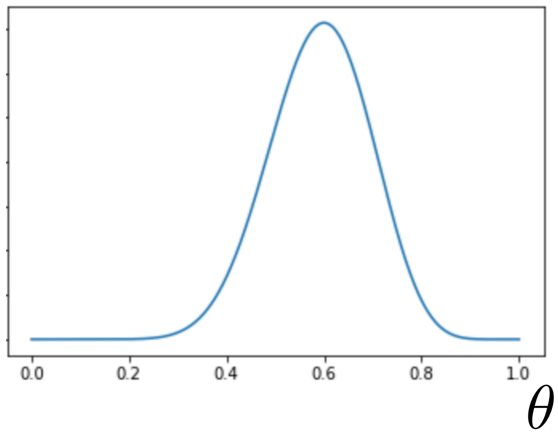
Lecture 2: MLE for Gaussian and linear regression

Recap: Maximum Likelihood Estimation

- **Observe** $\mathscr{D} = X_1, X_2, \ldots, X_n$ drawn i.i.d. from $P(X_i; \theta)$ for some ground truth $\theta = \theta^*$, unknown to us
- **Maximize log-likelihood** when we observe k heads in n flips

$\log P(\mathscr{D}; \theta) \;=$

$P(\mathscr{D}; \theta)$ when k heads obsereved in n flips

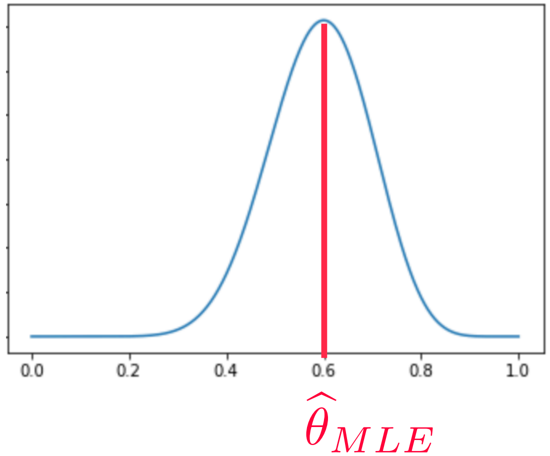


Recap: Maximum Likelihood Estimation

- **Observe** $\mathcal{D} = X_1, X_2, \ldots, X_n$ drawn i.i.d. from $P(X_i; \theta)$ for some ground truth $\theta = \theta^*$, unknown to us
- **Maximize log-likelihood** when we observe k heads in n flips

$\log P(\mathcal{D}; \theta) \;=$

$P(\mathcal{D}; \theta)$ when k heads obsereved in n flips



- Use the fact that derivative is zero at maxima (and also minima)
- Set derivative to zero,

  and find θ satisfying: $$\frac{d}{d\theta} \log P(\mathcal{D}; \theta) = 0$$

Maximum Likelihood Estimation

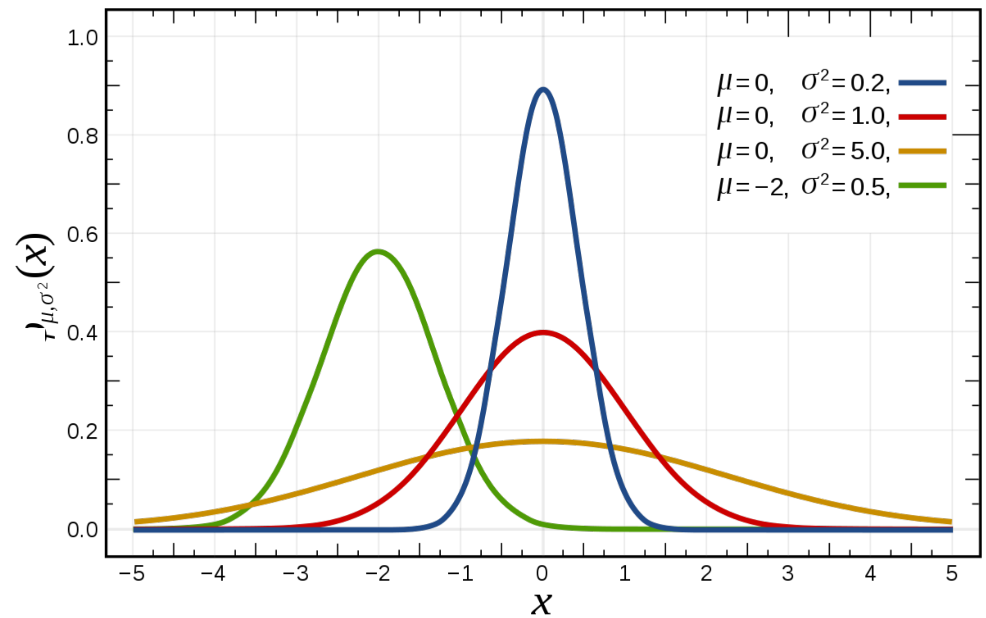
- **Observe** $X_1, X_2, \ldots, X_n$ drawn i.i.d. from $P(X_i; \theta)$ for some true $\theta = \theta^*$
- **Likelihood function**: $L_n(\theta) = \prod_{i=1}^{n} P(X_i; \theta)$
- **Log-likelihood function**: $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log P(X_i; \theta)$
- **Maximum Likelihood Estimator (MLE)**: $\widehat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \ell_n(\theta)$

