Regression: 

Predicting House Prices

Emily Fox
University of Washington
March 27, 2018

Predicting house prices

How much is my house worth?

How much is my house worth?

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 = \$)

⋮

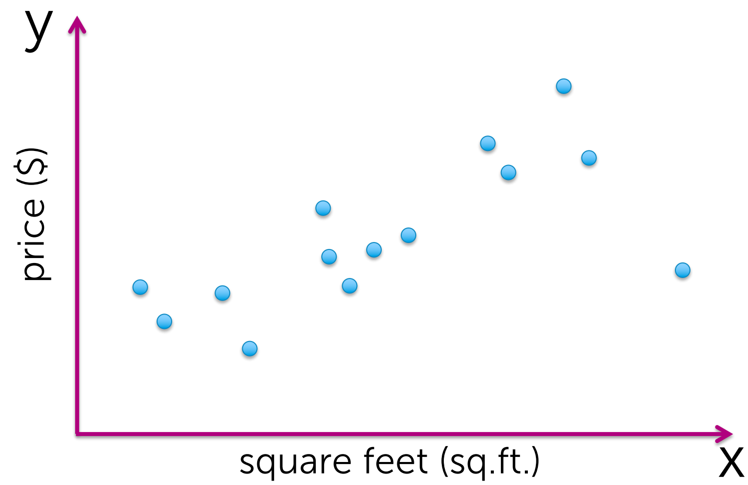
Input vs. Output:

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

Look at recent sales in my neighborhood

How much did they sell for?

Plot recent house sales
(Past 2 years)

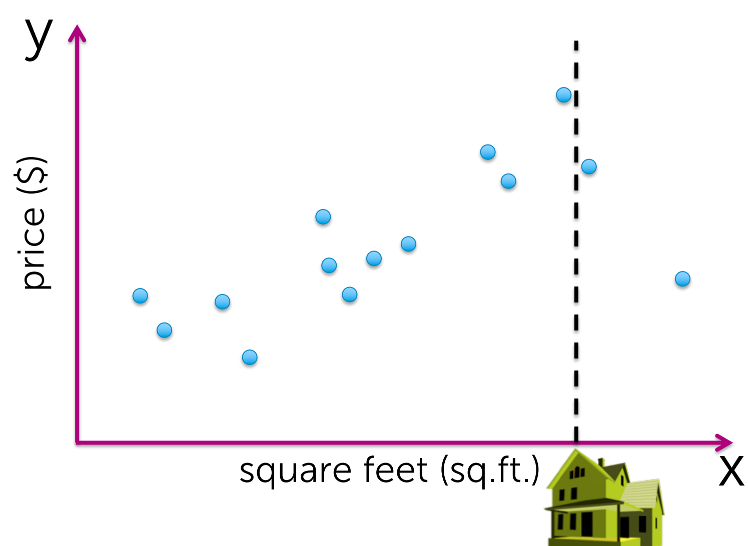


Terminology:

x – feature, covariate, or predictor

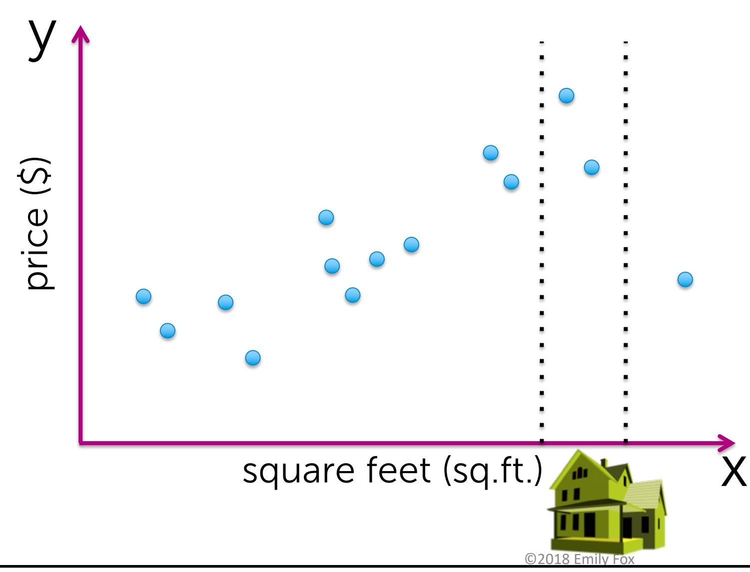
y – observation or response

Predict your house by similar houses



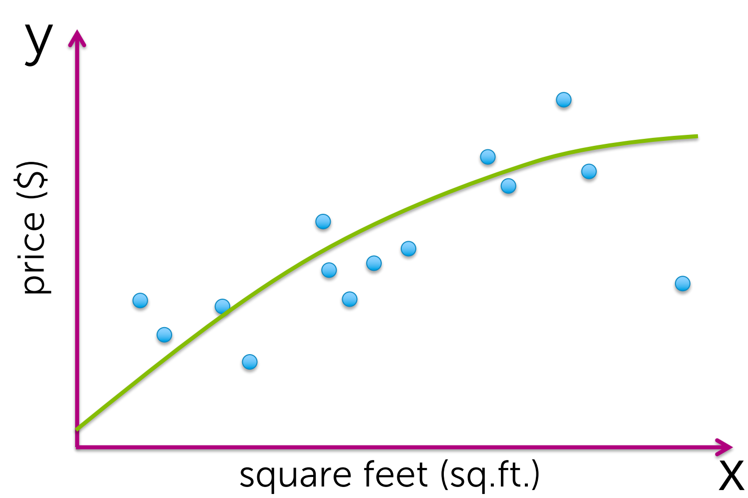
No house sold recently had *exactly* the same sq.ft.

Predict your house by similar houses



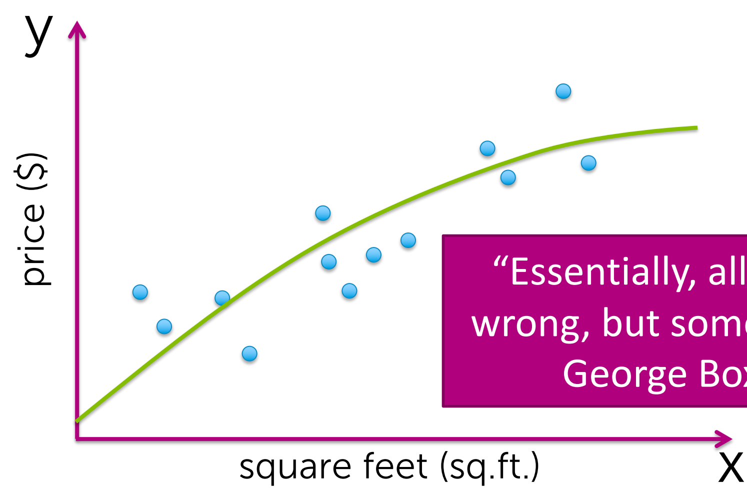
- Look at average price in range
- **Still only 2 houses!**
- Throwing out info from all other sales

