Regression

Spring 2018

Review

- Last week:
  - Why machine learning?
  - What does a machine learn?
  - Designing a “supervised” machine learning algorithm

- Assignments (Canvas)
  - Problem Set 1 due yesterday
  - Lab Assignment 1 out

- Questions?

Today's Topics

- Regression Applications
- Discussion: Building Regression Datasets
- Linear Regression
- Polynomial Regression
- Regularization (Ridge Regression and Lasso Regression)
- Evaluating Regression Models
- Lab

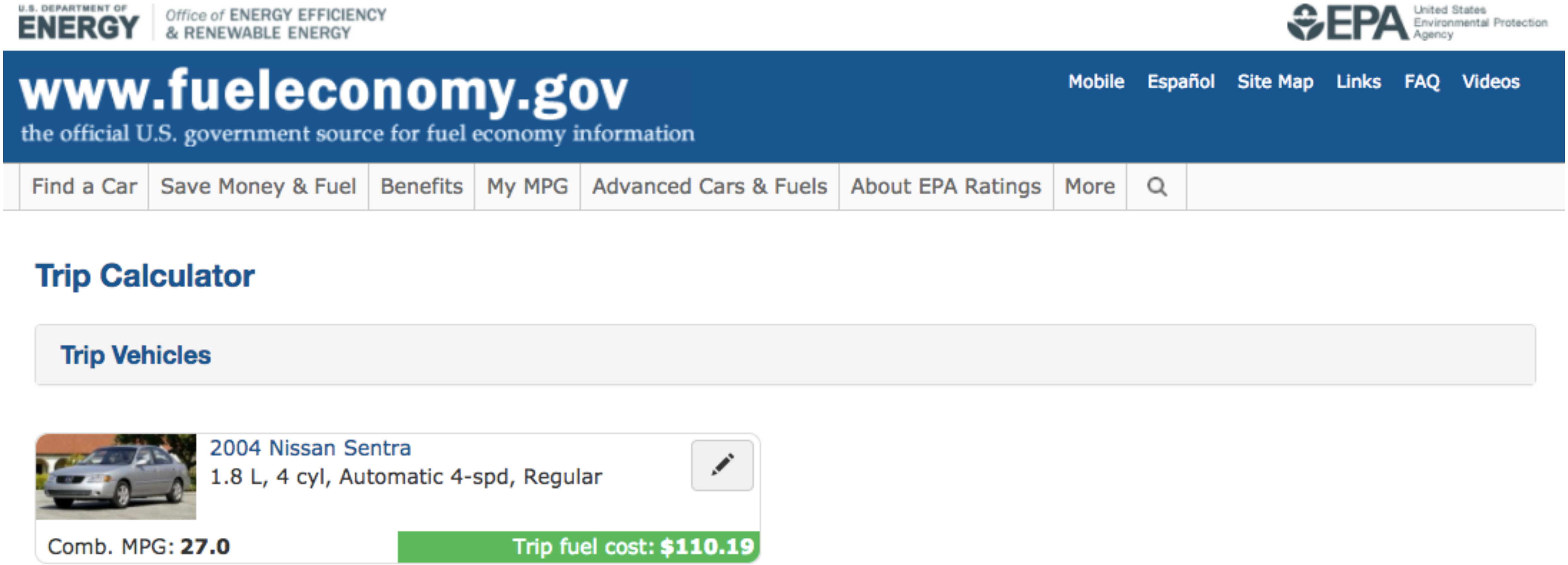
Today’s Topics

- Regression Applications
- Discussion: Building Regression Datasets
- Linear Regression
- Polynomial Regression
- Regularization (Ridge Regression and Lasso Regression)
- Evaluating Regression Models
- Lab

