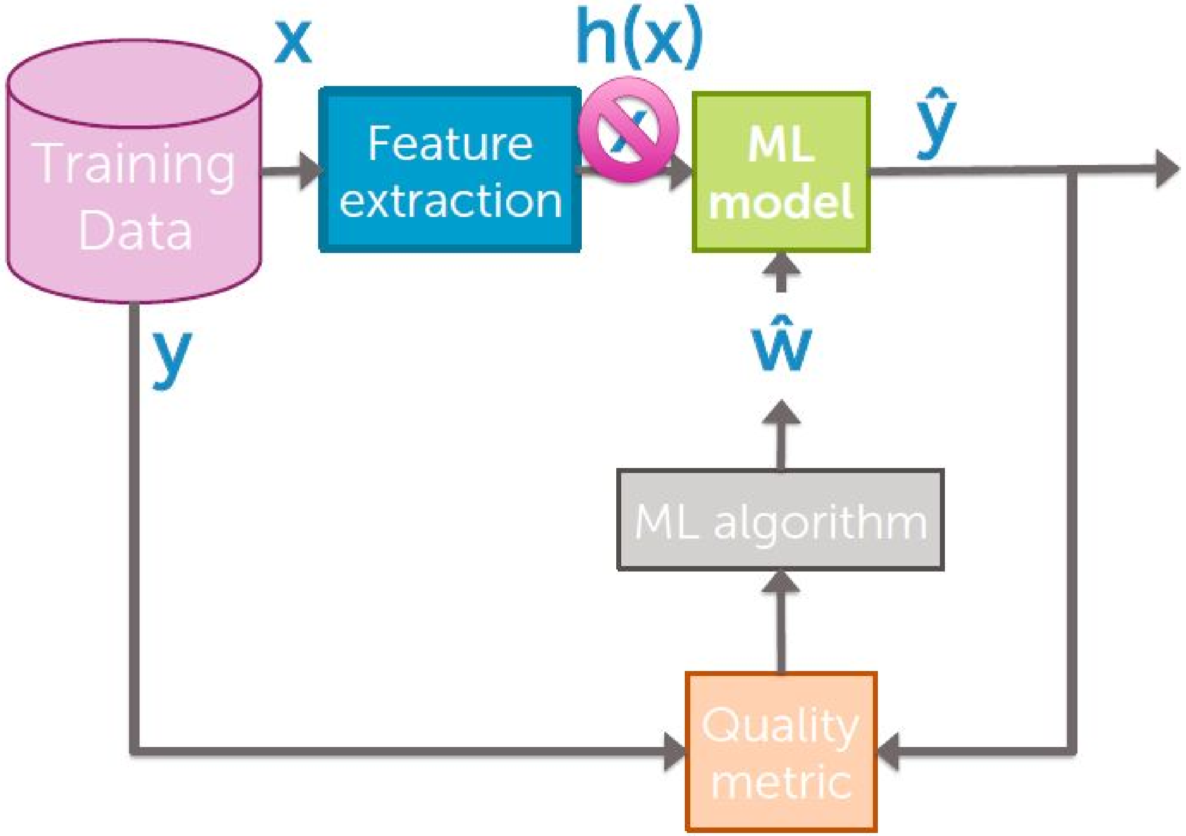
feature 2 = $h_1(x)$... e.g., x

feature 3 = $h_2(x)$... e.g., x^2

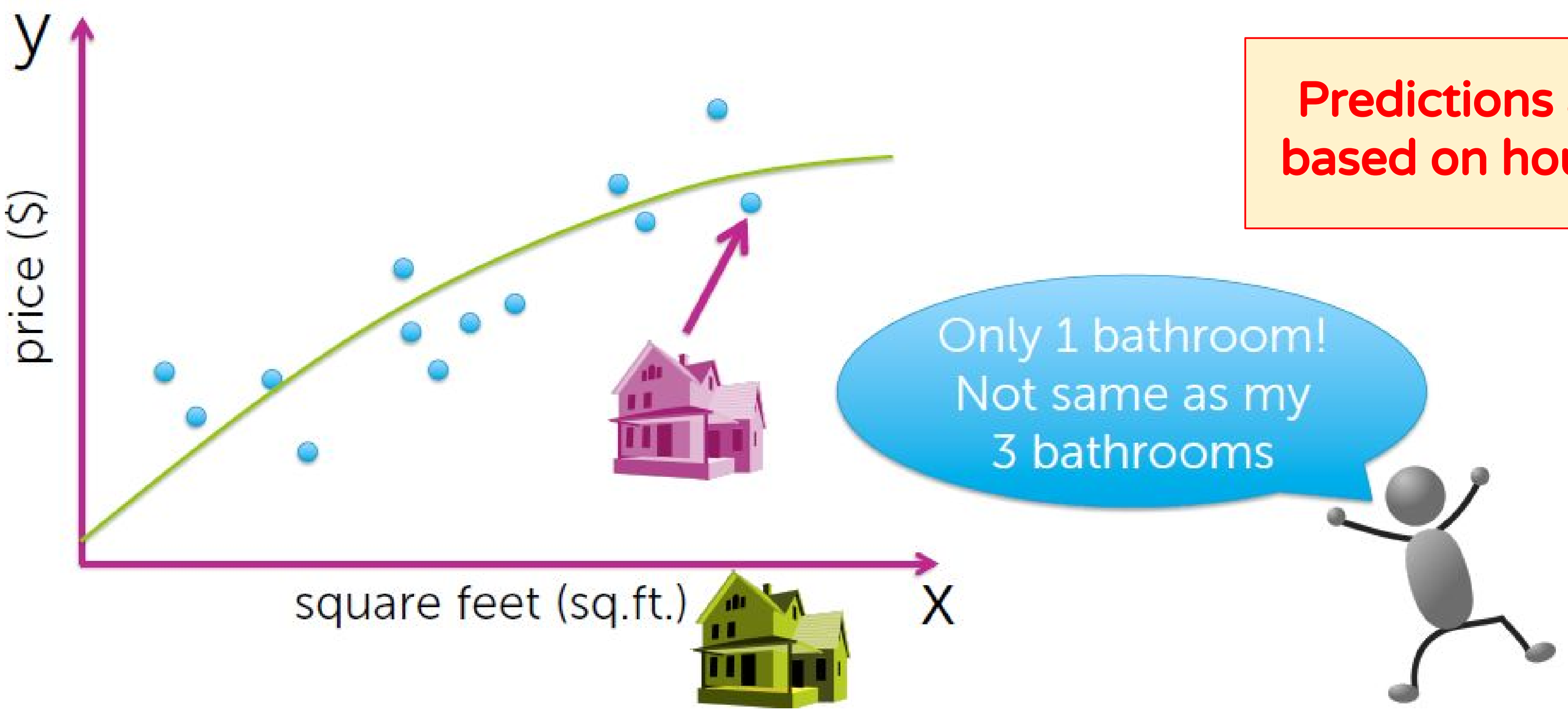
...