- Warning when setting the derivative to zero to find the MLE:
  - The solution includes maxima, minima, and stationary points ⇒ needs to be checked
  - It does not always lead to an explicit expression in a closed form ⇒ alternative methods

What about continuous variables?

- *Client*: What if I am measuring a **continuous variable**?
- ***You*: Let me tell you about Gaussians…**
  - A Gaussian random variable is written as $X \sim \mathcal{N}(\mu, \sigma^2)$ with mean $\mu \triangleq \mathbb{E}[X]$ and variance $\sigma^2 \triangleq \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$
  - The p.d.f. (Probability Density Function) of X is

$$P(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



[EdDiscussion Question: What distributions do we need to memorize?]

Some useful properties of Gaussians

- Affine transformation
  (multiplying by scalar and adding a constant)
  - $X \sim \mathcal{N}(\mu, \sigma^2)$
  - $Y = aX + b \implies Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

- Sum of Gaussians
  - $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$
  - $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$
  - $Z = X + Y \implies Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

- [HW0 Questions A3 and A4]

MLE for Gaussian

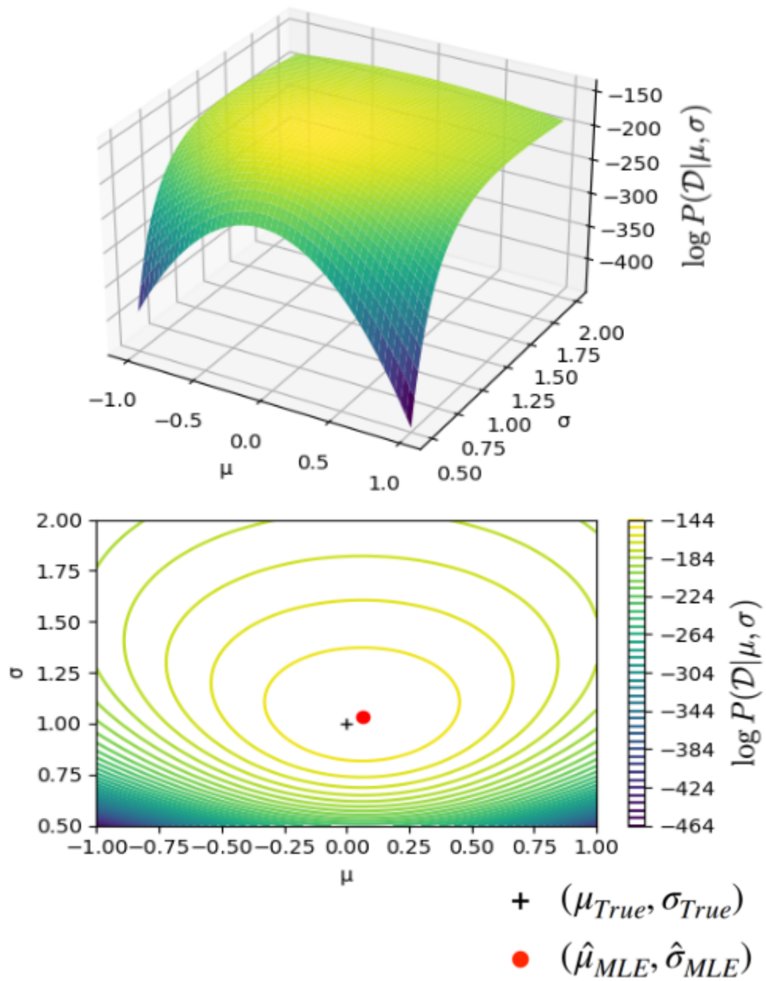
- **Hypothesis**: i.i.d. samples $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ from $\mathcal{N}(\mu, \sigma^2)$

$$P(\mathcal{D}; \mu, \sigma^2) = P(x_1, \ldots, x_n; \mu, \sigma^2)$$
$$= P(x_1; \mu, \sigma^2) \times P(x_2; \mu, \sigma^2) \times \cdots P(x_n; \mu, \sigma^2)$$
$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

- **Log-likelihood** of data:

$$\log P(\mathcal{D}; \mu, \sigma^2) =$$

- What is $\widehat{\theta}_{\mathrm{MLE}}$ for $\theta = (\mu, \sigma^2)$?