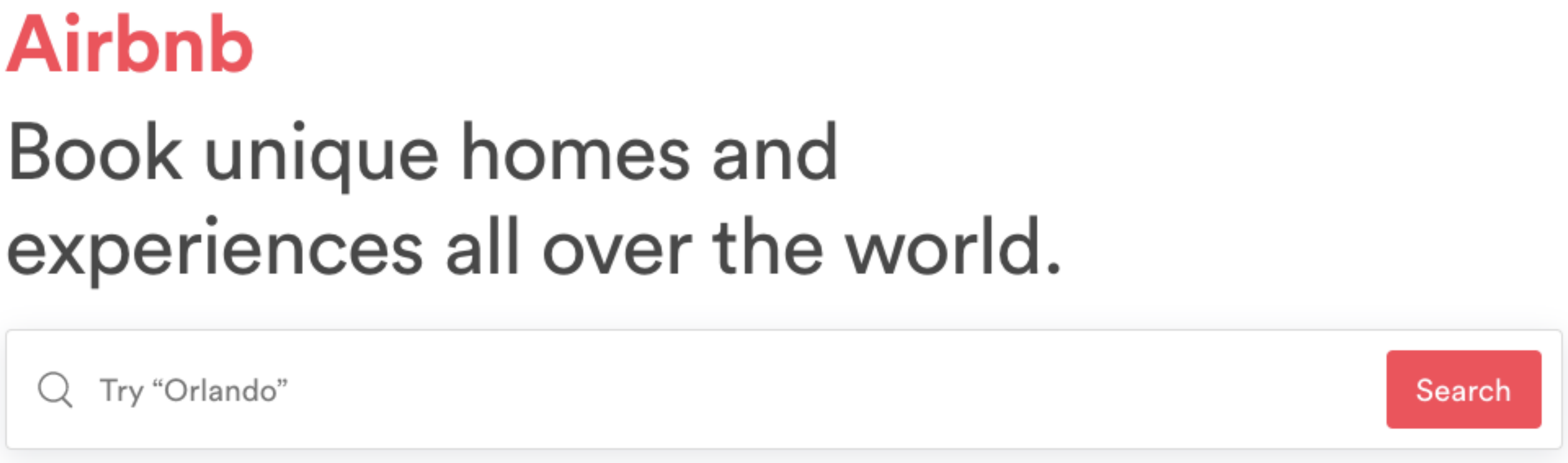
Today’s Focus: Regression

Predict **continuous** value

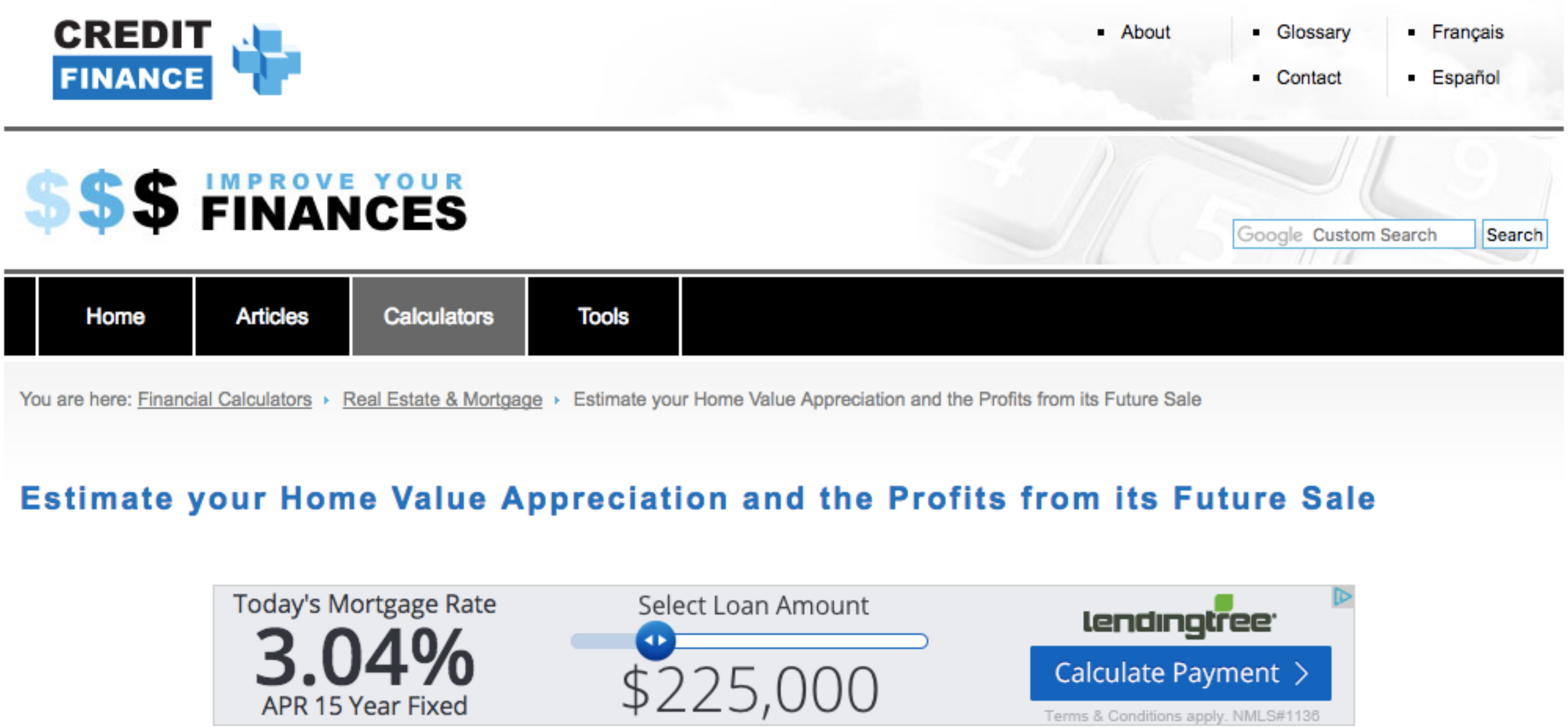
Predict Road Trip Fuel Cost



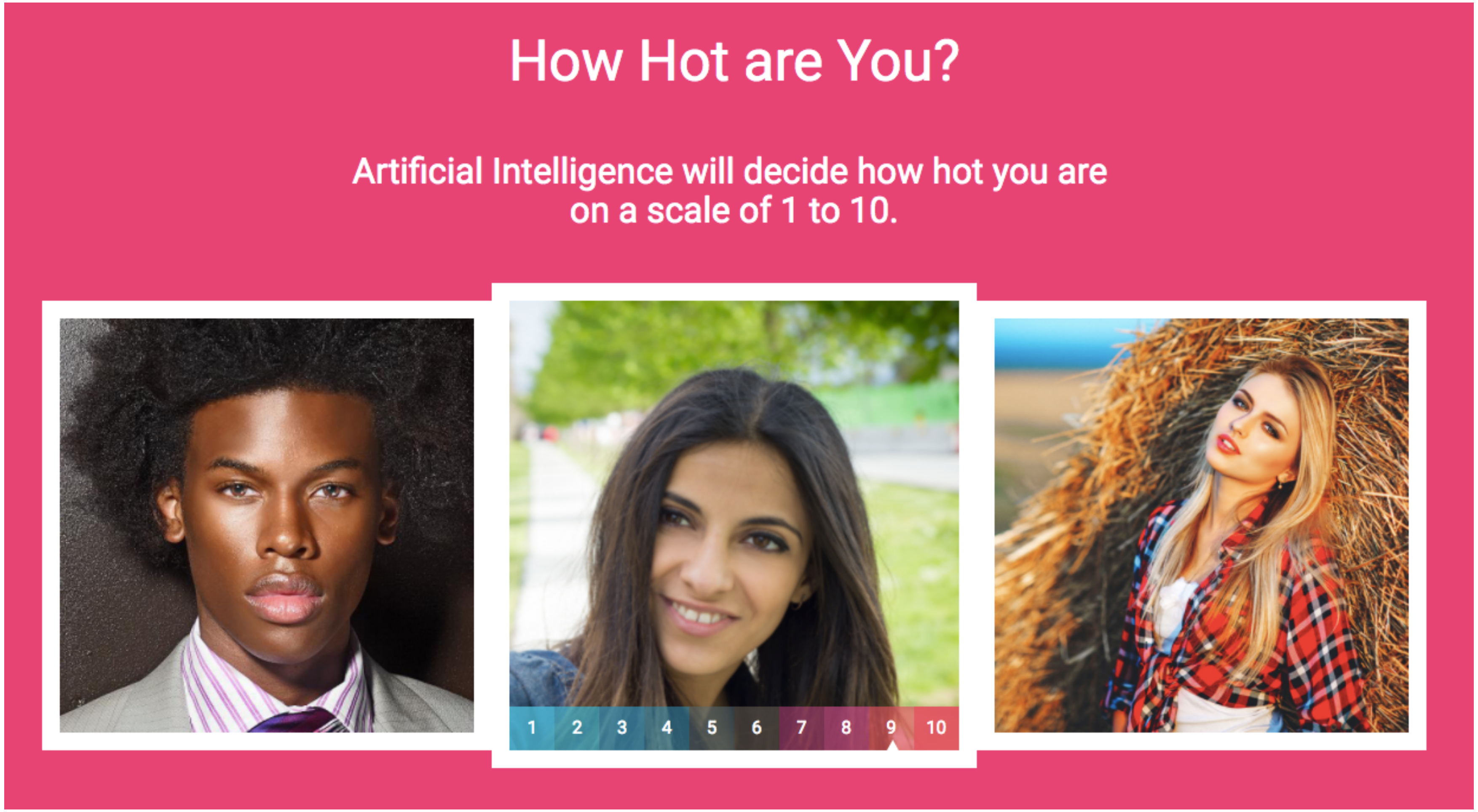
Predict Price to Charge for Your Home



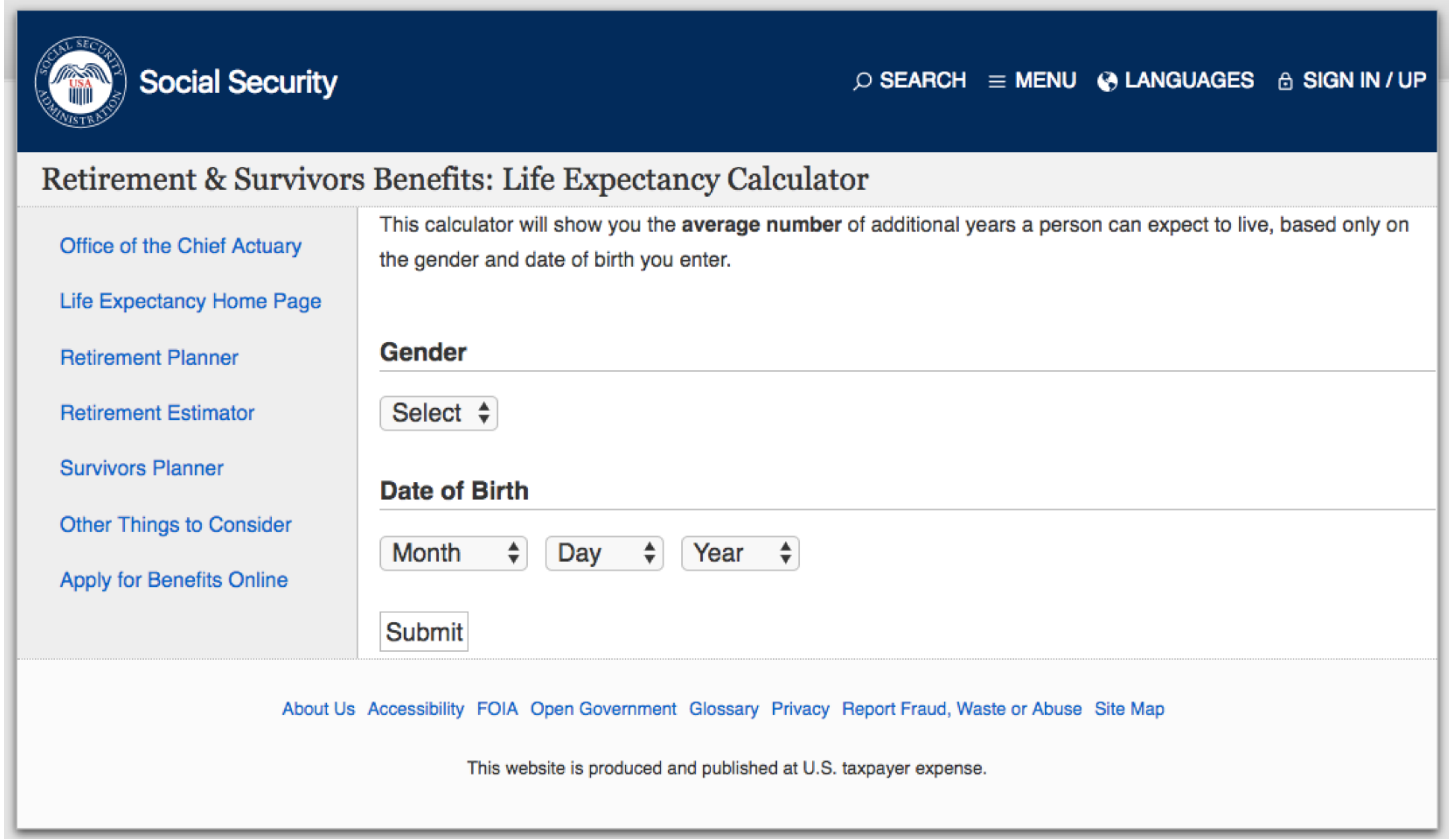
Predict Future Value of a House You Buy



Predict Perceived "Hot"-ness



Predict Life Expectancy



Predict Future Stock Price

QUANT INSTI GO ALGO Pioneer Institute in Algo Trading

HOME ABOUT US EPAT ™ PLACEMENT RESOURCES WEBINARS BLOG CONTACT US

Home > Blog > Trading Strategies

Machine Learning For Trading – How To Predict Stock Prices Using Regression?

What Else to Predict?