Model – How we *assume* the world works

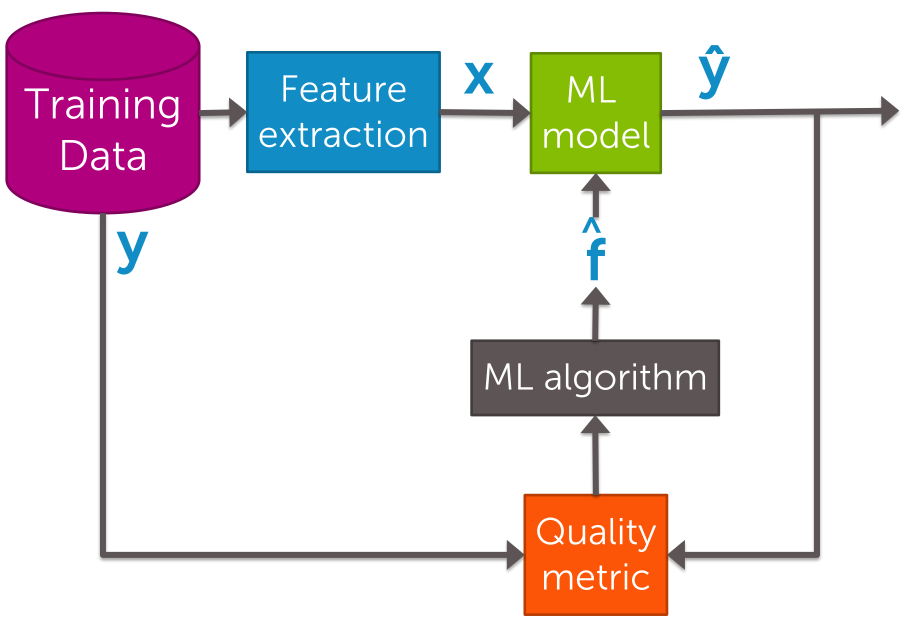


Regression model:

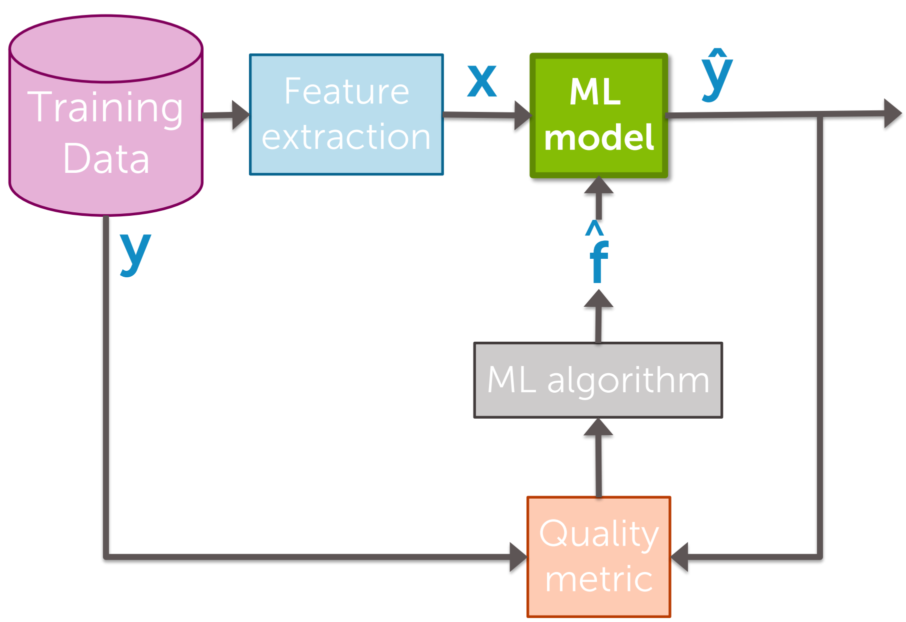
Model – How we *assume* the world works



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

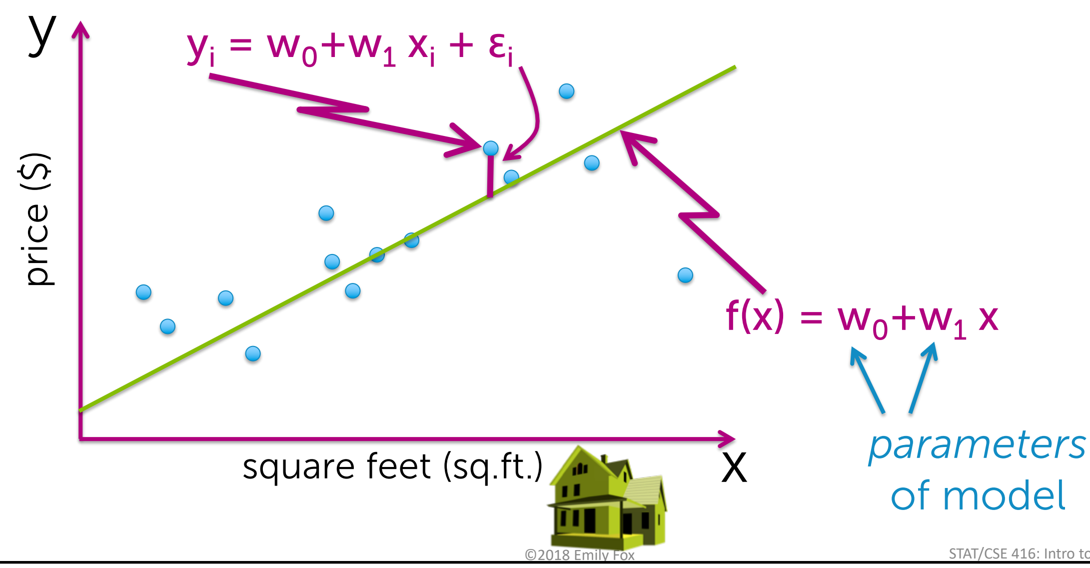


Linear regression

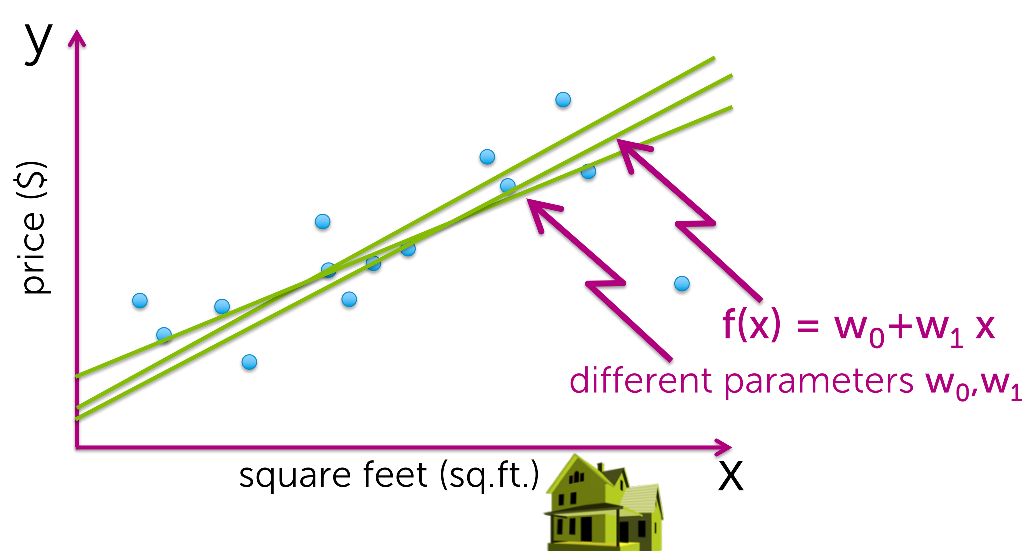


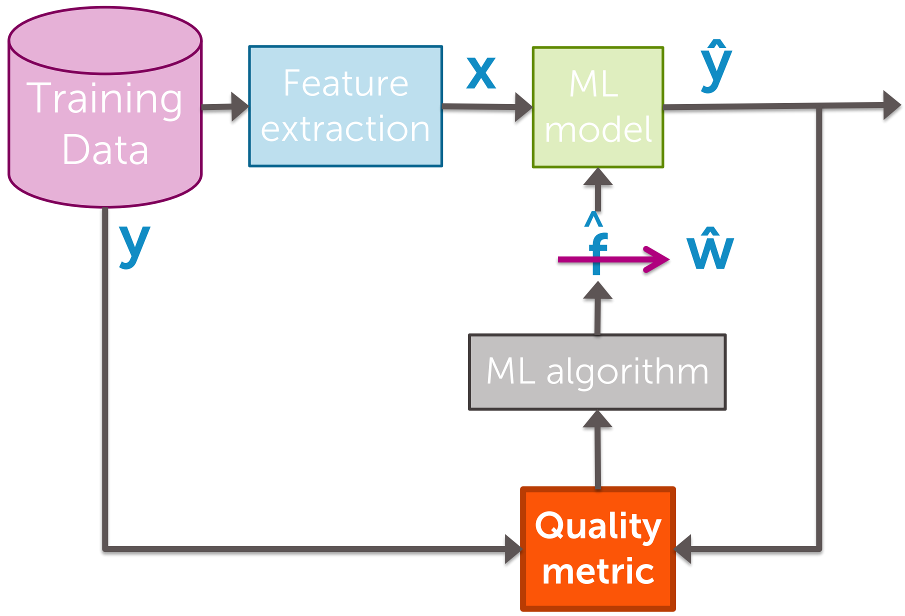
Use a simple **linear** regression model