feature D+1 = $h_D(x)$... e.g., x^p

No longer a single input

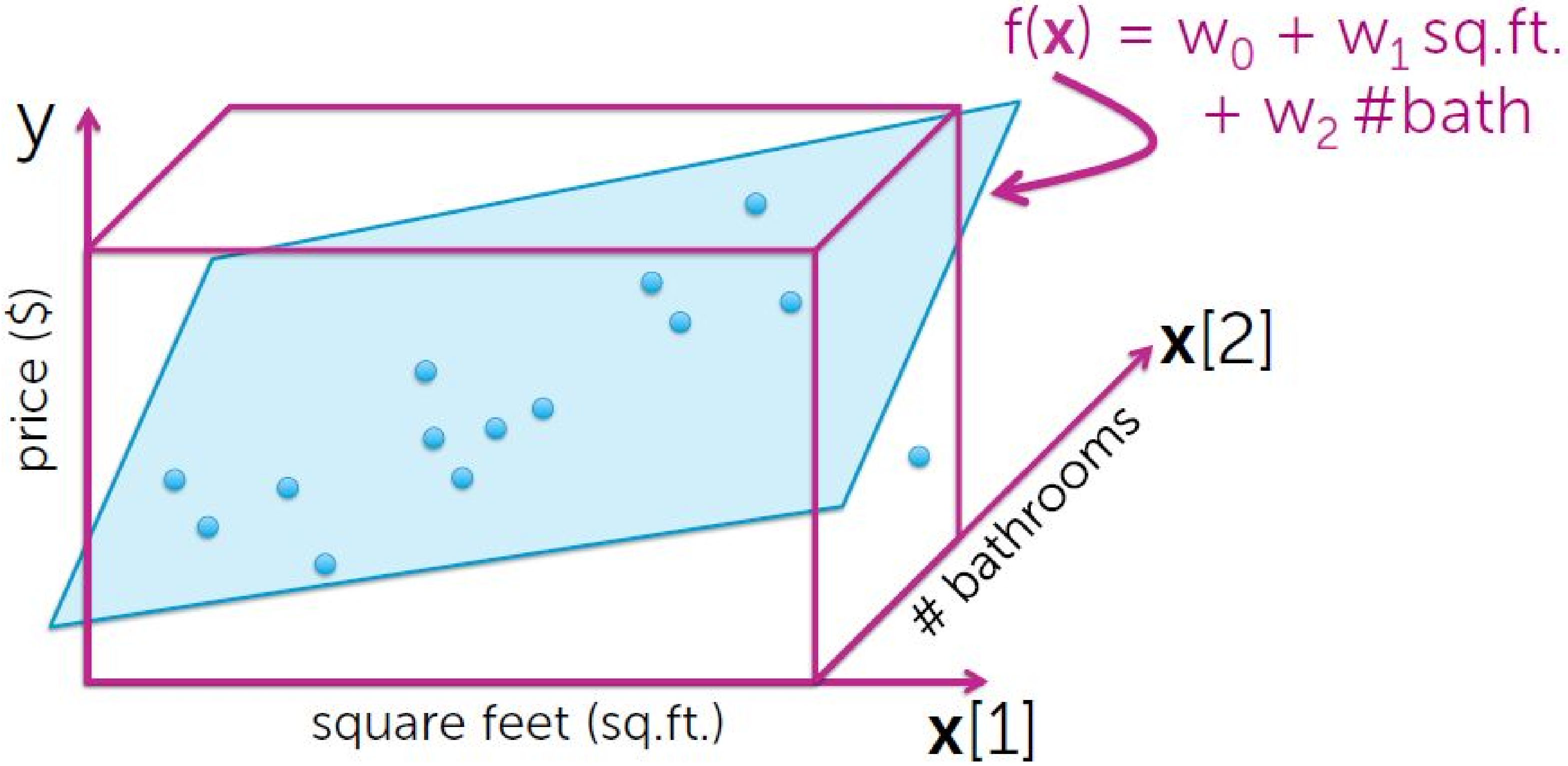


Big house, but…



Predictions are just based on house size!