Insurance Cost

Popularity of Social Media Posts

Public Opinion

Factory Analysis

Political Party Preference

Call Center Complaints

Weather

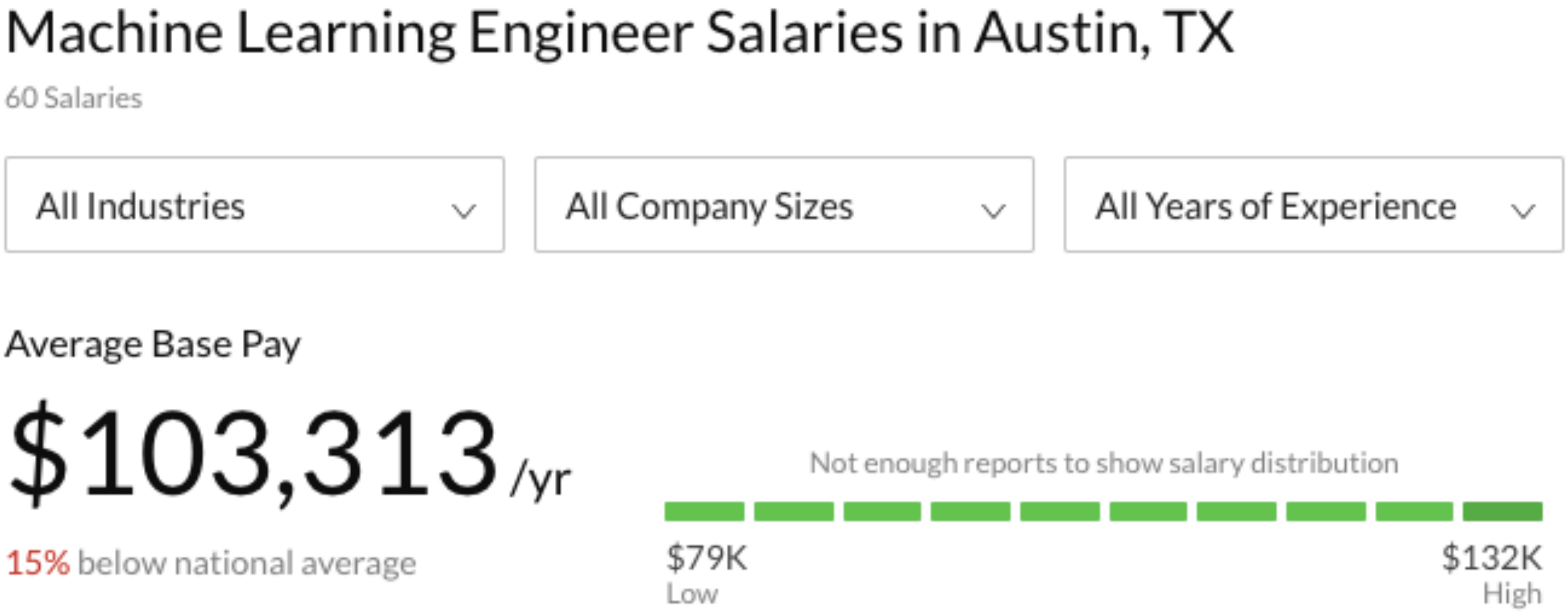
Class Ratings

Animal Behavior

Today’s Topics

- Regression Applications
- Discussion: Building Regression Datasets
- Linear Regression
- Polynomial Regression
- Regularization (Ridge Regression and Lasso Regression)
- Evaluating Regression Models
- Lab

Class Task: Predict Your Salary if You Become a Machine Learning Engineer



https://www.glassdoor.com/Salaries/machine-learning-engineer-salary-SRCH_KO0,25.htm

Class Task: Predict Your Salary if You Become a Machine Learning Engineer

- What cues would be predictive of your salary?
- Where can you find the data (predictive cues + true labels)?
- What would introduce noise to your data?

Class Task: Predict Your Salary if You Become a Machine Learning Engineer

Each person enter 4 data samples into the following spreadsheet:

https://docs.google.com/spreadsheets/d/1M_-qPmvWmA_uCAOWCuRKi18upaZILUmD2zsa6yTb5H8/edit?usp=sharing

Today’s Topics

- Regression Applications
- Discussion: Building Regression Datasets
- Linear Regression
- Polynomial Regression
- Regularization (Ridge Regression and Lasso Regression)
- Evaluating Regression Models
- Lab

Linear Models for Regression

- General formula:

$$\boxed{\hat{y}} = \boxed{w[0]} * \boxed{x[0]} + \boxed{w[1]} * \boxed{x[1]} + \ldots + \boxed{w[p]} * \boxed{x[p]} + \boxed{b}$$

Feature vector: **x** = x[0], x[1], …, x[p]

- How many features are there?
  - p+1

Parameter vector to learn: **w** = w[0], w[1], …, w[p]

- How many parameters are there?
  - p+2

Predicted value

“Simple” Linear Regression Model

- Formula:

$$\boxed{\widehat{y}} = \boxed{w[0]} * \boxed{x[0]} + \boxed{b}$$

Feature vector

- How many features are there?
  - 1

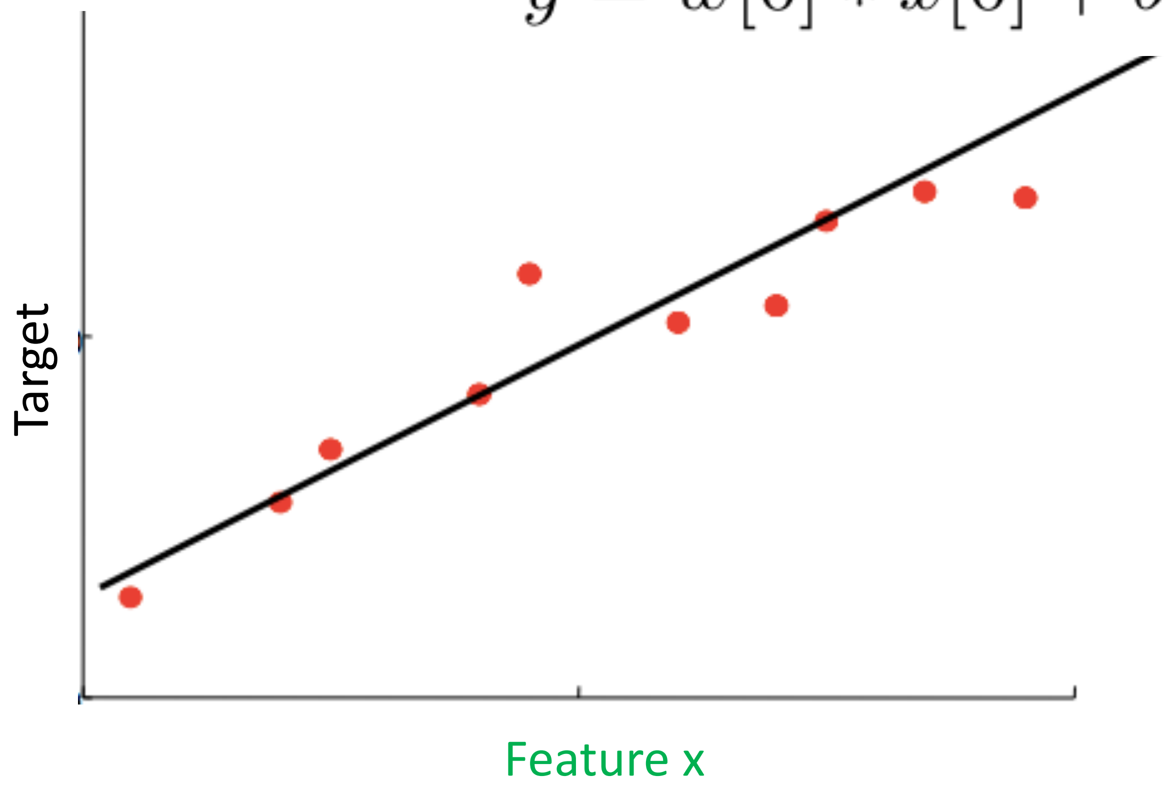
Parameter vector to learn

- How many parameters are there?
  - 2

Predicted value

(Line)

$$\widehat{y} = w[0] * x[0] + b$$



“Multiple” Linear Regression Model

- Formula:

$$\boxed{\hat{y}} = \boxed{w[0]} * \boxed{x[0]} + \boxed{w[1]} * \boxed{x[1]} + \boxed{b}$$

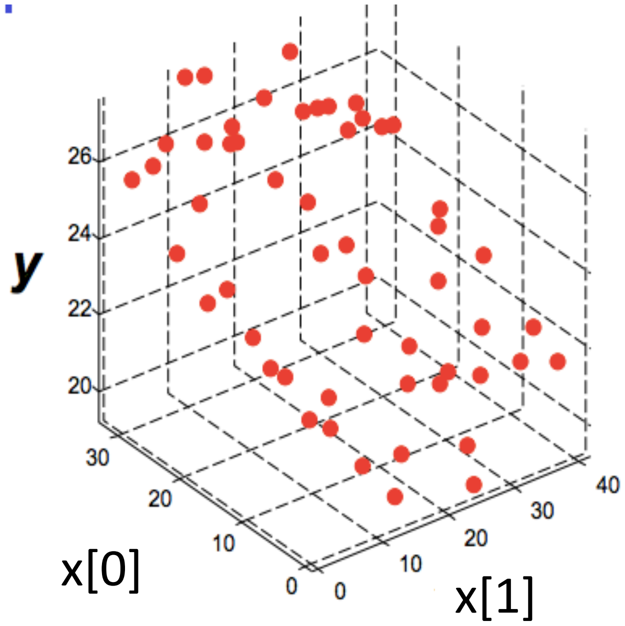
Feature vector

- How many features are there?
  - 2

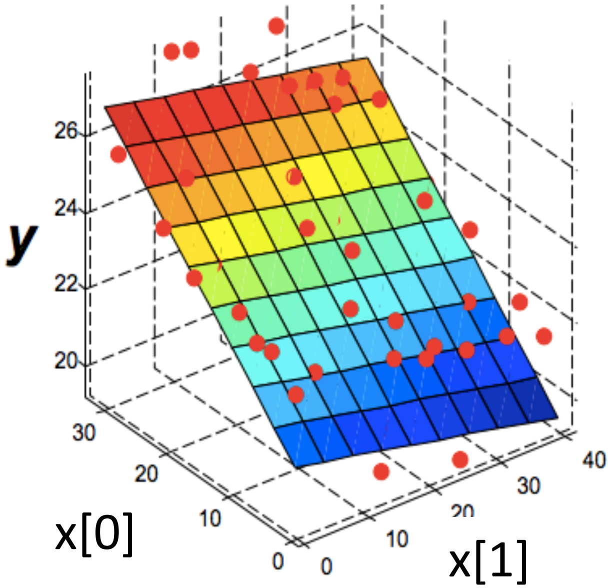
Parameter vector to learn

- How many parameters are there?
  - 3

Predicted value



(Plane)



Linear Model: Predict Salary as a ML Engineer

(Solution is a hyperplane)

$$\widehat{y} = \boxed{w[0]} * x[0] + \boxed{w[1]} * x[1] + \ldots + \boxed{w[p]} * x[p] + \boxed{b}$$

- How would you write the linear model equation?
- How would you weight the different predictive cues?