Fit a line through the data



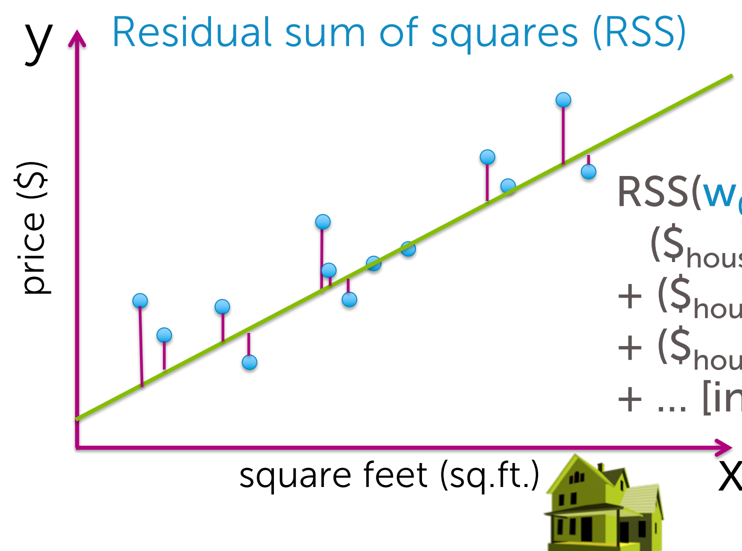
Which line?





"Cost" of using a given line

Residual sum of squares (RSS)



$RSS(w_0,w_1) =$

$(\$_{house\ 1}-[w_0+w_1 sq.ft._{house\ 1}])^2$

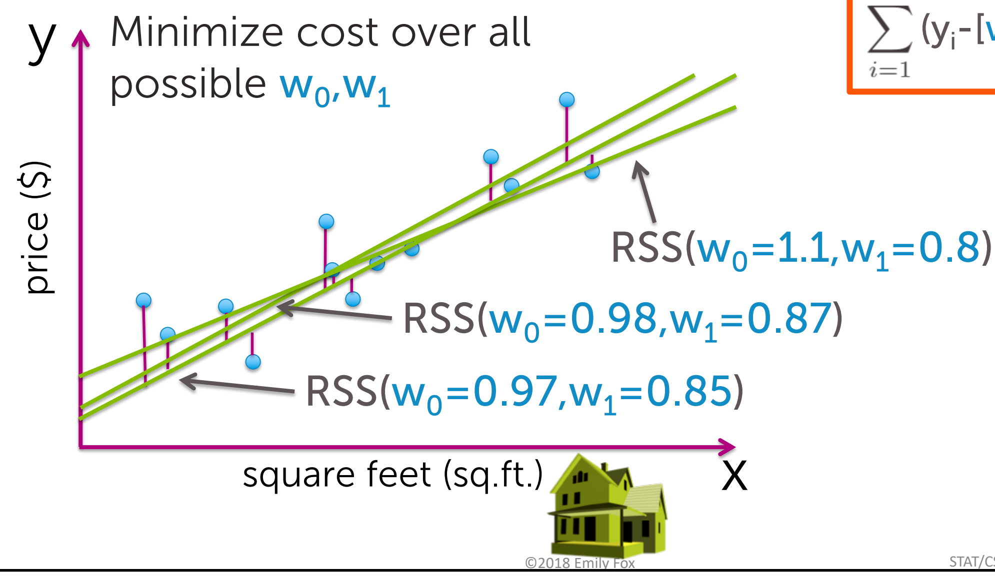
$+ (\$_{house\ 2}-[w_0+w_1 sq.ft._{house\ 2}])^2$

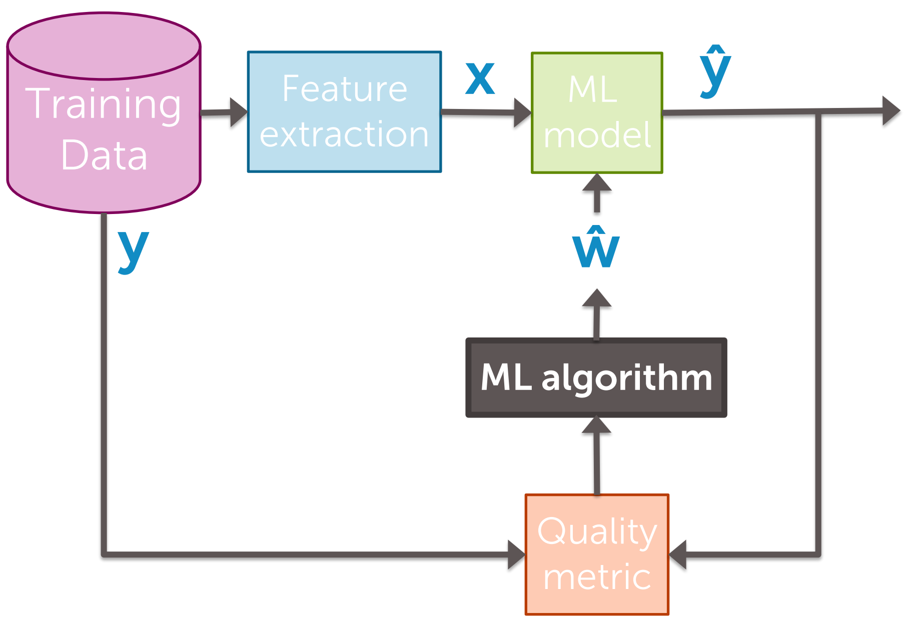
$+ (\$_{house\ 3}-[w_0+w_1 sq.ft._{house\ 3}])^2$

$+ \ldots$ [include all houses]

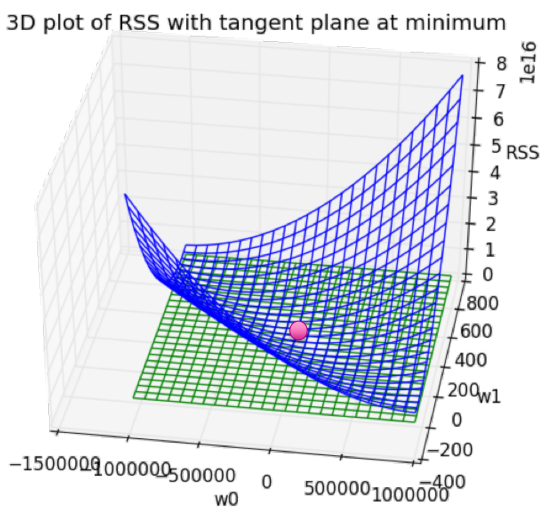
Find “best” line

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





Minimizing the cost

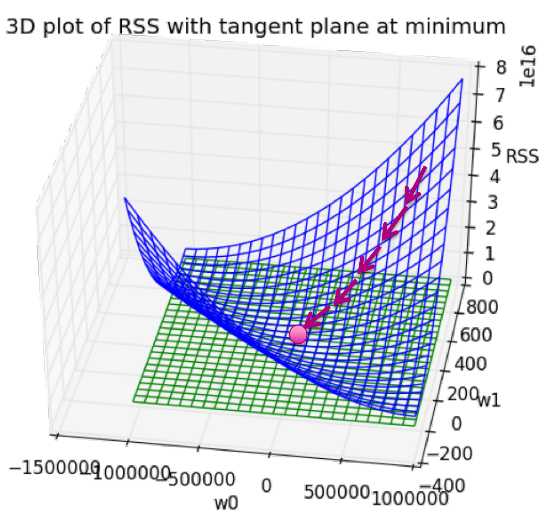


Minimize function over all possible w_0, w_1

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

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

Gradient descent



Algorithm:

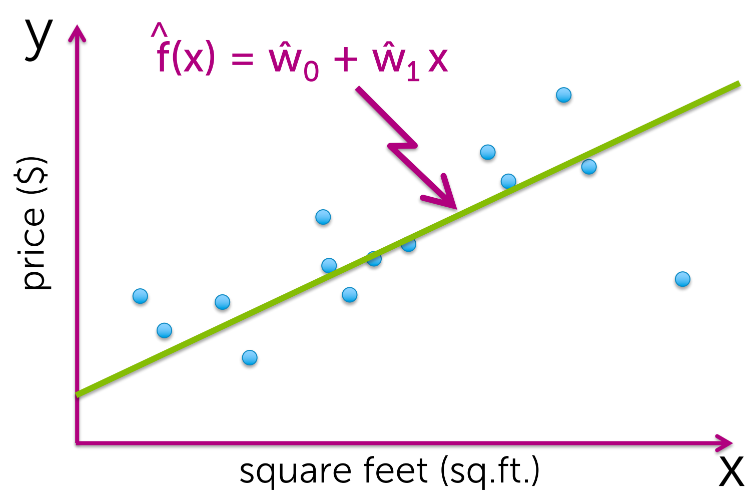
while not converged

$w^{(t+1)} \leftarrow w^{(t)} - \eta \nabla RSS(w^{(t)})$

$\hat{w}$

The fitted line: use + interpretation

Model vs. fitted line



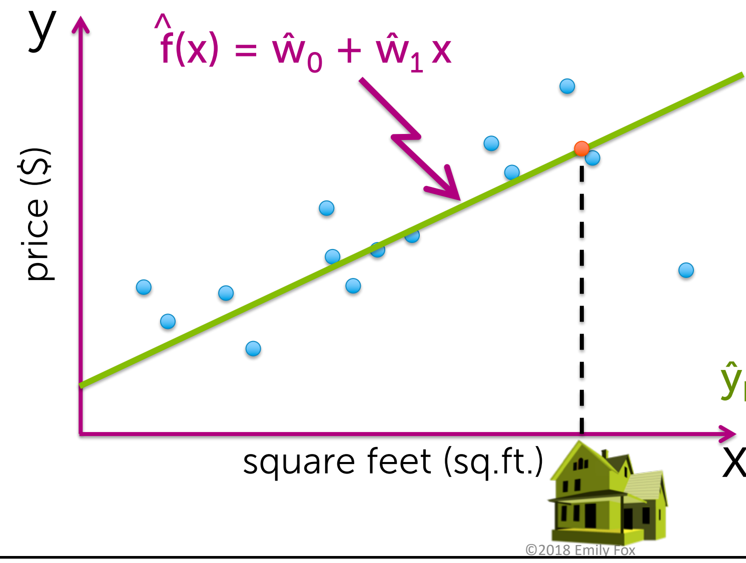
Regression model:

$y_i = w_0 + w_1 x_i + \varepsilon_i$

Estimated parameters:

$\hat{w}_0$, $\hat{w}_1$

Seller:
Predicting your house price



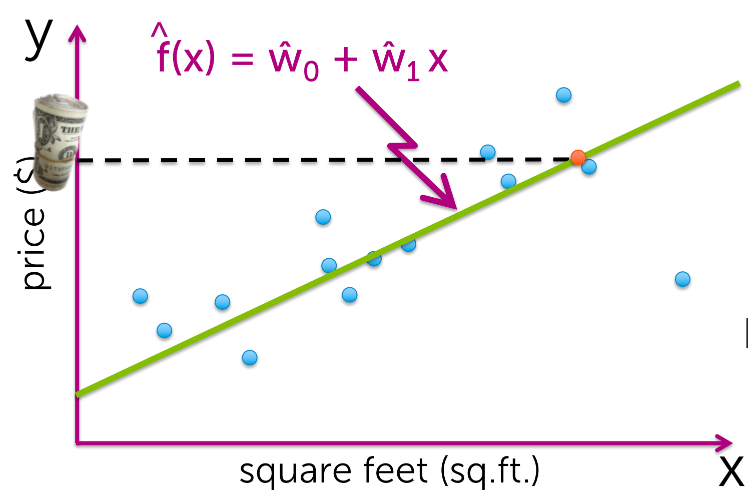
Regression model:

$y_i = w_0 + w_1 x_i + \varepsilon_i$

Best guess of your house price:

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

Buyer:
Predicting size of house



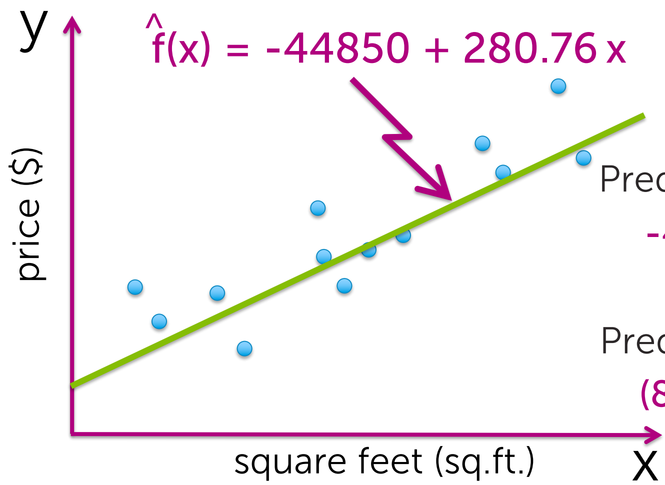
Regression model:

$y_i = w_0 + w_1 x_i + \varepsilon_i$

Best guess of size of house you can afford:

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

A concrete example



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

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

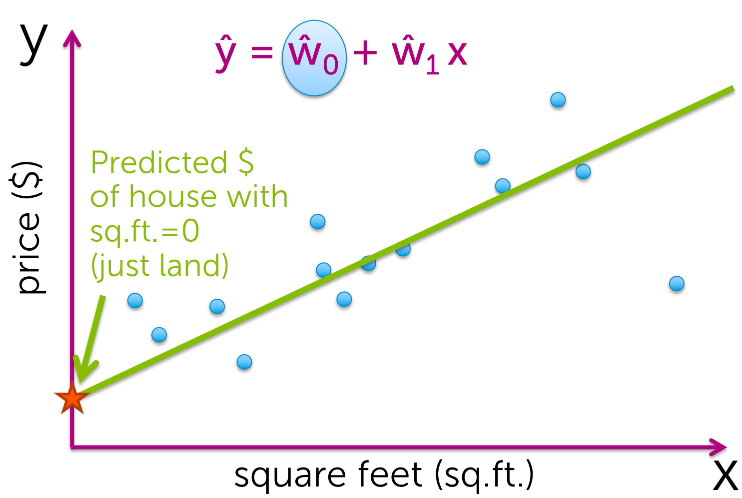
$= \$696{,}356$

Predict sq.ft. of $859,000 sale:

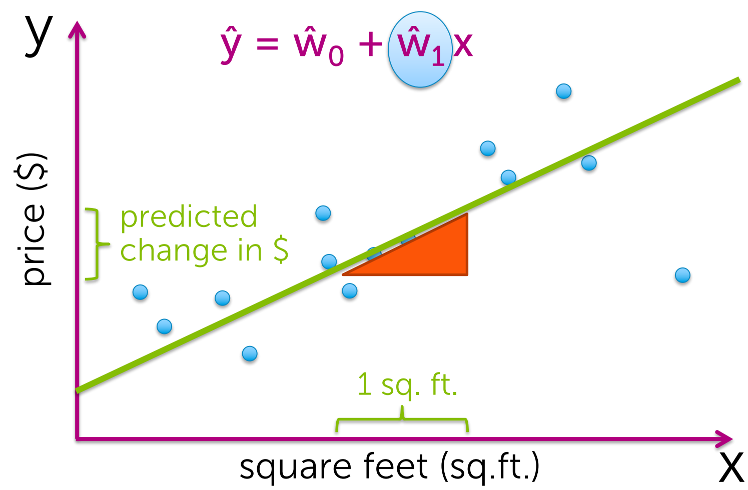
$(859000+44850)/\ 280.76$

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

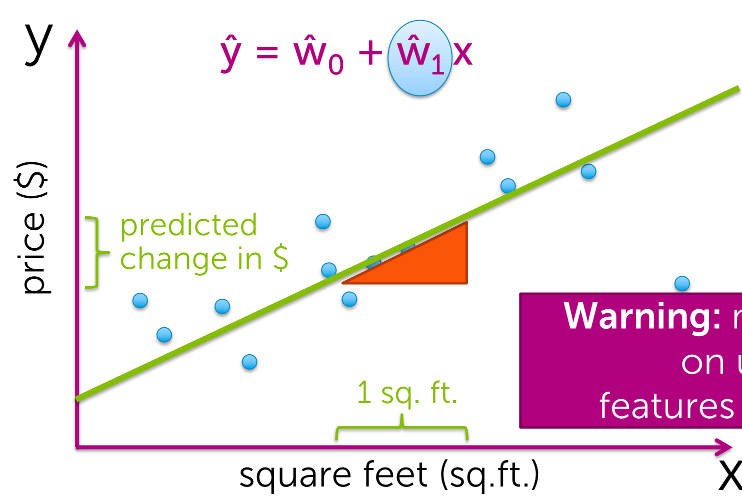
Interpreting the coefficients



Interpreting the coefficients



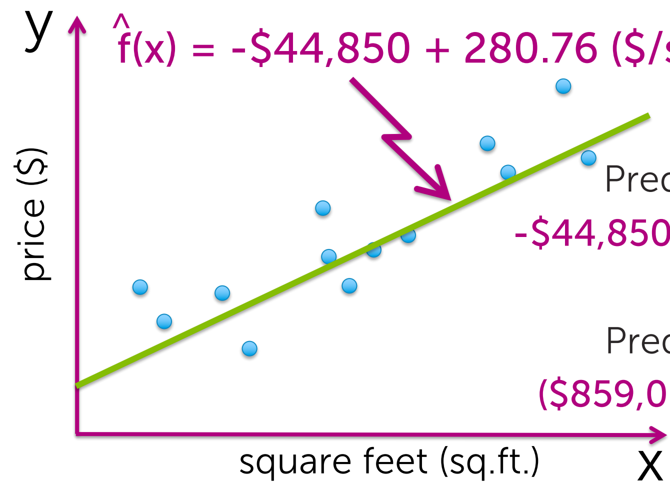
Interpreting the coefficients



Warning: magnitude depends on units of both features and observations

A concrete example

$\hat{f}(x)$ = -$44,850 + 280.76 ($/sq.ft.) x



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

-$44,850 + 280.76 ($/sq.ft.) * 2,640 sq.ft.

= $696,356

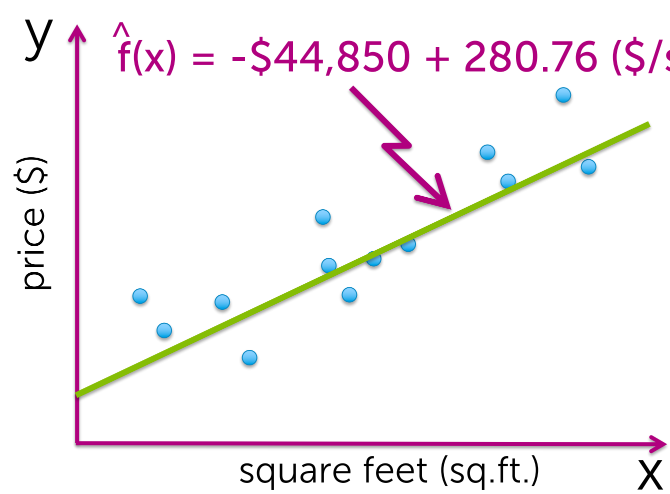
Predict sq.ft. of $859,000 sale:

($859,000+$44,850)/ 280.76 ($/sq.ft.)

= 3,219 sq.ft.

A concrete example

$\hat{f}(x)$ = -$44,850 + 280.76 ($/sq.ft.) x

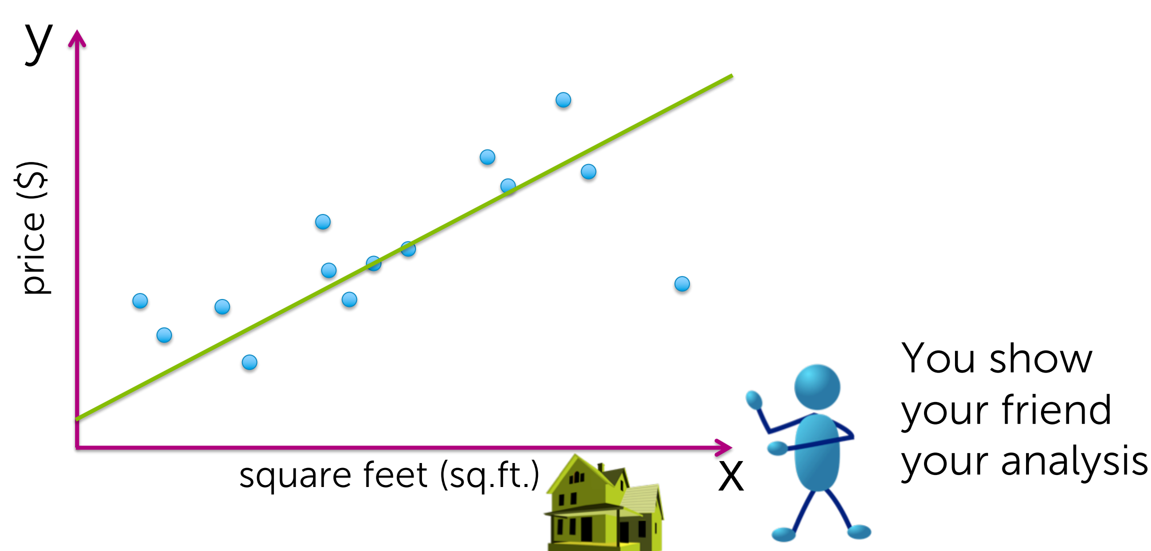


But what if:

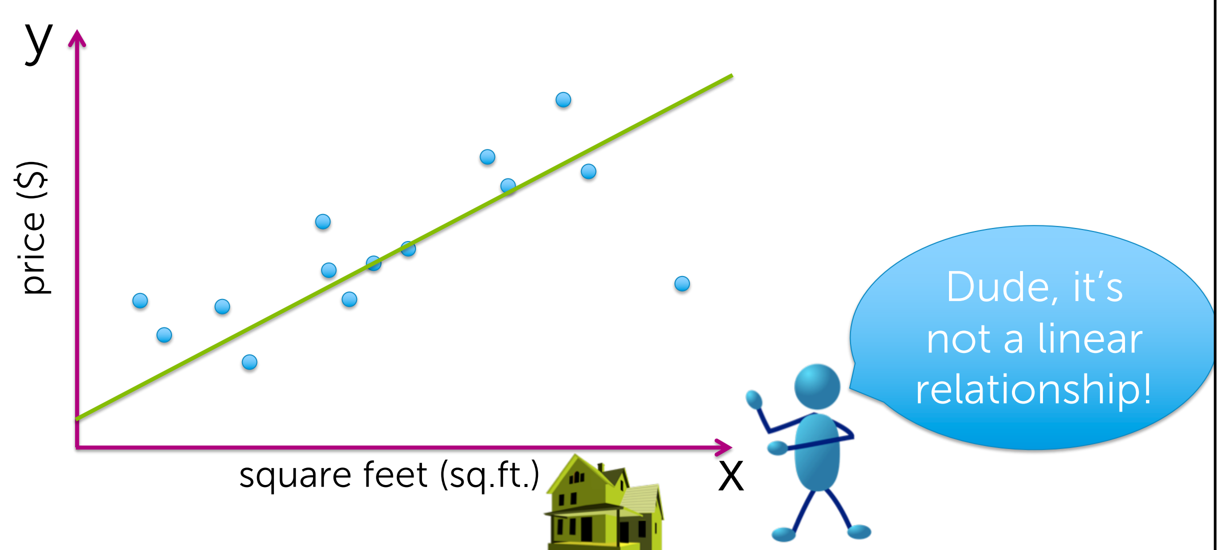
- House was measured in square meters?
- Price was measured in RMB?

Adding higher order effects

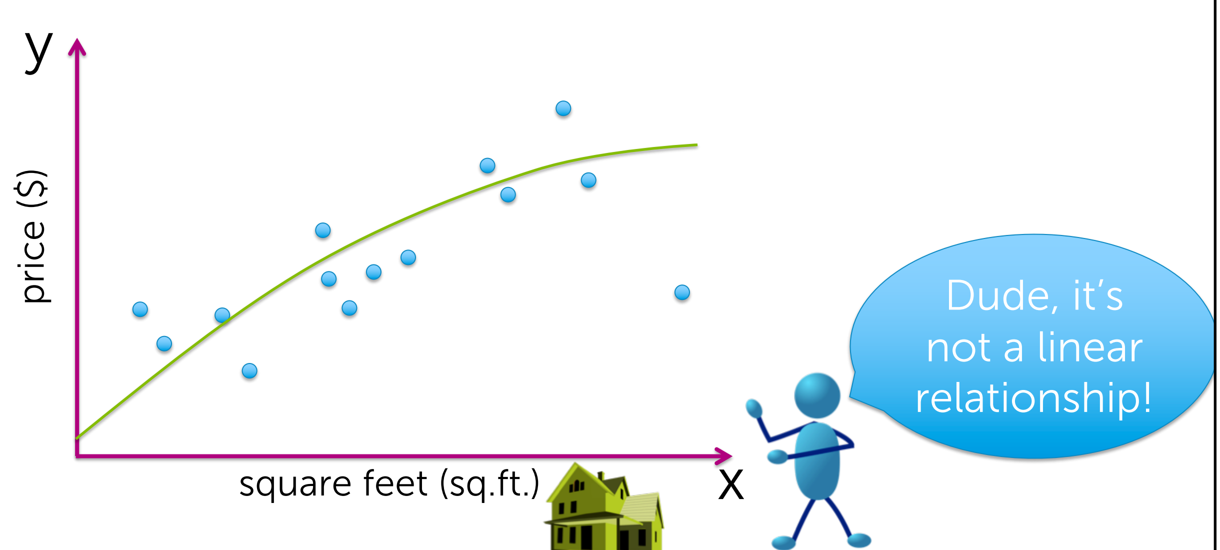
Fit data with a line or … ?



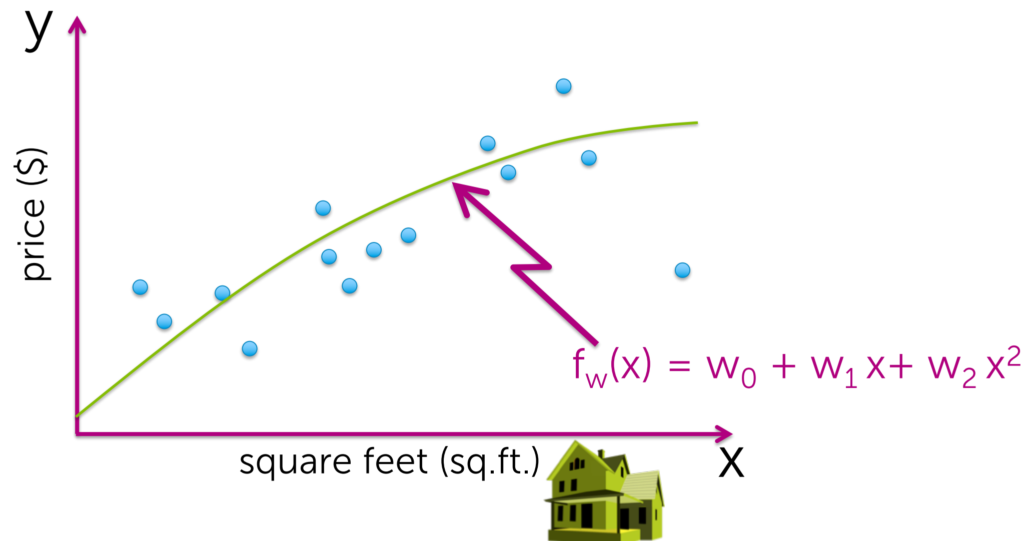
Fit data with a line or … ?



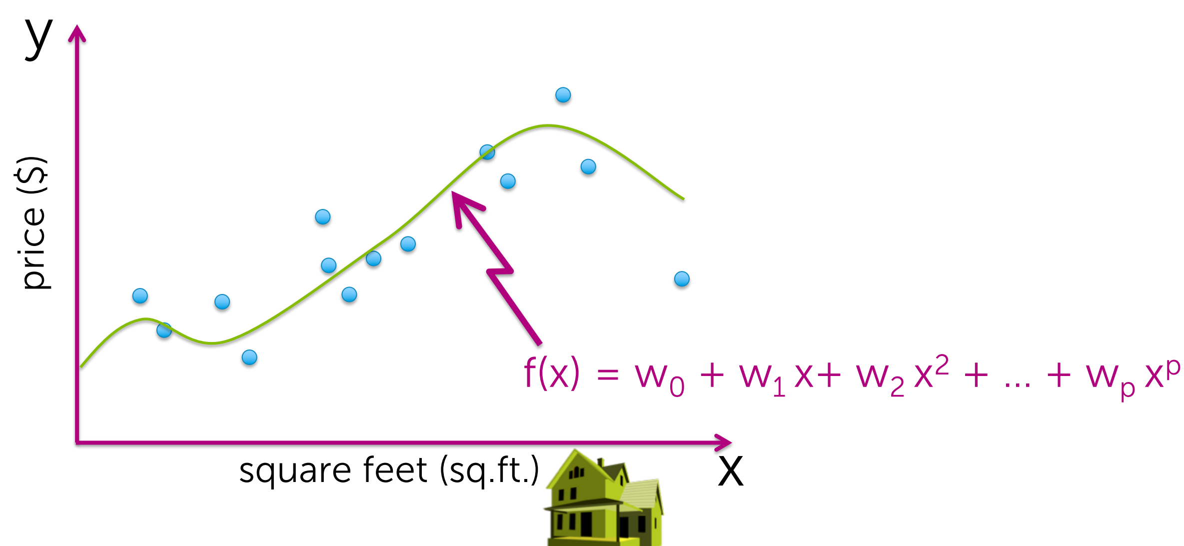
What about a quadratic function?