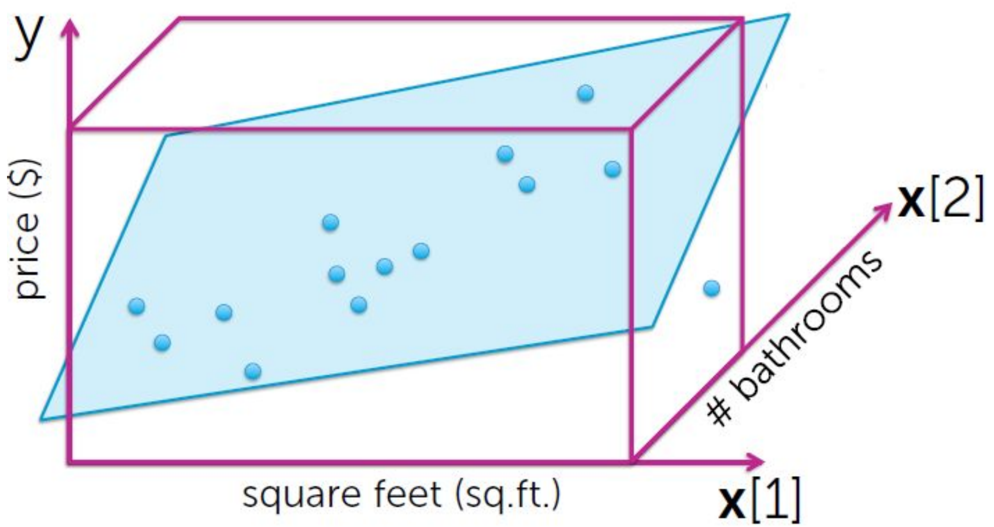
Add more inputs



Any many more

- Square feet
- # bathrooms
- # bedrooms
- Lot size
- Year built
- …

Planes and Hyperplanes



Model:

$y_i = w_0 + w_1\,\mathbf{x}_i[1] + \ldots + w_d\,\mathbf{x}_i[d] + \varepsilon_i$

feature 1 = 1

feature 2 = $\mathbf{x}[1]$ … e.g., sq. ft.

feature 3 = $\mathbf{x}[2]$ … e.g., #bath

…

feature d+1 = $\mathbf{x}[d]$ … e.g., lot size

Generically...a D-dimensional curve

Model:

$$y_i = w_0\, h_0(\mathbf{x}_i) + w_1\, h_1(\mathbf{x}_i) + \ldots + w_D\, h_D(\mathbf{x}_i) + \varepsilon_i$$

$$= \sum_{j=0}^{D} w_j\, h_j(\mathbf{x}_i) + \varepsilon_i$$

feature 1 = $h_0(\mathbf{x})$ ... e.g., 1

feature 2 = $h_1(\mathbf{x})$ ... e.g., $\mathbf{x}[1]$ = sq. ft.

feature 3 = $h_2(\mathbf{x})$ ... e.g., $\mathbf{x}[2]$ = #bath
or, $\log(\mathbf{x}[7])\ \mathbf{x}[2]$ = log(#bed) x #bath

...

feature D+1 = $h_D(\mathbf{x})$ ... some other function of $\mathbf{x}[1],\ldots,\mathbf{x}[d]$