How to Learn Model Parameters?

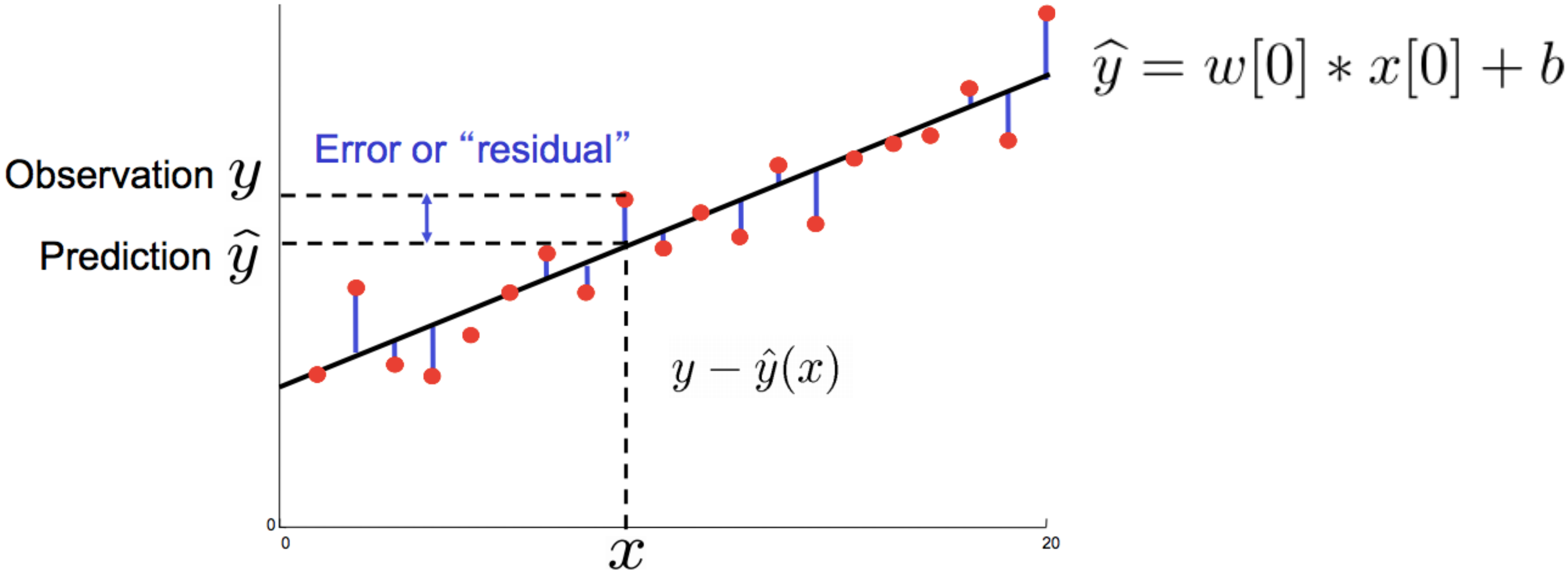
- Given: dataset (split into train and test partitions!)

Convention: X is $n \times d$ design matrix of sample pts
y is n-vector of scalars [constants]

$$\begin{array}{cc} \left[\begin{array}{cccccc} X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1d} \\ X_{21} & X_{22} & & X_{2j} & & X_{2d} \\ \vdots & & & & & \\ X_{i1} & X_{i2} & & X_{ij} & & X_{id} \\ \vdots & & & & & \\ X_{n1} & X_{n2} & & X_{nj} & & X_{nd} \end{array}\right] & \begin{array}{l} \\ \\ \\ \leftarrow \text{point } X_i^\top \\ \\ \\ \end{array} \\ \begin{array}{c} \uparrow \\ \text{feature column } X_{*j} \end{array} & \end{array} \qquad \begin{array}{c} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \\ \uparrow \\ y \end{array}$$

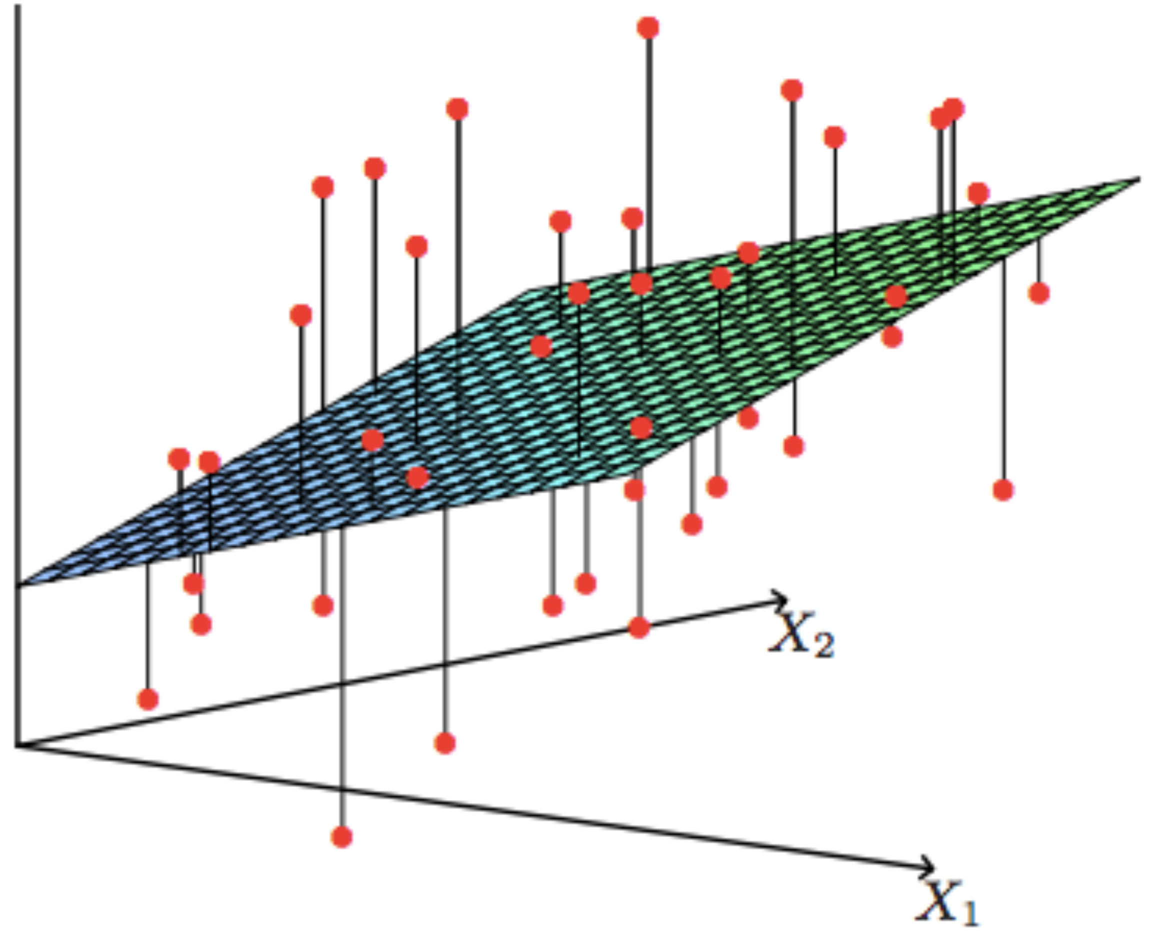
How to Learn Model Parameters?

- (Gauss, 1801) Least squares: *minimize* total squared error ("residual")
  - Why square the error?



How to Learn Model Parameters?

- (Gauss, 1801) Least squares: *minimize* total squared error ("residual")



How to Learn Model Parameters?

- (Gauss, 1801) Least squares: *minimize* total squared error ("residual")
  - Take derivatives, set to zero, and solve for parameters

$$\frac{\partial}{\partial w}\sum_i (y_i - wx_i)^2 = 2\sum_i - x_i(y_i - wx_i) \Rightarrow$$