What about a quadratic function?



Even higher order polynomial



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*

Generic basis expansion

Model:

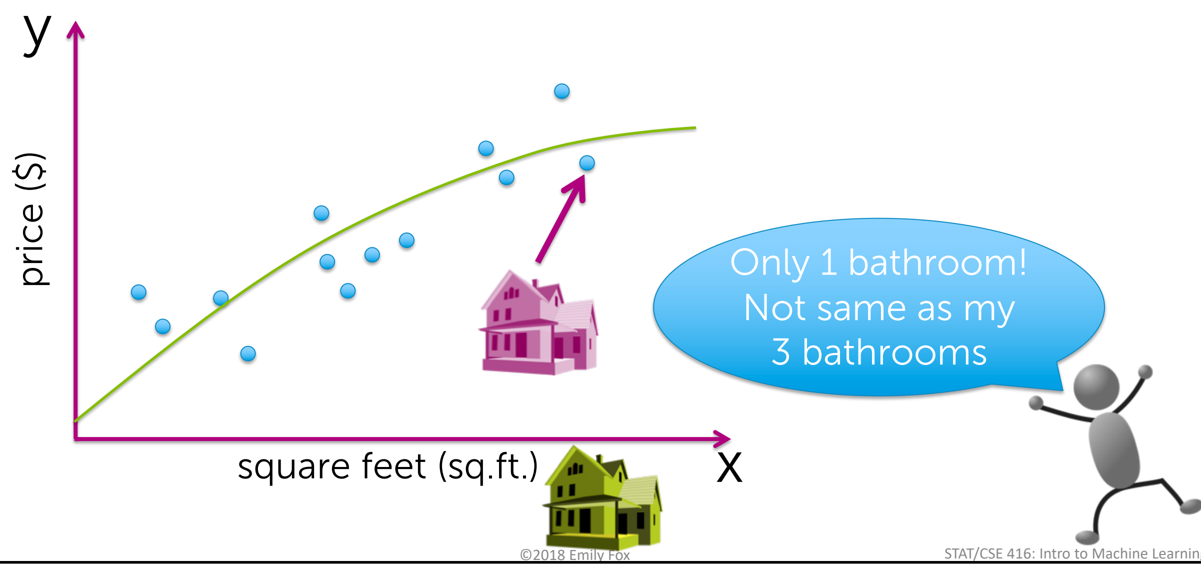
$y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + \ldots + w_D h_D(x_i) + \varepsilon_i$

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

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 or $\sin(2\pi x/12)$ or $\log(x)$
…
feature D+1 = $h_D(x)$… e.g., x^p

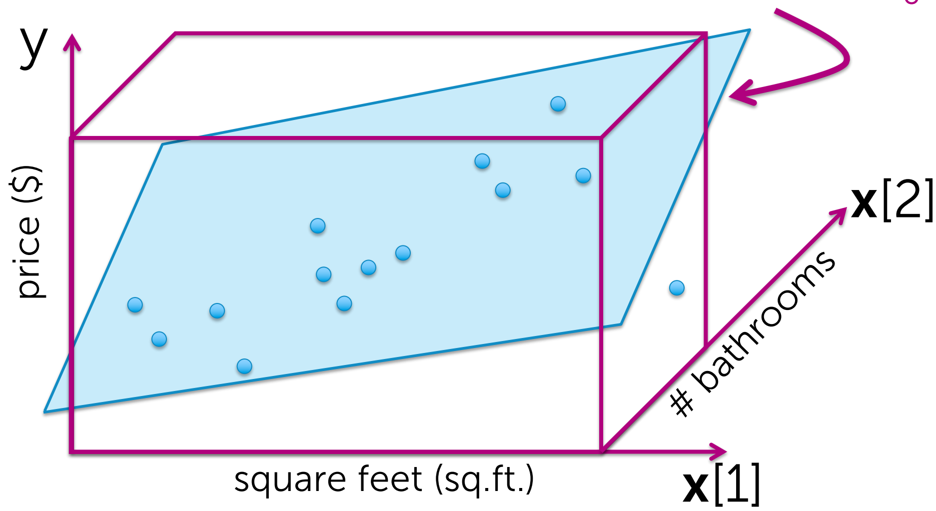
Adding other features

Predictions just based on house size



Add more inputs

$$f(\mathbf{x}) = w_0 + w_1 \text{ sq.ft.} + w_2 \text{ \#bath}$$



Many possible inputs

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

General notation

Output: y ← scalar
Inputs: $\mathbf{x} = (\mathbf{x}[1], \mathbf{x}[2], \ldots, \mathbf{x}[d])$
d-dim vector

Notational conventions:
$\mathbf{x}[j]$ = j^{th} input (*scalar*)
$h_j(\mathbf{x})$ = j^{th} feature (*scalar*)
$\mathbf{x}_i$ = input of i^{th} data point (*vector*)
$\mathbf{x}_i[j]$ = j^{th} input of i^{th} data point (*scalar*)

Generic linear regression model

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 $\mathbf{x}[1],\ldots,\mathbf{x}[d]$

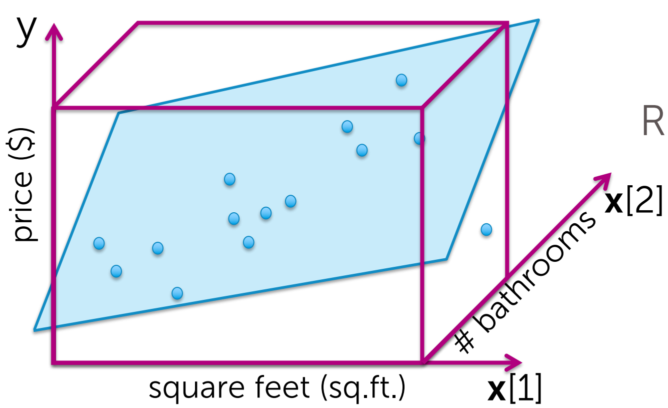
More on notation

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

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

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

RSS for multiple regression



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

How many features to use?

- More on this soon!

A compact representation

Representing function using vectors

$$f(\mathbf{x}_i) = w_0\, h_0(\mathbf{x}_i) + w_1\, h_1(\mathbf{x}_i) + \ldots + w_D\, h_D(\mathbf{x}_i) = \sum_{j=0}^{D} w_j\, h_j(\mathbf{x}_i)$$

1	0	0	0	5	3	0	0	1	0	0	0	0

3	0	0	0	2	0	0	1	0	1	0	0	0

As an algorithm…

$$f(\mathbf{x}_i) = w_0\, h_0(\mathbf{x}_i) + w_1\, h_1(\mathbf{x}_i) + \ldots + w_D\, h_D(\mathbf{x}_i) = \sum_{j=0}^{D} w_j\, h_j(\mathbf{x}_i)$$