Some common notations

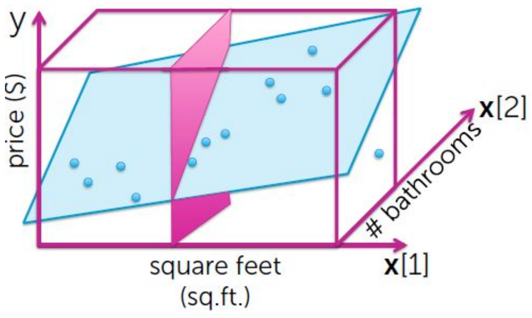
observations $(\mathbf{x}_i, y_i)$: N

inputs $\mathbf{x}[j]$: d

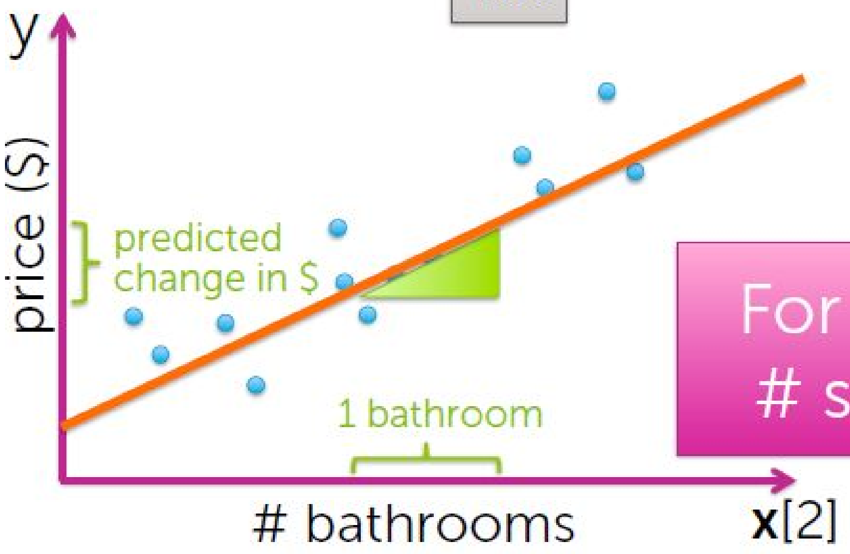
features $h_j(\mathbf{x})$: D

Interpreting the coefficients

For 2 linear features



$$\hat{y} = \hat{w}_0 + \hat{w}_1 \underbrace{\boxed{\mathbf{x}[1]}}_{\text{fix}} + \hat{w}_2 \mathbf{x}[2]$$



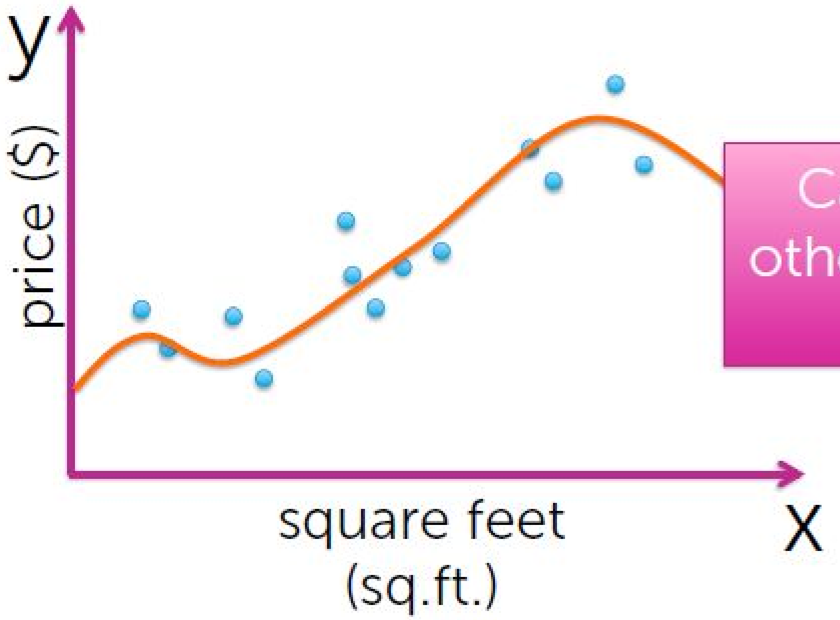
For fixed
sq.ft.!

We can fix one of the two features to see the other

Interpreting the coefficients

For multiple linear features

$$\hat{y} = \hat{w}_0 + \hat{w}_1 \underset{\text{fix}}{\boxed{\mathbf{x}[1]}} \underset{\text{fix}}{\boxed{+ \ldots}} + \hat{w}_j \mathbf{x}[j] \underset{\text{fix}}{\boxed{+ \ldots}} + \hat{w}_d \underset{\text{fix}}{\boxed{\mathbf{x}[d]}}$$



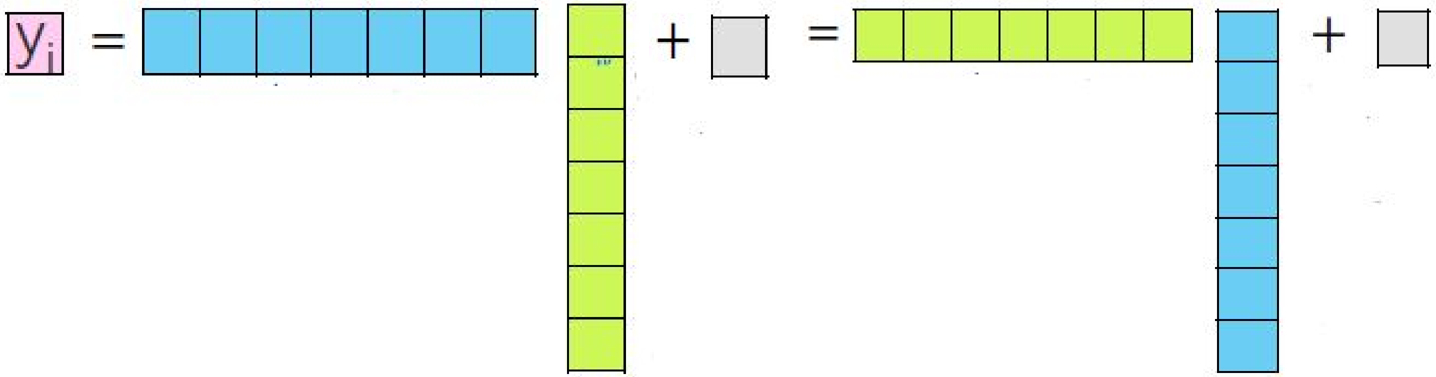
Can't hold other features fixed!