$$2\sum_i x_i(y_i - wx_i) = 0 \Rightarrow$$

$$\sum_i x_i y_i = \sum_i wx_i^2 \Rightarrow$$

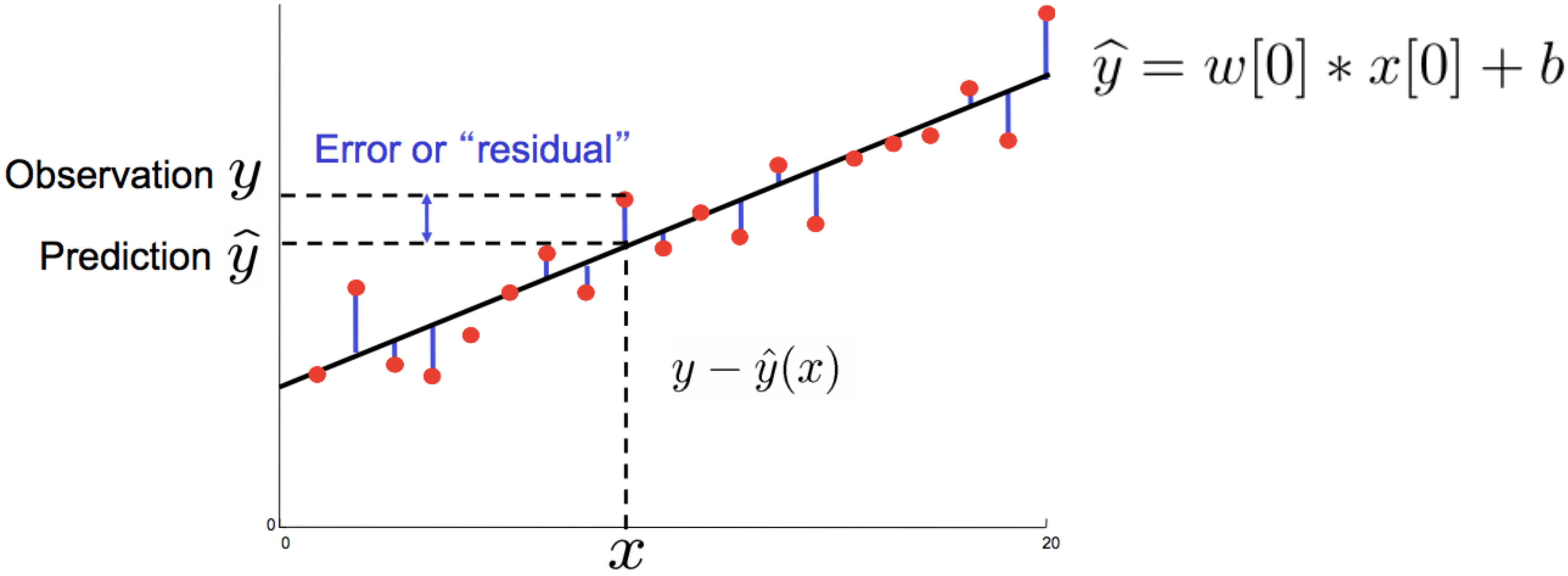
$$w = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

How to Learn Model Parameters?

- (Gauss, 1801) Least squares: *minimize* total squared error (“residual”)
  - Gradient Descent: online learning solution
    - Slower
    - Not guaranteed to find optimal solution
    - But works for large datasets!

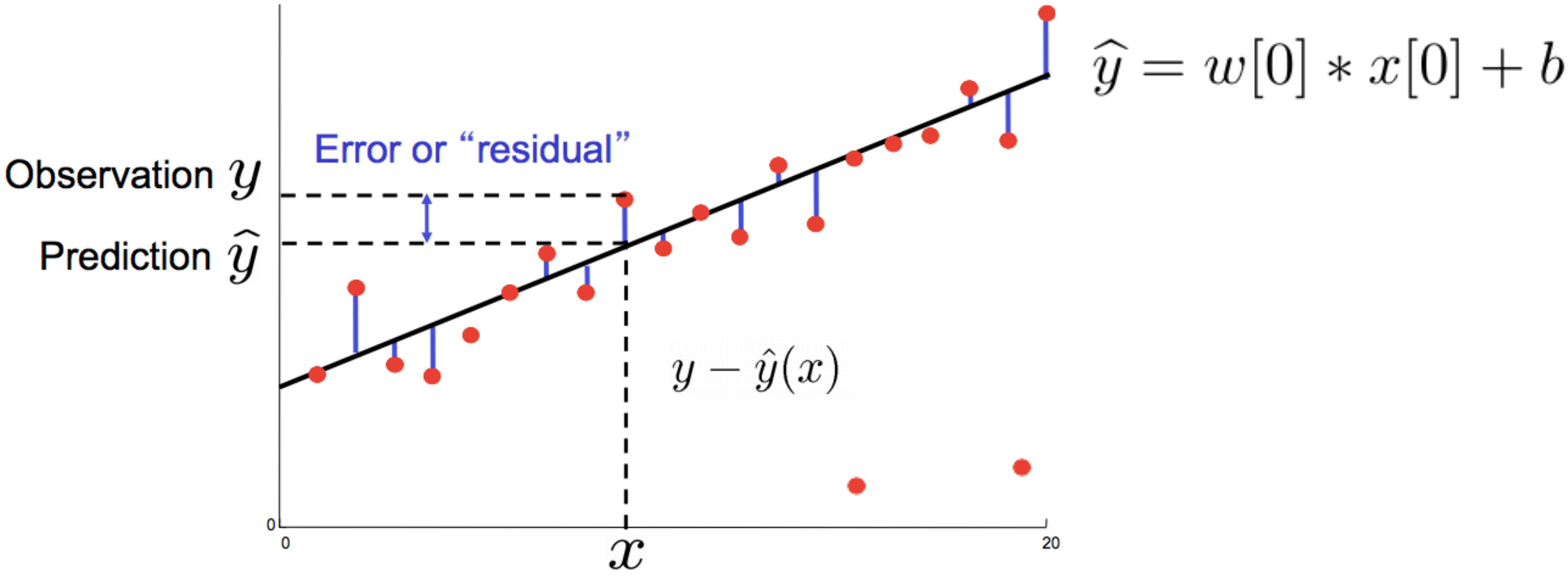
What Does the Error Represent?

- (Gauss, 1801) Least squares: *minimize* total squared error ("residual")
  - Noise



What Happens to the Learned Model in the Presence of Outliers?

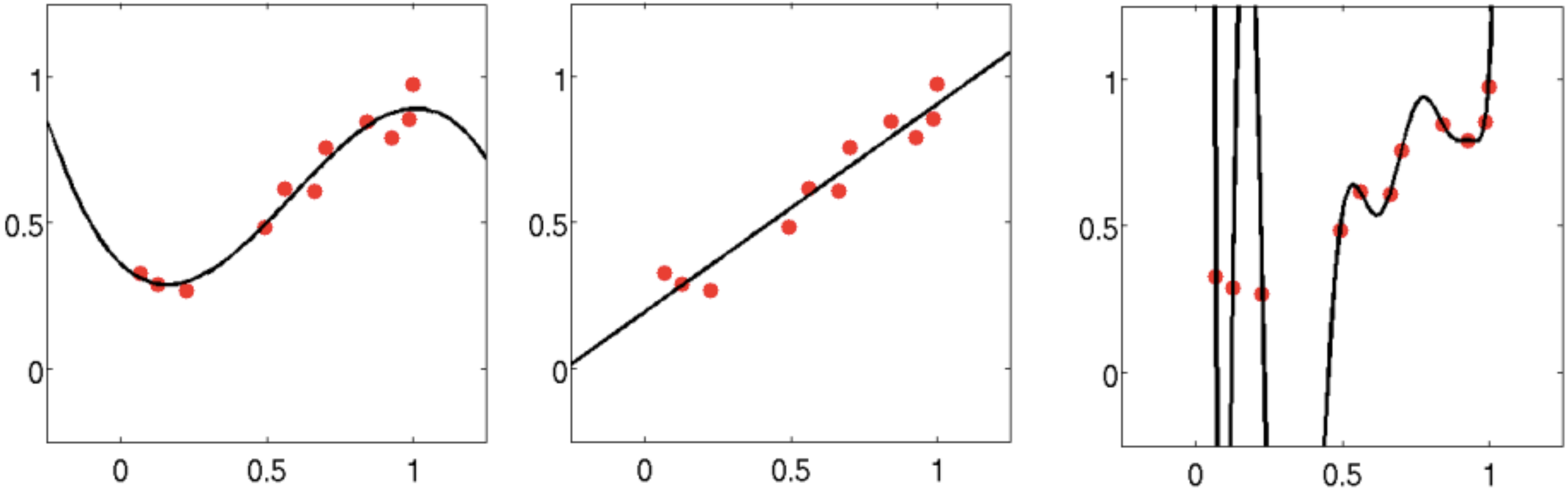
- (Gauss, 1801) Least squares: *minimize* total squared error (“residual”)



Today’s Topics

- Regression Applications
- Discussion: Building Regression Datasets
- Linear Regression
- Polynomial Regression
- Regularization (Ridge Regression and Lasso Regression)
- Evaluating Regression Models
- Lab

What if Linear Models Are Not Good Enough?



Model Non-Linear Relationships by Transforming Features

- e.g., (Recall) Formula:

$$\boxed{\hat{y}} = \boxed{w[0]} * \boxed{x[0]} + \boxed{w[1]} * \boxed{x[1]} + \boxed{b}$$

- e.g., New Formula:

$$\hat{y} = \boxed{w[0]} * \boxed{x[0]} + \boxed{w[1]} * \boxed{x[0]^2} + \boxed{b}$$

Predicted value

Parameter vector

Feature vector

  - **Same model parameters...**
  - **Still a linear model!**
  - **But can now model more complex relationships!!**

Model Non-Linear Relationships by Transforming Features

- e.g., feature conversion for polynomial degree 3

$$D = \{(x^{(j)}, y^{(j)})\} \longrightarrow D = \{([x^{(j)}, (x^{(j)})^2, (x^{(j)})^3], y^{(j)})\}$$

- e.g., What is the new feature vector with polynomial degree up to 3?

$$\begin{array}{l} \textit{Example 1:} \\ \textit{Example 2:} \\ \textit{Example 3:} \end{array} \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \longrightarrow \begin{array}{l} \textit{Example 1:} \\ \textit{Example 2:} \\ \textit{Example 3:} \end{array} \begin{bmatrix} 2 & 4 & 8 \\ 3 & 9 & 27 \\ 4 & 16 & 64 \end{bmatrix}$$