Your second learning algorithm: MLE for mean of a Gaussian distribution

- What's MLE for mean? Set partial derivative to zero:

$$\frac{d}{d\mu} \log P(\mathscr{D}; \mu, \sigma^2) = \frac{d}{d\mu}\left[-n \log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$

MLE for variance of a Gaussian distribution

- Again, set partial derivative to zero:

$$\frac{d}{d\sigma}\log P(\mathscr{D};\mu,\sigma^2) = \frac{d}{d\sigma}\left[-n\log(\sigma\sqrt{2\pi}) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

What can we say about the MLE?

- MLE for the mean of a Gaussian is **unbiased**
  - $\widehat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- MLE for the variance of a Gaussian is **biased**
  - $\widehat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \widehat{\mu}_{\mathrm{MLE}})^2$
  - $\mathbb{E}[\widehat{\sigma}^2_{\mathrm{MLE}}] \neq \sigma^2$

Maximum Likelihood Estimation

- **Observe** $X_1, X_2, \ldots, X_n$ drawn i.i.d. from $P(X_i; \theta)$ for some true $\theta = \theta^*$
- **Likelihood function**: $L_n(\theta) = \prod_{i=1}^{n} P(X_i; \theta)$
- **Log-likelihood function**: $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log P(X_i; \theta)$
- **Maximum Likelihood Estimator (MLE)**: $\widehat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \ell_n(\theta)$

- Properties (under benign regularity conditions—smoothness, identifiability, etc.):
  - MLE converges to the ground truths θ^* as the number of samples $n \to \infty$

Linear Regression

Maximum Likelihood Estimation

- **Observe** $X_1, X_2, \ldots, X_n$ drawn i.i.d. from $P(X_i; \theta)$ for some true $\theta = \theta^*$
- **Likelihood function**: $L_n(\theta) = \prod_{i=1}^{n} P(X_i; \theta)$
- **Log-likelihood function**: $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log P(X_i; \theta)$
- **Maximum Likelihood Estimator (MLE)**: $\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell_n(\theta)$

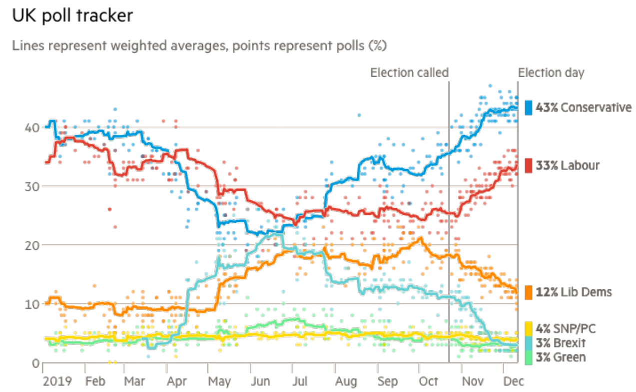
- Why do we care about recovering the "true" parameter θ^*?
  - **Estimation** of the parameter θ^* can be a goal.
  - Help **Interpret** or summarize large datasets.
  - Make **predictions** about future data.
  - **Generate** new data $X \sim f(\cdot\,; \hat{\theta}_{\text{MLE}})$

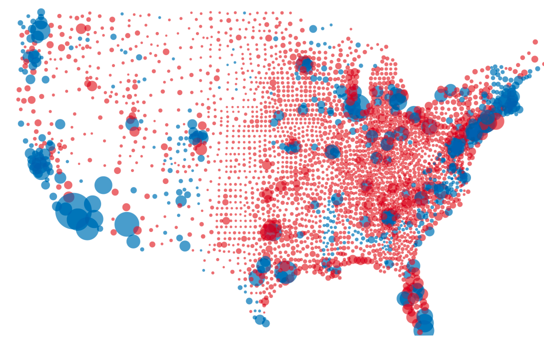
Estimation

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Opinion polls
How does the greater population feel about an issue? Correct for over-sampling?