1	0	0	0	5	3	0	0	1	0	0	0	0

3	0	0	0	2	0	0	1	0	1	0	0	0

Algorithm:

for j=0,…,D

Compact notation

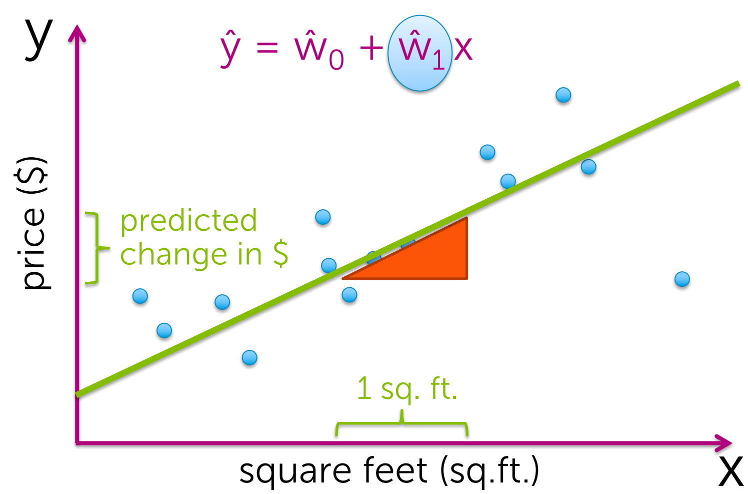
$$f(\mathbf{x}_i) = w_0\, h_0(\mathbf{x}_i) + w_1\, h_1(\mathbf{x}_i) + \ldots + w_D\, h_D(\mathbf{x}_i) = \sum_{j=0}^{D} w_j\, h_j(\mathbf{x}_i)$$

1	0	0	0	5	3	0	0	1	0	0	0	0

3	0	0	0	2	0	0	1	0	1	0	0	0

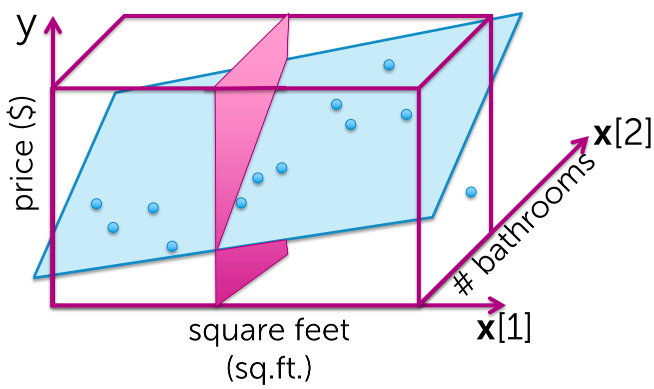
Interpreting the fitted function

Interpreting the coefficients – Simple linear regression



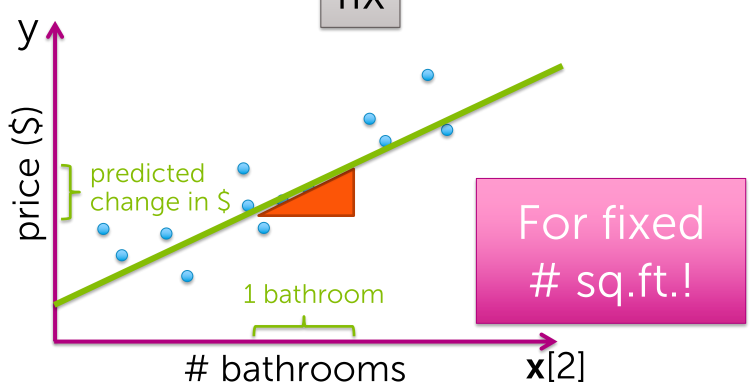
Interpreting the coefficients – Two linear features

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



Interpreting the coefficients – Two linear features

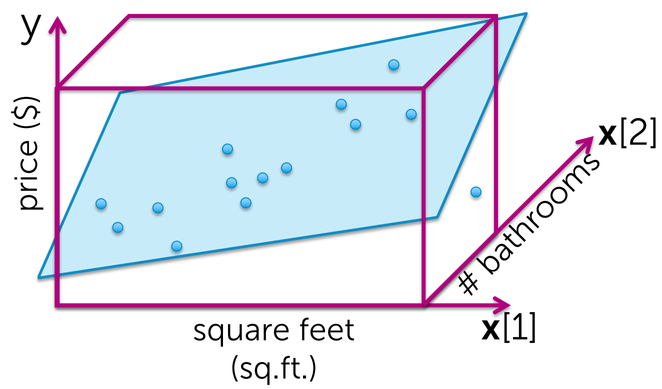
$$\hat{y} = \hat{w}_0 + \hat{w}_1 \underbrace{\boxed{\mathbf{x}[1]}}_{\text{fix}} + \hat{w}_2\, \mathbf{x}[2]$$



Interpreting the coefficients – Multiple linear features

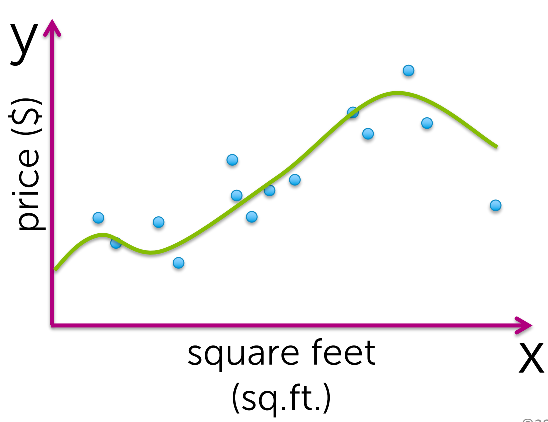
$$\hat{y} = \hat{w}_0 + \hat{w}_1 \boxed{\mathbf{x}[1]} + \boxed{\ldots} + \hat{w}_j \mathbf{x}[j] + \boxed{\ldots} + \hat{w}_d \boxed{\mathbf{x}[d]}$$

fix fix fix fix



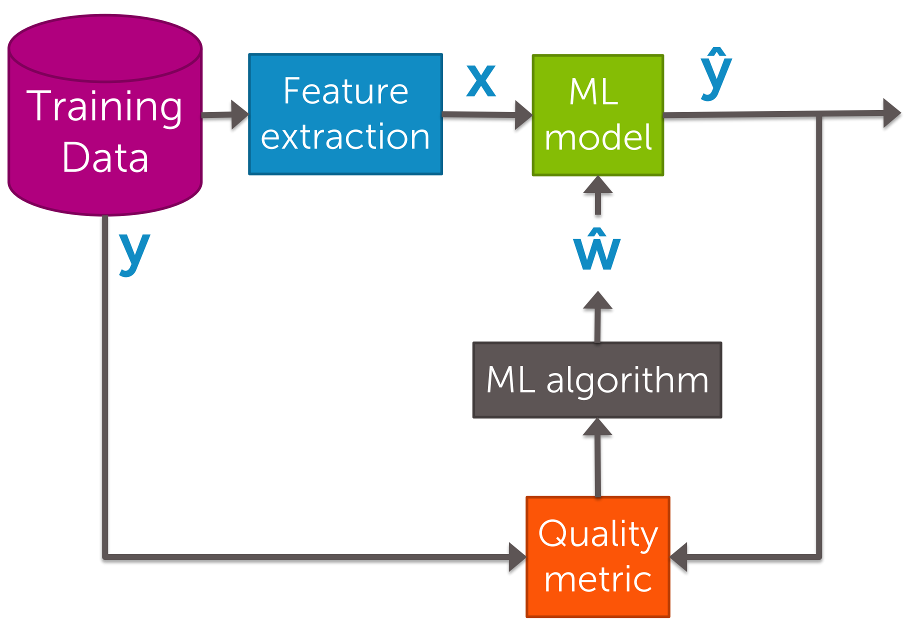
Interpreting the coefficients- Polynomial regression

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x + \ldots + \hat{w}_j x^j + \ldots + \hat{w}_p x^p$$



Can't hold other features fixed!

Summary for regression



What you can do now…

- Describe the input (features) and output (real-valued predictions) of a regression model
- Calculate a goodness-of-fit metric (e.g., RSS)
- Understand how gradient descent is used to estimate model parameters by minimizing RSS
- Exploit the estimated model to form predictions
- Describe a regression model using multiple features
- Interpret coefficients in a regression model with multiple features
- Describe other applications where regression is useful