Fitting D-dimensional curves

Rewrite in matrix notation

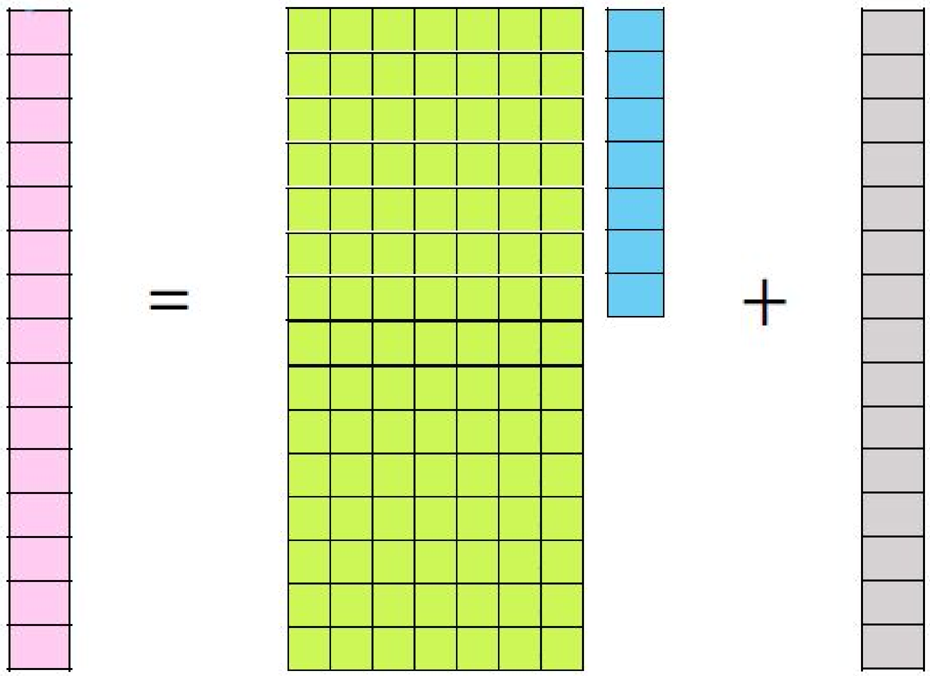
For observation i

$$y_i = \sum_{j=0}^{D} w_j h_j(\mathbf{x}_i) + \varepsilon_i$$



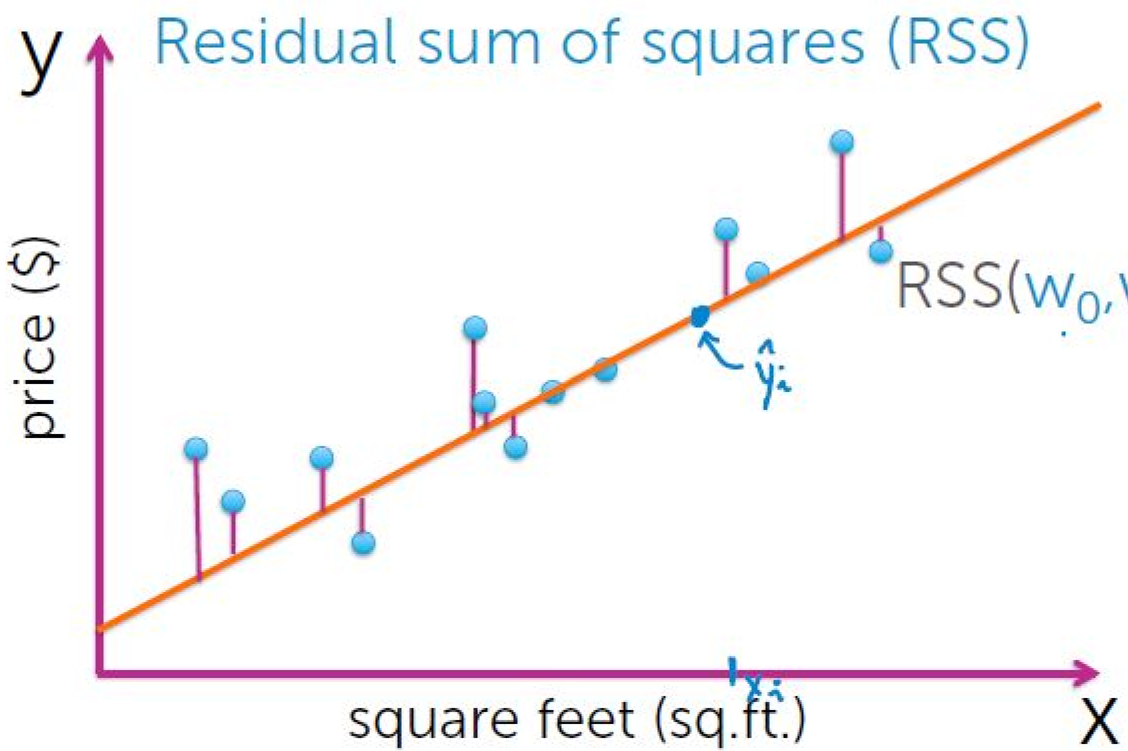
Fitting D-dimensional curves

For all observations together



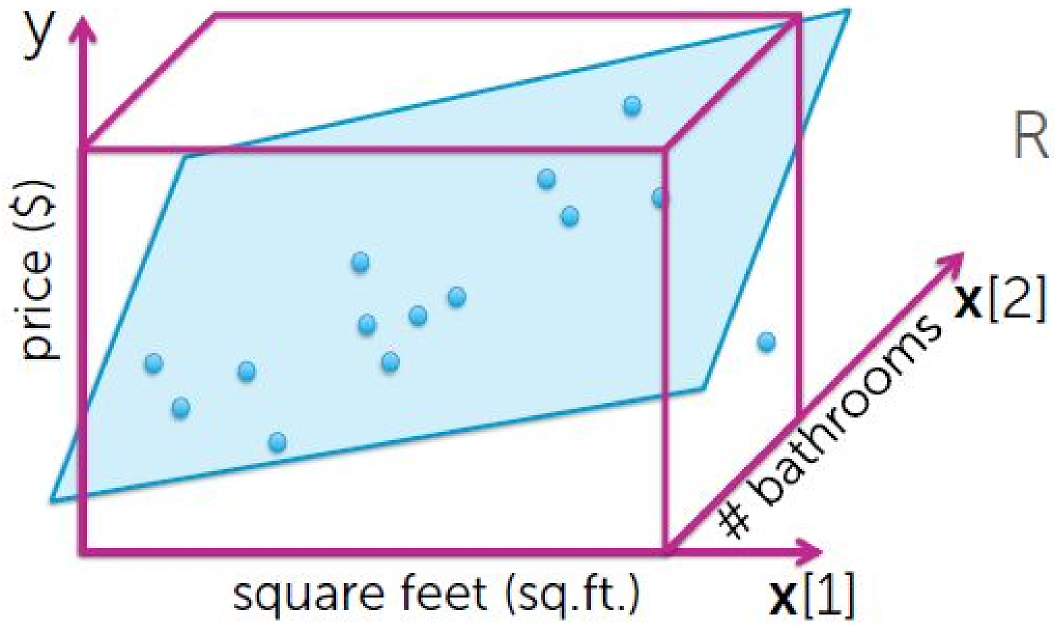
$y = Hw + \varepsilon$

Recap: Cost of using a line

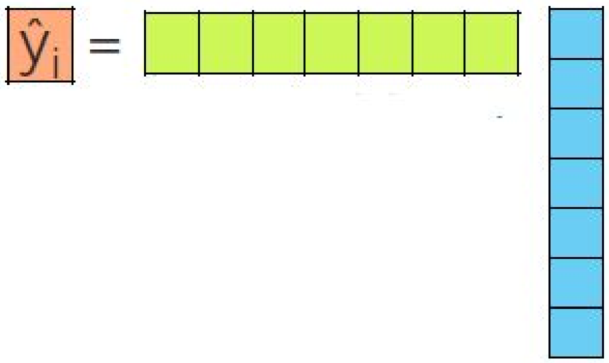


$$RSS(w_0,w_1) = \sum_{i=1}^{N} (y_i - [w_0 + w_1 x_i])^2$$

RSS for multiple regression



$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \qquad)^2$$



RSS in matrix notation

$$\mathrm{RSS}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - h(\mathbf{x}_i)^T \mathbf{w})^2$$

$$= (\mathbf{y} - \mathbf{H}\mathbf{w})^T (\mathbf{y} - \mathbf{H}\mathbf{w})$$

Why?

$$\begin{bmatrix} \text{residual}_1 & \text{residual}_2 & \text{residual}_3 & \cdots & \text{residual}_N \end{bmatrix} \begin{bmatrix} \text{residual}_1 \\ \text{residual}_2 \\ \text{residual}_3 \\ \cdots \\ \text{residual}_N \end{bmatrix}$$

Gradient of RSS

$$\nabla \text{RSS}(\mathbf{w}) = \nabla[(\mathbf{y}-\mathbf{H}\mathbf{w})^T(\mathbf{y}-\mathbf{H}\mathbf{w})]$$

$$= -2\mathbf{H}^T(\mathbf{y}-\mathbf{H}\mathbf{w})$$

Why? By analogy to 1D case:

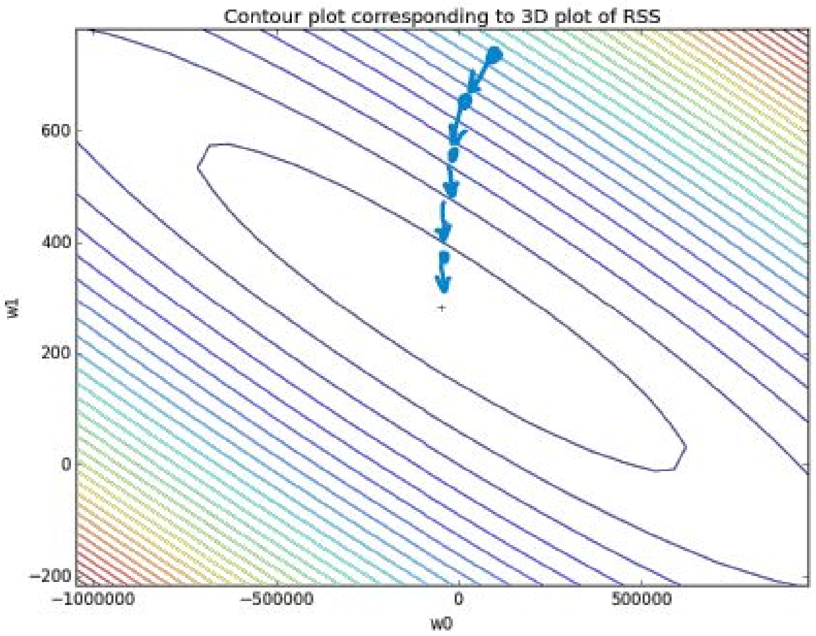
Approaches to get w

Approach 1: Set gradient = 0 and solve it to get closed-form solution

Answer: $\hat{\mathbf{w}} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}$

Gradient Descent

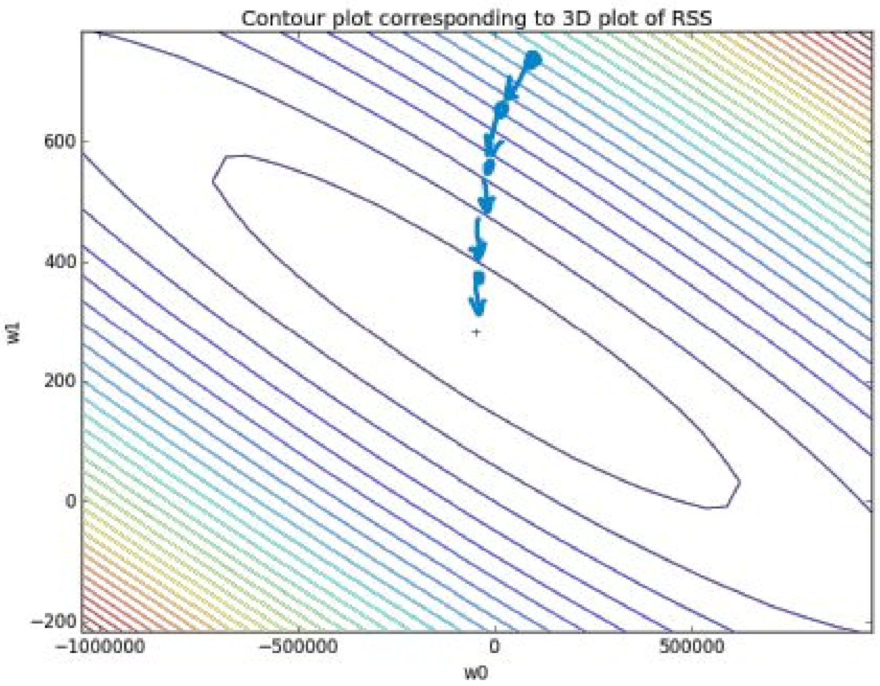
Approach 2: Use Gradient Descent to optimize value of w



while not converged

$$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \underbrace{\nabla \mathrm{RSS}(\mathbf{w}^{(t)})}_{-2\mathbf{H}^T(\mathbf{y}-\mathbf{H}\mathbf{w})}$$

Summary of GD for multiple regression



init $\mathbf{w}^{(1)}=0$ (or randomly, or smartly), $t=1$

while $||\nabla \mathrm{RSS}(\mathbf{w}^{(t)})|| > \varepsilon$

for $j=0,\ldots,D$

$\mathrm{partial}[j] = -2\sum_{i=1}^{N} h_j(\mathbf{x}_i)(y_i-\hat{y}_i(\mathbf{w}^{(t)}))$

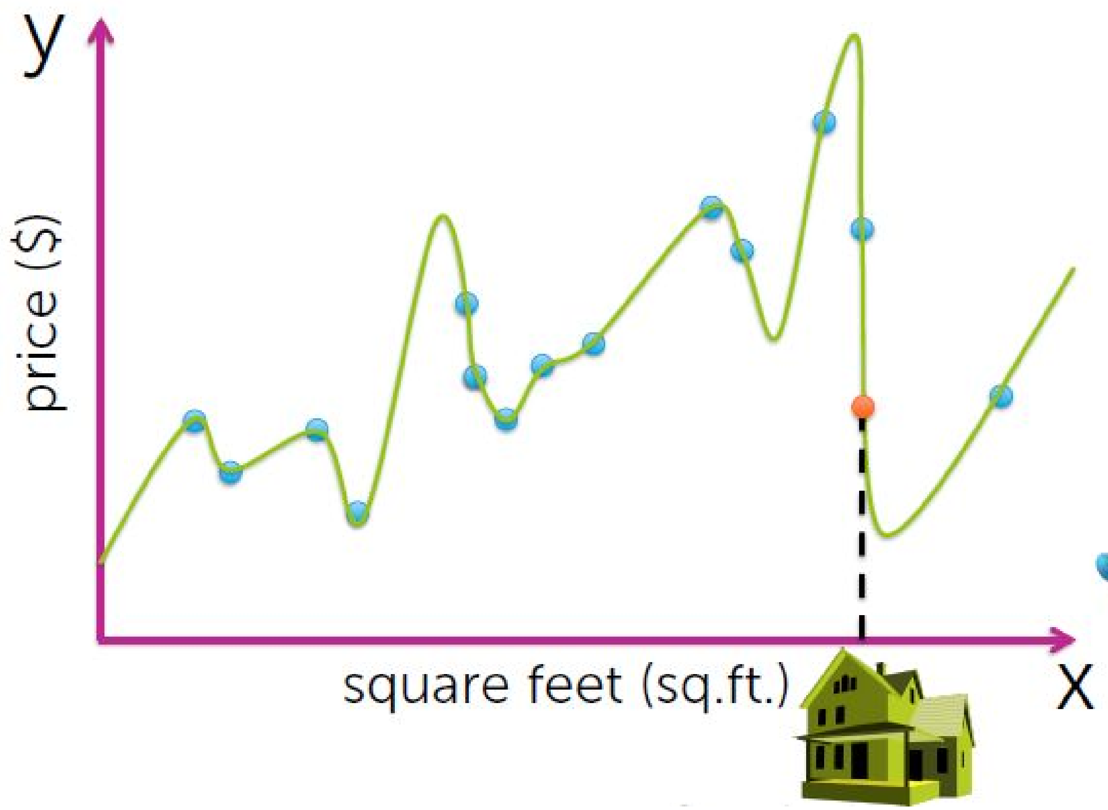
$\mathbf{w}_j^{(t+1)} \leftarrow \mathbf{w}_j^{(t)} - \eta\ \mathrm{partial}[j]$