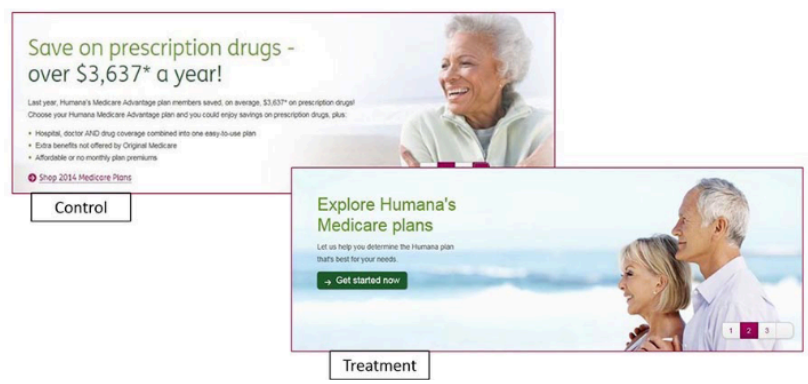
- θ_* is "true" average opinion
- $X_1, X_2, \ldots$ are sample calls





A/B testing
How do we figure out which ad results in more click-through?

- θ_* are the "true" average rates
- $X_1, X_2, \ldots$ are binary "clicks"

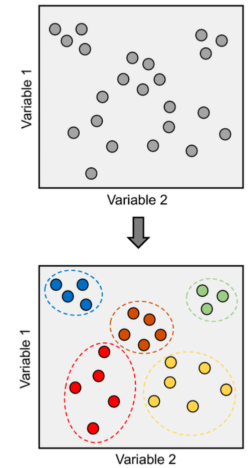


Interpret

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

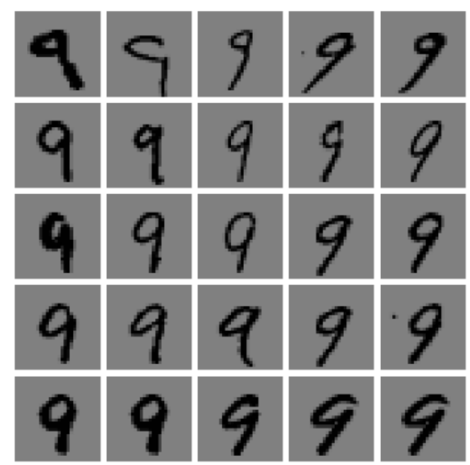
Customer segmentation / clustering
Can we identify distinct groups of customers by their behavior?

- θ_* describes "center" of distinct groups
- $X_1, X_2, \ldots$ are individual customers



Data exploration
What are the degrees of freedom of the dataset?

- θ_* describes the principle directions of variation
- $X_1, X_2, \ldots$ are the individual images

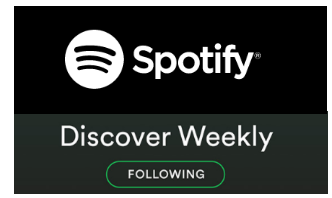


Predict

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Content recommendation
Can we predict how much someone will like a movie based on past ratings?

- θ_* describes user's preferences
- $X_1, X_2, \ldots$ are (movie, rating) pairs



Object recognition / classification
Identify a flower given just its picture?

- θ_* describes the characteristics of each kind of flower
- $X_1, X_2, \ldots$ are the (image, label) pairs

(a) *(b)* *(c)*

Figure 1.1: Three types of Iris flowers: Setosa, Versicolor and Virginica. Used with kind permission of Dennis Kramb and SIGNA.

index	sl	sw	pl	pw	label
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
	...				
50	7.0	3.2	4.7	1.4	Versicolor
	...				
149	5.9	3.0	5.1	1.8	Virginica

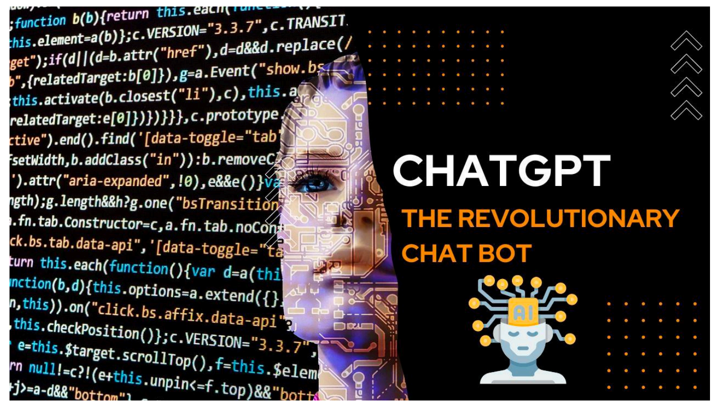
Generate

Observe $X_1, X_2, \ldots, X_n$ drawn IID from $f(x;\theta)$ for some "true" $\theta = \theta_*$

Text generation
Can AI generate text that could have been written like a human?