Basis Functions to Transform Features: Polynomial and Beyond…

- General idea: **project data into a higher dimension** to fit more complicated relationships to a linear fit
- How to **project data into a higher dimension?**

e.g.,

Polynomial: $\phi_j(x) = x^j$ for j=0 … n

Gaussian: $$\phi_j(x) = \frac{(x-\mu_j)}{2\sigma_j^2}$$

Sigmoid: $$\phi_j(x) = \frac{1}{1+\exp(-s_j x)}$$

Polynomial Function to Transform Features

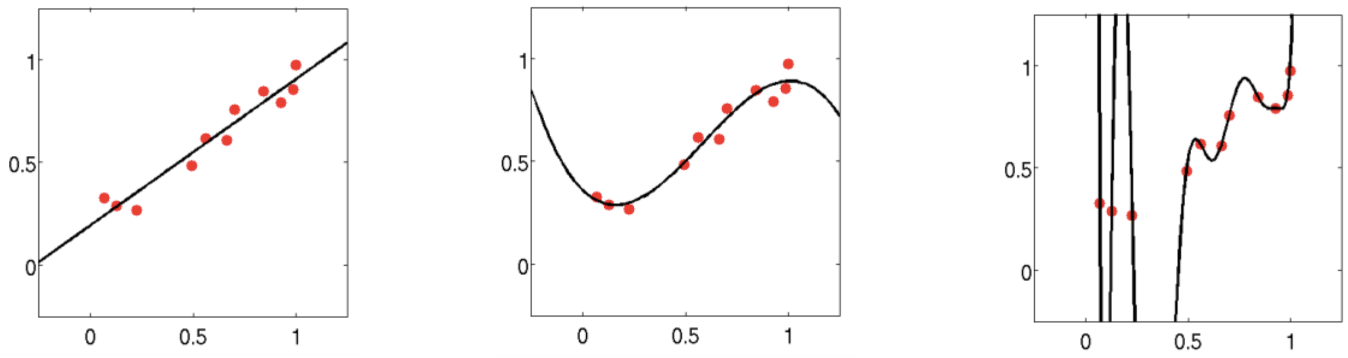
- M-th order polynomial function:

$$y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j x^j$$

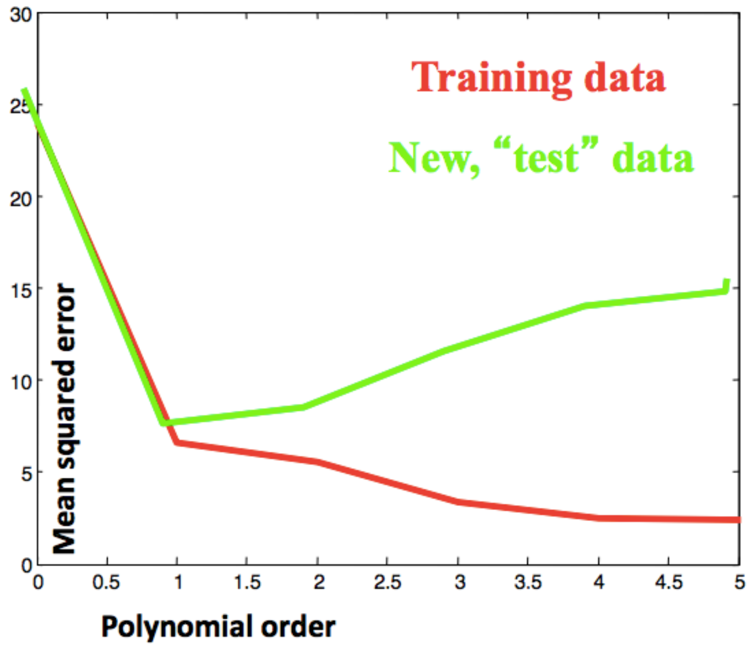
- Linear model… so still learn model parameters by solving same analytical/gradient descent methods as discussed in previous section

Choosing a Feature Transformation

- Are more features better?

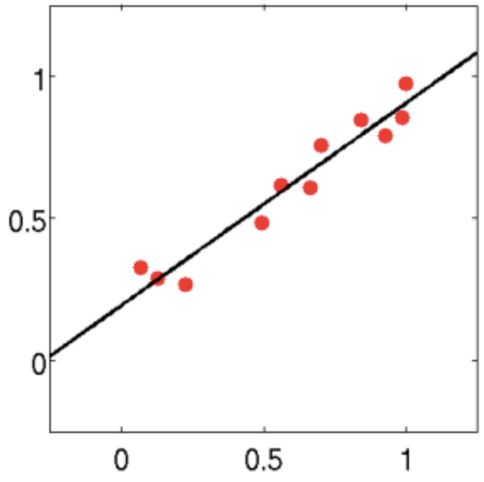


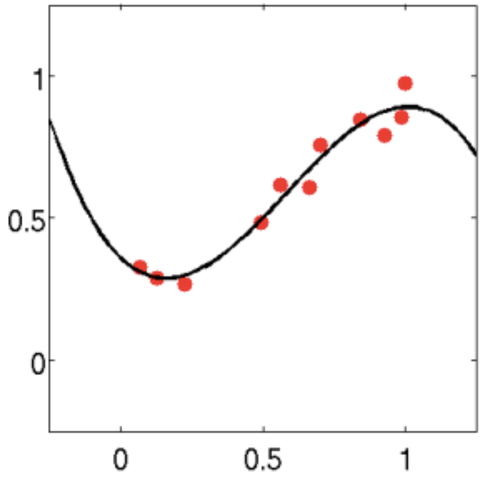
- Example plot of error on training dataset and testing dataset:
    - What happens to training data error with larger polynomial order?
    - What happens to testing data error with larger polynomial order?
        - Recall data has noise
        - Higher order is more able to model this noise!
        - Higher order is more likely therefore to **"overfit"** to training data and so not generalize to new unobserved test data

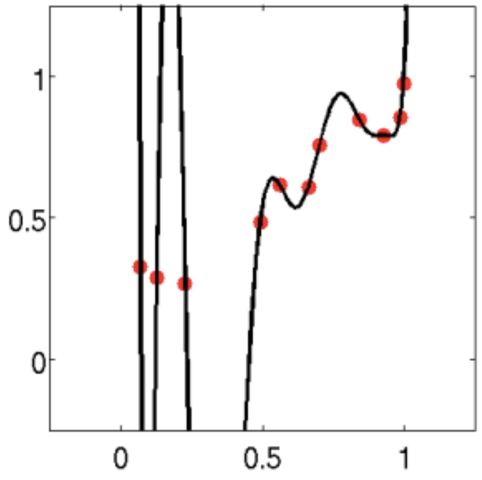


How to Avoid Overfitting?

- Use lower degree polynomial:



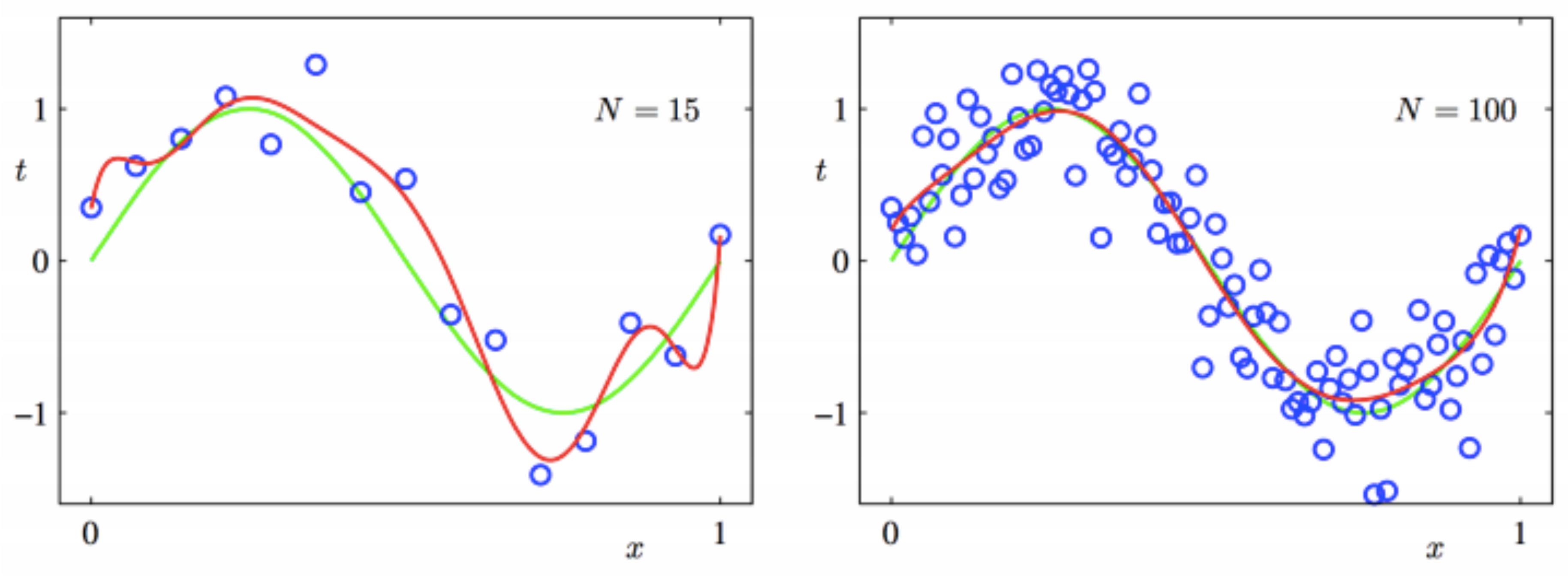




- Risk: may be underfitting again

How to Avoid Overfitting?

- Add more training data



- What are the challenges/costs with collecting more training data?

Today’s Topics

- Regression Applications
- Discussion: Building Regression Datasets
- Linear Regression
- Polynomial Regression
- Regularization (Ridge Regression and Lasso Regression)
- Evaluating Regression Models
- Lab

How to Avoid Overfitting?