$t \leftarrow t + 1$

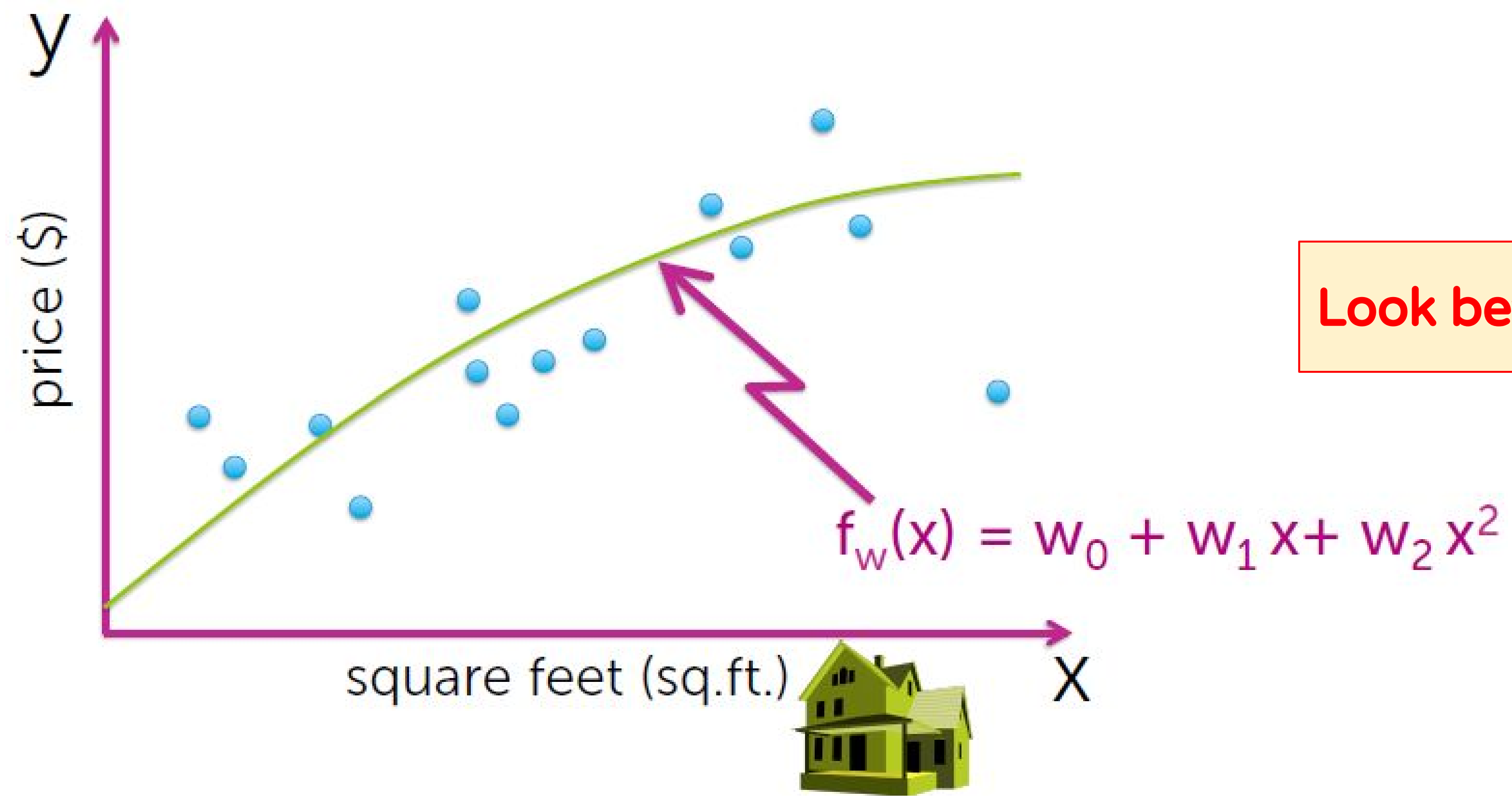
Evaluating overfitting via training/test split

Do you believe this fit?

Well, model minimizes RSS well, but it may make bad predictions…



What about a quadratic function?



Look better now?

How to choose model order/complexity

- Want good predictions, but can't observe future
- **Simulate predictions**

1. Remove some houses
2. Fit model on remaining
3. Predict heldout houses

Training/test split

Terminology:
- training set
- test set

Training error

Terminology: – training set

– test set

Go To Exercises

Anyone tried Kaggle?