- θ_* describes language structure
- $X_1, X_2, \ldots$ are text snippets found online

"Kaia the dog wasn't a natural pick to go to mars. No one could have predicted she would..."



https://chat.openai.com/chat

Image to text generation
Can AI generate an image from a prompt?

- θ_* describes the coupled structure of images and text
- $X_1, X_2, \ldots$ are the (image, caption) pairs found online

"dog talking on cell phone under water, oil painting"

https://labs.openai.com/

CSE 446/546

- week 1: **Estimation**
    - Maximum Likelihood Estimation
- week 1~8: **Prediction**
    - week 1~4: Linear regression models
    - week 4~5: Linear classification models (also called Logistic regression)
    - Mideterm
    - week 6~7: Non-linear models
- week 8~9: **Interpretation**
- week 10: **Generation**…(?)

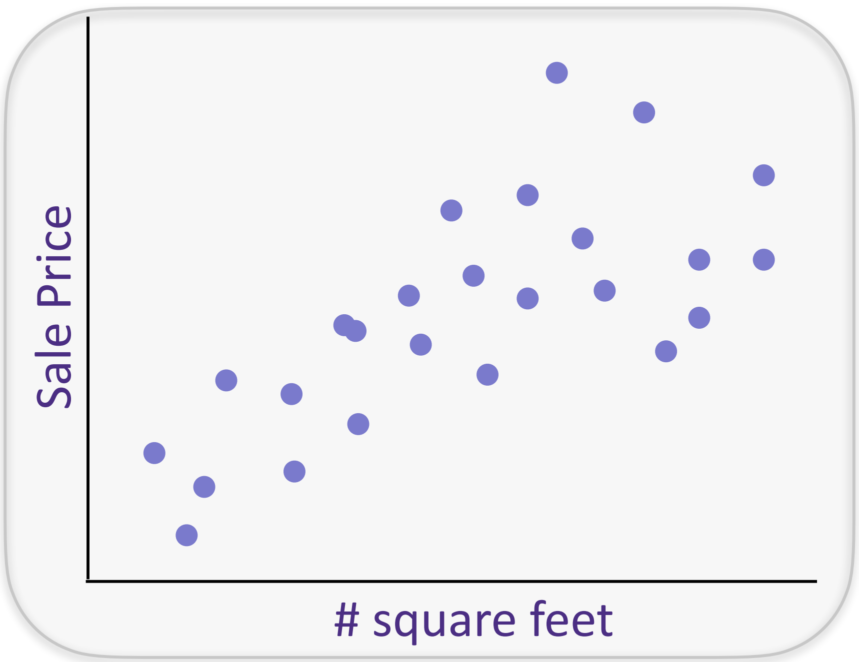
Linear regression model, 1-dimensional

You want to sell your house that is 2,500 sq.ft.

Q. What is the right price?

Collect past sales data on zillow.com:

***y* = House sale price *and x* = {# sq. ft.}**



Training Data: $x_i \in \mathbb{R}$ $y_i \in \mathbb{R}$

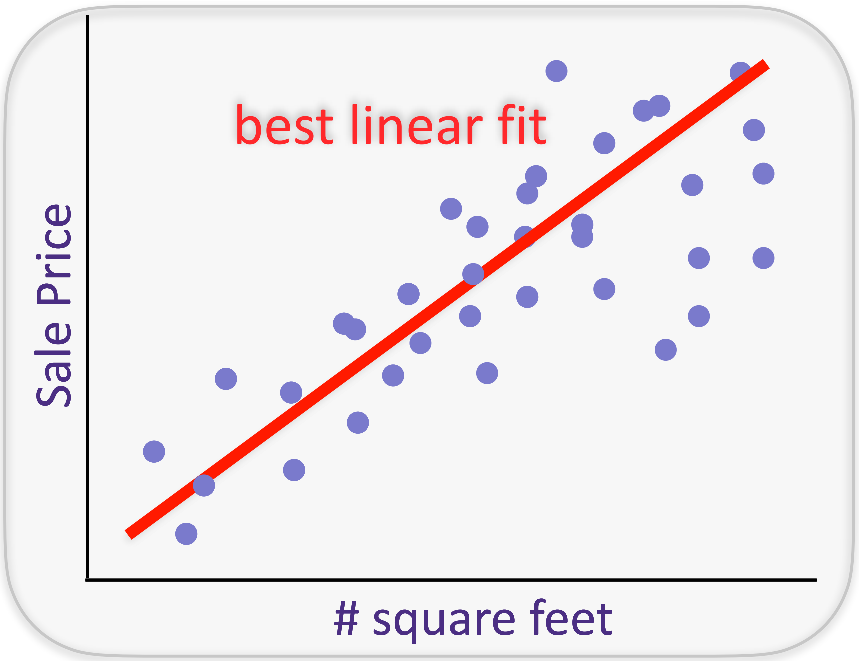
$$\{(x_i, y_i)\}_{i=1}^n$$

Linear regression model, 1-dimension

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft.}**



1. Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$

2. Hypothesis/Model: linear

$$y_i = w \cdot x_i + \epsilon_i$$

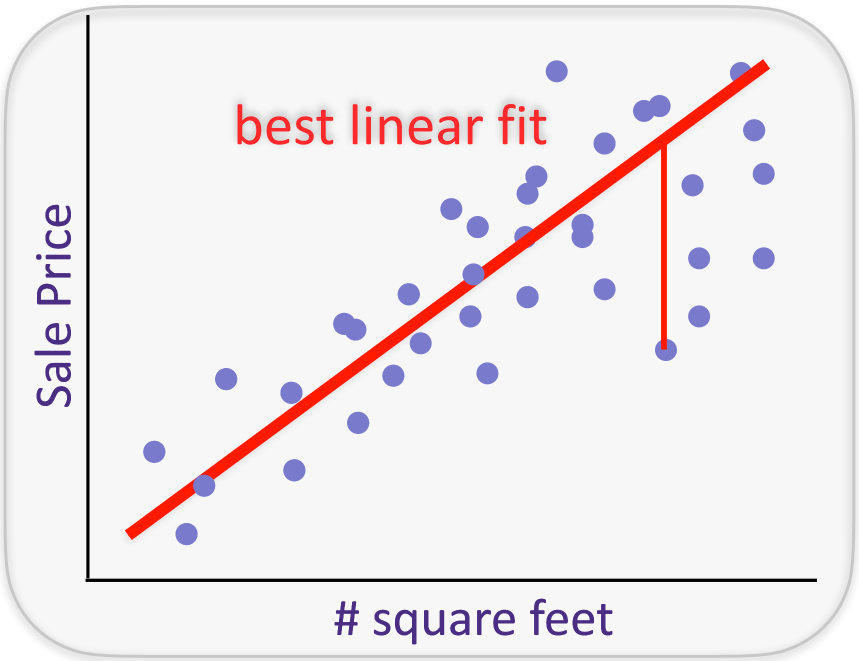
For now we assume there is no y-intercept in the model, and will handle it later

Linear regression model, 1-dimension

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft.}**



1. Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}$, $y_i \in \mathbb{R}$

2. Hypothesis/Model: linear

$$y_i = w \cdot x_i + \epsilon_i$$

3. Noise: i.i.d. Gaussian with

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

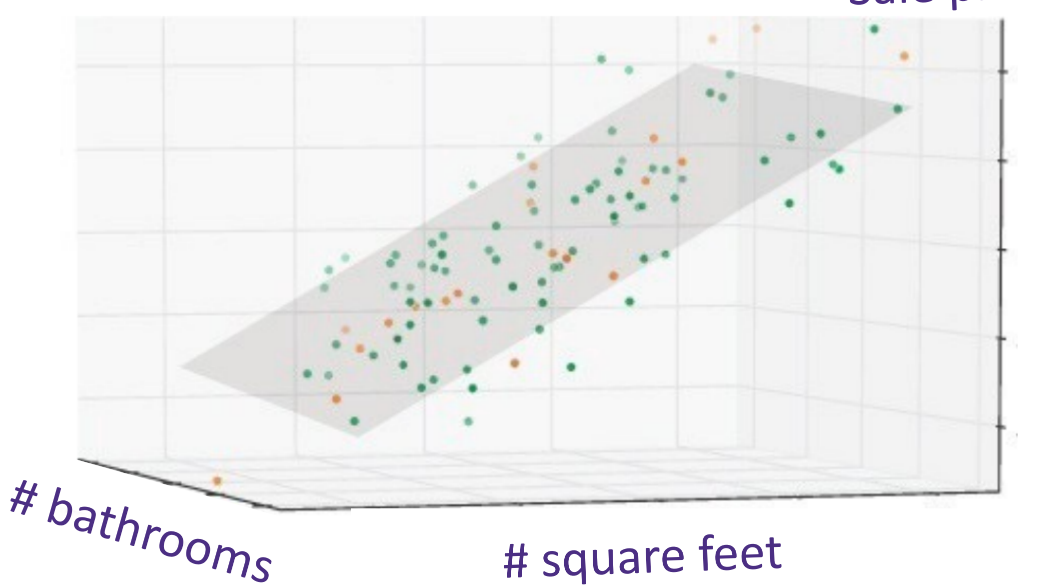
For now we assume there is no y-intercept in the model, and will handle it later