- Regularize model (add constraints)… but how?
  - e.g., weights learned for fitting a model to a sine wave function (polynomial degrees 0, 1, …, 9)

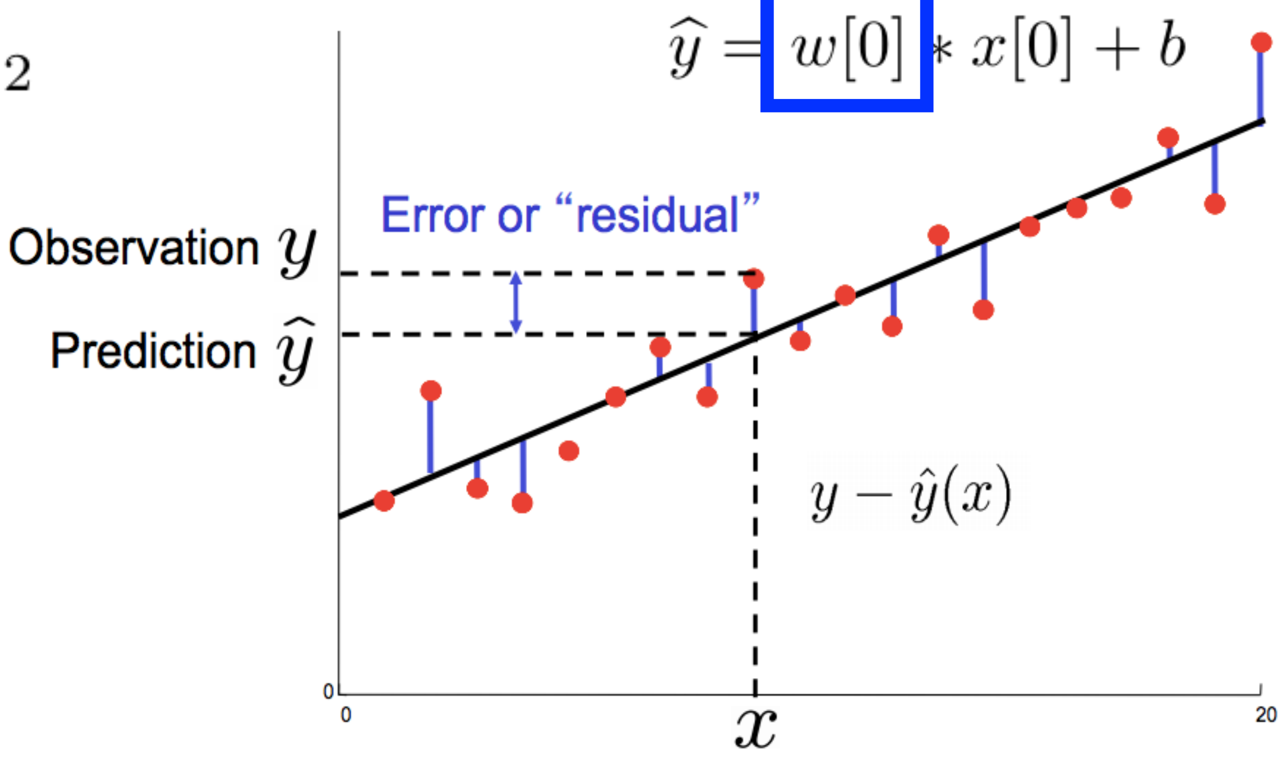
	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

  - Sign of overfitting: weights blow up and cancel each other out to fit the training data
- Idea: add constraint to minimize presence of large weights in models!

How to Avoid Overfitting?

- Idea: add constraint to minimize presence of large weights in models
- Recall: we previously learned models by *minimizing* **s**um of **s**quared **e**rrors (SSE) for all n training examples:

$$SSE = \sum_{i=1}^{n} (y^{(i)} - \widehat{y}^{(i)})^2$$



How to Avoid Overfitting?

- Idea: add constraint to $\boxed{\text{minimize presence of large weights in models}}$
- Recall: we previously learned models by *minimizing* **s**um of **s**quared **e**rrors (SSE) for all n training examples:

$$SSE = \sum_{i=1}^{n} (y^{(i)} - \widehat{y}^{(i)})^2$$

- Ridge Regression: add constraint to penalize squared weight values

$$Error = \sum_{i=1}^{n} (y^{(i)} - \widehat{y}^{(i)})^2 + \alpha \sum_{j=1}^{m} \boxed{w_j^2}$$

- Lasso Regression: add constraint to penalize absolute weight values

$$Error = \sum_{i=1}^{n} (y^{(i)} - \widehat{y}^{(i)})^2 + \alpha \sum_{j=1}^{m} \boxed{|w_j|}$$

How to Set Alpha?

Recall: $\hat{y} = \sum_{j=1}^{m} \boxed{w_j} x_j + w_0$

What happens when you set alpha to 0?

What happens when you set alpha to 1?

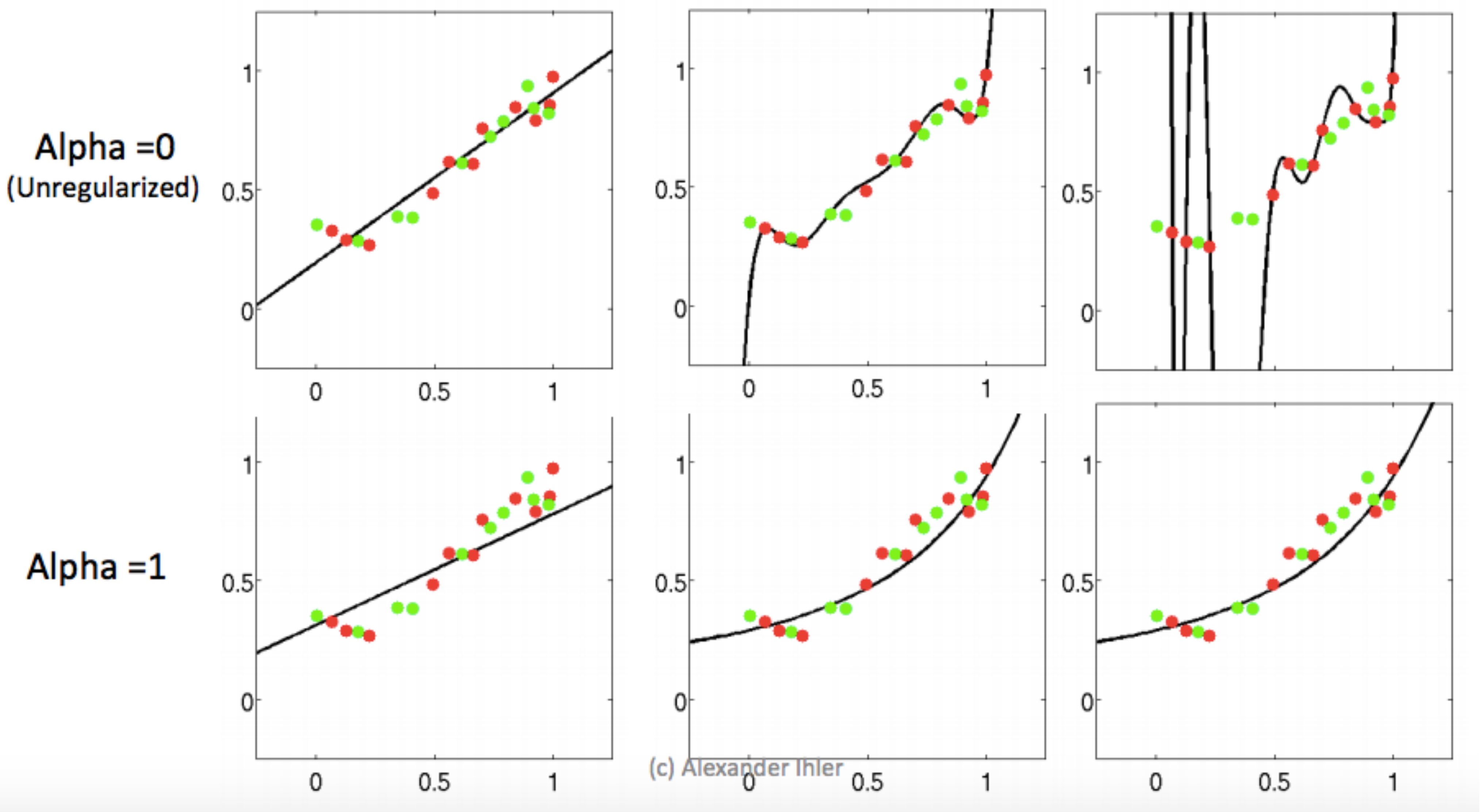
- Ridge Regression: add constraint to penalize squared weight values

$$Error = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 + \boxed{\alpha} \sum_{j=1}^{m} \boxed{w_j^2}$$

- Lasso Regression: add constraint to penalize absolute weight values

$$Error = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 + \boxed{\alpha} \sum_{j=1}^{m} \boxed{|w_j|}$$

How to Set Alpha?



Why Choose Lasso Instead vs Ridge Regression

- Lasso: typically creates sparse weight vectors (sets weights to 0)
  - Good to use when there are MANY features and few believed to be relevant
  - Increases interpretability

Today’s Topics

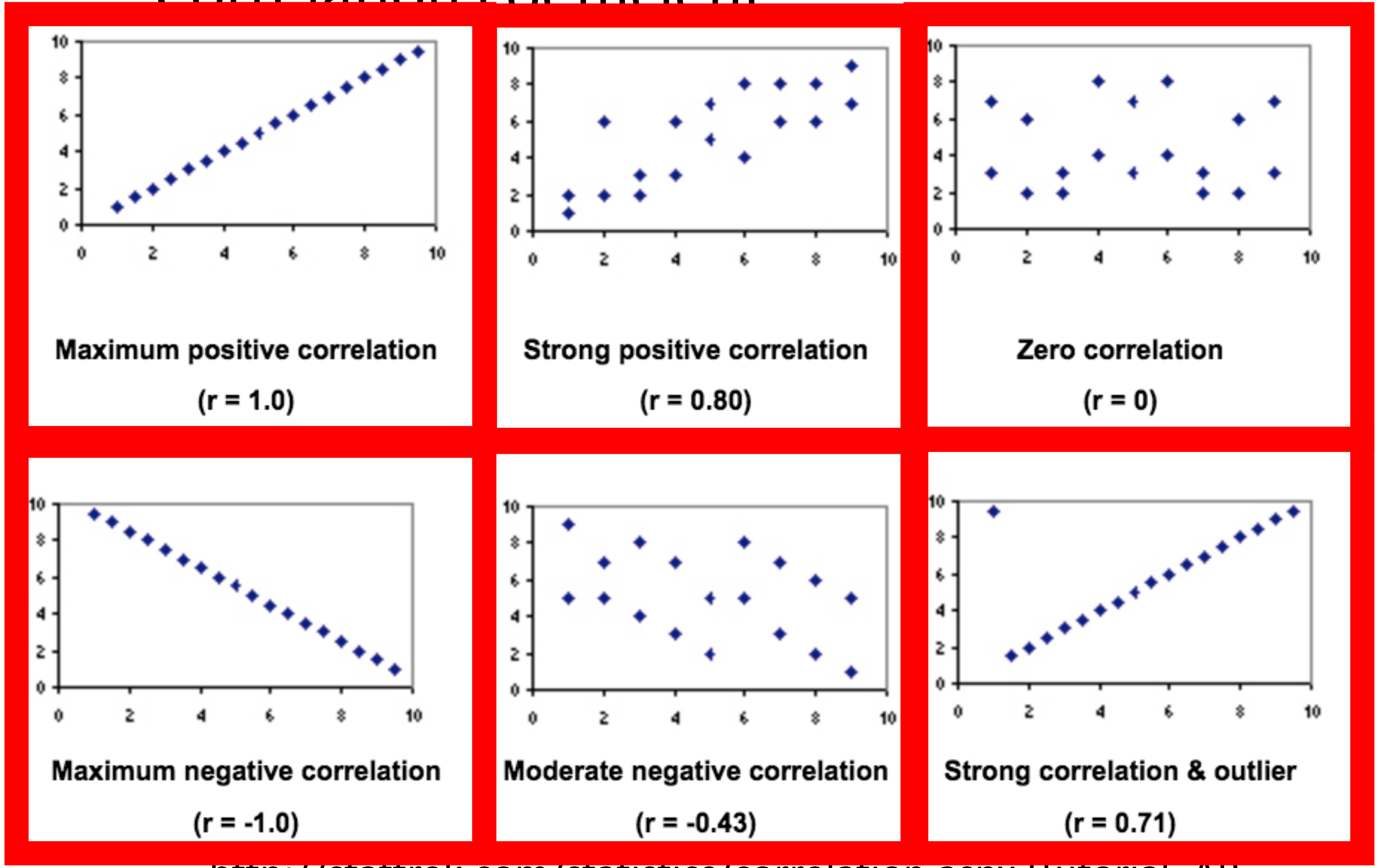
- Regression Applications
- Discussion: Building Regression Datasets
- Linear Regression
- Polynomial Regression
- Regularization (Ridge Regression and Lasso Regression)
- Evaluating Regression Models
- Lab

Evaluating regressors: descriptive statistics

Results: e.g.,

inst#	actual	predicted	error
1	0.18	0.272	0.092
2	0.122	0.434	0.312
3	0.088	0.344	0.256
4	0.125	0.238	0.112
5	0	0.232	0.232
6	0	0.092	0.092
7	0.907	0.367	-0.54
8	0.216	0.227	0.011
9	0	0.367	0.367
10	0.048	0.108	0.061
11	0.198	0.145	-0.053
12	0	0.159	0.159
13	0.505	0.28	-0.225
14	0.273	0.097	-0.175
15	0.12	0.178	0.058
16	0.254	0.235	-0.018

- Correlation coefficient



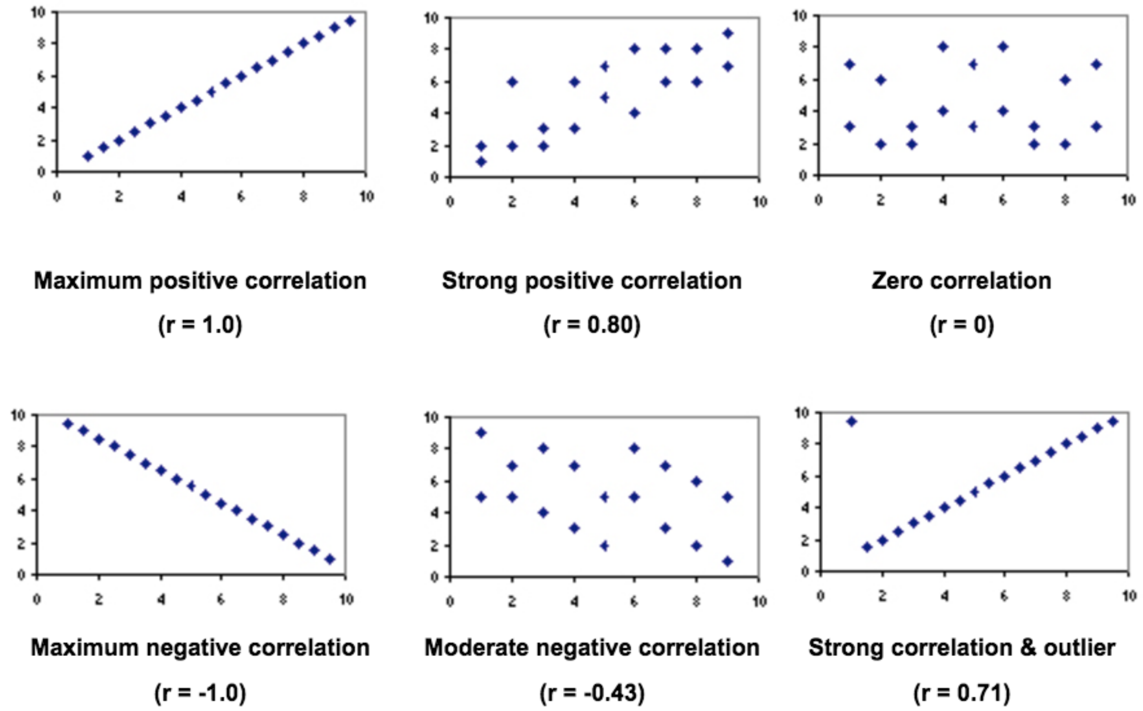
http://stattrek.com/statistics/correlation.aspx?Tutorial=AP

Evaluating regressors: descriptive statistics

Results: e.g.,

inst#	actual	predicted	error
1	0.18	0.272	0.092
2	0.122	0.434	0.312
3	0.088	0.344	0.256
4	0.125	0.238	0.112
5	0	0.232	0.232
6	0	0.092	0.092
7	0.907	0.367	-0.54
8	0.216	0.227	0.011
9	0	0.367	0.367
10	0.048	0.108	0.061
11	0.198	0.145	-0.053
12	0	0.159	0.159
13	0.505	0.28	-0.225
14	0.273	0.097	-0.175
15	0.12	0.178	0.058
16	0.254	0.235	-0.018

- Correlation coefficient
  - What is the range of possible values?
  - Are larger values better or worse?



http://stattrek.com/statistics/correlation.aspx?Tutorial=AP

Evaluating regressors: descriptive statistics

Results: e.g.,

inst#	actual	predicted	error
1	0.18	0.272	0.092
2	0.122	0.434	0.312
3	0.088	0.344	0.256
4	[illegible]	[illegible]	[illegible]
5	0	0.232	0.232
6	[illegible]	[illegible]	[illegible]
7	0.907	0.367	-0.54
8	0.216	0.227	0.011
9	0	0.367	0.367
10	0.048	0.108	0.061
11	0.198	0.145	-0.053
12	[illegible]	[illegible]	[illegible]
13	0.505	0.28	-0.225
14	[illegible]	[illegible]	[illegible]
15	0.12	0.178	0.058
16	0.254	0.235	-0.018

- Correlation coefficient (Pearson's)
- Mean absolute error
  - What is the range of possible values?
  - Are larger values better or worse?

Today’s Topics

- Regression Applications
- Discussion: Building Regression Datasets
- Linear Regression
- Polynomial Regression
- Regularization (Ridge Regression and Lasso Regression)
- Evaluating Regression Models
- Lab

Lab Assignment 1 Out

- “Unofficial TA” – Brandon Dang
- Piazza for class questions/discussion
- Office hours from Microsoft Azure representative this Friday Jan. 26:
    - Choose time slot here: http://slottd.com/events/izg73js2o8/slots

More Resources I Used for Today's Slides

- http://www.cs.utoronto.ca/~fidler/teaching/2015/slides/CSC411/
- - http://www.cs.cmu.edu/~epxing/Class/10701/lecture.html
- - http://web.cs.ucla.edu/~sriram/courses/cs188.winter-2017/html/index.html
- - https://people.eecs.berkeley.edu/~jrs/189/
- - http://alex.smola.org/teaching/cmu2013-10-701/
- - http://sli.ics.uci.edu/Classes/2015W-273a

Imagine Cup Competition

- Microsoft is hosting the annual Imagine Cup competition – both in the US and internationally.
  - The top prize is $100,000.
  - Learn more at https://imagine.microsoft.com/en-us/Category/11

- What to do? Build a project using Microsoft technologies