Linear regression model, d-dim

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft., zip code, date of sale, etc.}**



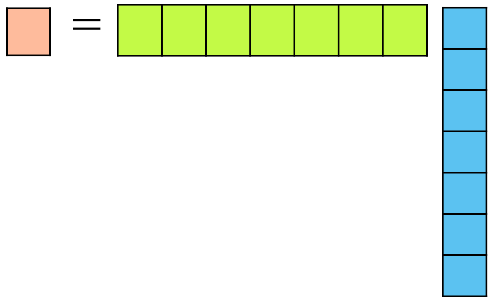
Training Data: $\{(x_i, y_i)\}_{i=1}^n$

$x_i \in \mathbb{R}^d$

$y_i \in \mathbb{R}$

Hypothesis/Model: linear

$$y_i = x_i^T w + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

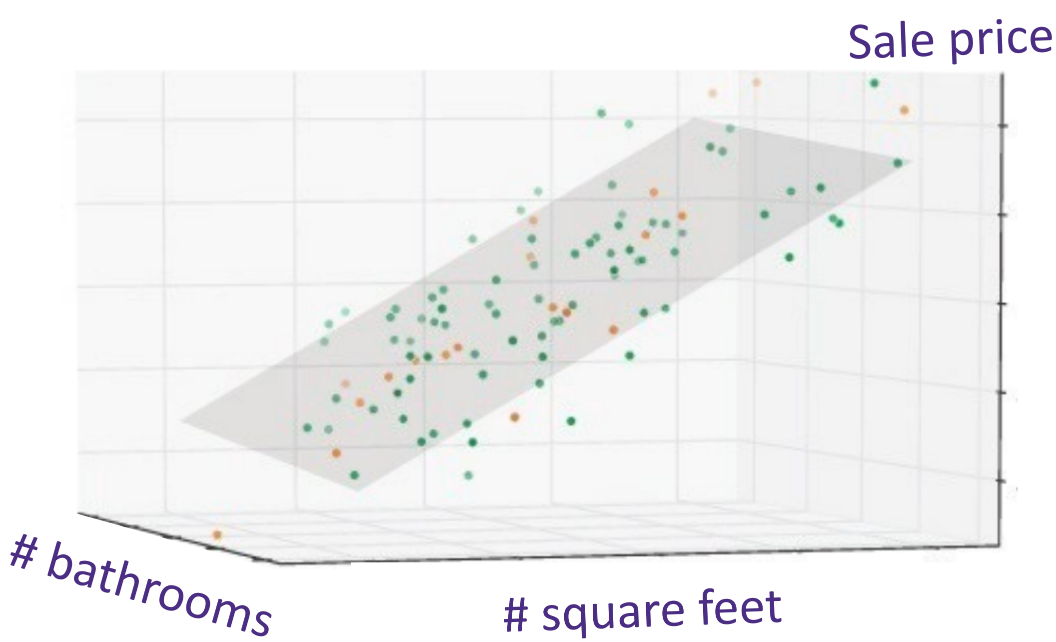


Linear regression model, d-dim

Given past sales data on zillow.com, predict:

y = **House sale price** *from*

x = **{# sq. ft., zip code, date of sale, etc.}**



Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Hypothesis/Model: linear

$$y_i = x_i^T w + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

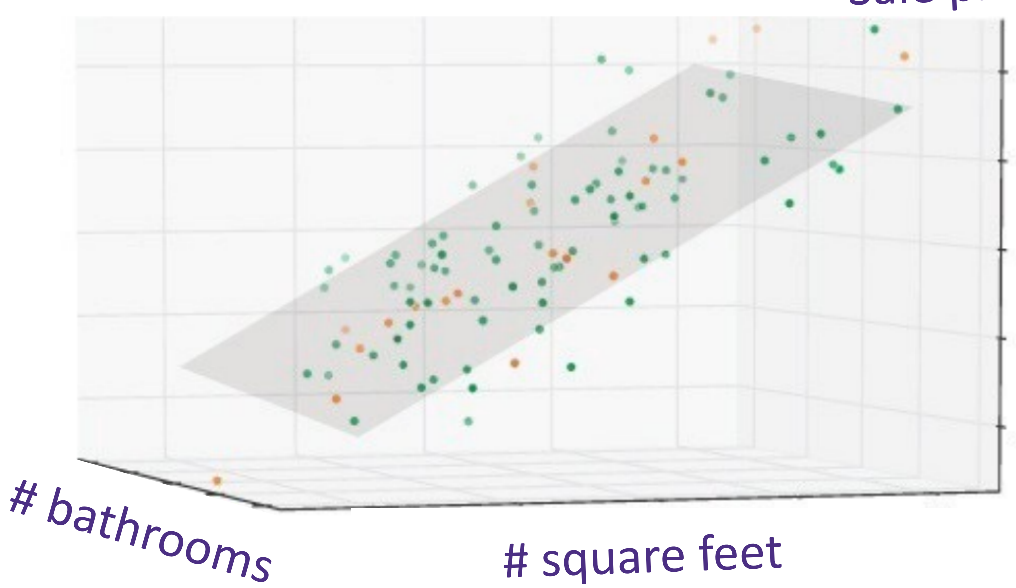
$$p(y|x, w, \sigma) =$$

Linear regression model, d-dim

Given past sales data on zillow.com, predict:

y = House sale price *from*

x = {# sq. ft., zip code, date of sale, etc.}



Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Hypothesis/Model: linear

$$y_i = x_i^T w + \epsilon_i \qquad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y - x^\top w)^2 / 2\sigma^2}$$

Maximizing log-likelihood

Training Data:
$\{(x_i, y_i)\}_{i=1}^n$

$x_i \in \mathbb{R}^d$
$y_i \in \mathbb{R}$

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y - x^\top w)^2 / 2\sigma^2}$$

Likelihood: $$P(\mathcal{D}|w, \sigma) = \prod_{i=1}^n p(y_i|x_i, w, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - x_i^\top w)^2 / 2\sigma^2}$$

Maximum Likelihood Estimation

- **Observe** $X_1, X_2, \ldots, X_n$ drawn i.i.d. from $P(X_i; \theta)$ for some true $\theta = \theta^*$
- **Likelihood function**: $L_n(\theta) = \prod_{i=1}^{n} P(X_i; \theta)$
- **Log-likelihood function**: $\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log P(X_i; \theta)$
- **Maximum Likelihood Estimator (MLE)**: $\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell_n(\theta)$

- Properties (under benign regularity conditions—smoothness, identifiability, etc.):
  - MLE converges to the ground truths θ^* as the number of samples $n \to \infty$

Maximizing log-likelihood

Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$$p(y|x, w, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y - x^\top w)^2 / 2\sigma^2}$$

Likelihood: $$P(\mathcal{D}|w, \sigma) = \prod_{i=1}^n p(y_i|x_i, w, \sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - x_i^\top w)^2 / 2\sigma^2}$$

Maximize (wrt w): $$\log P(\mathcal{D}|w, \sigma) = \log \left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - x_i^\top w)^2 / 2\sigma^2} \right)$$

Maximizing log-likelihood

Training Data: $x_i \in \mathbb{R}^d$ $y_i \in \mathbb{R}$

$\{(x_i, y_i)\}_{i=1}^n$

$$p(y|x,w,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-x^\top w)^2/2\sigma^2}$$

Likelihood: $$P(\mathcal{D}|w,\sigma) = \prod_{i=1}^n p(y_i|x_i,w,\sigma) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2}$$

Maximize (wrt w): $$\log P(\mathcal{D}|w,\sigma) = \log\left(\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i-x_i^\top w)^2/2\sigma^2}\right)$$

$$\widehat{w}_{MLE} = \arg\min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

Maximizing log-likelihood

$$\widehat{w}_{MLE} = \arg\min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

Set gradient=0, solve for w

Maximizing log-likelihood

$$\widehat{w}_{MLE} = \arg\min_w \sum_{i=1}^n (y_i - x_i^\top w)^2$$

Set gradient=0, solve for w

$$\widehat{w}_{MLE} = \left(\sum_{i=1}^n x_i x_i^\top \right)^{-1} \sum_{i=1}^n x_i y_i$$

The regression problem in matrix notation

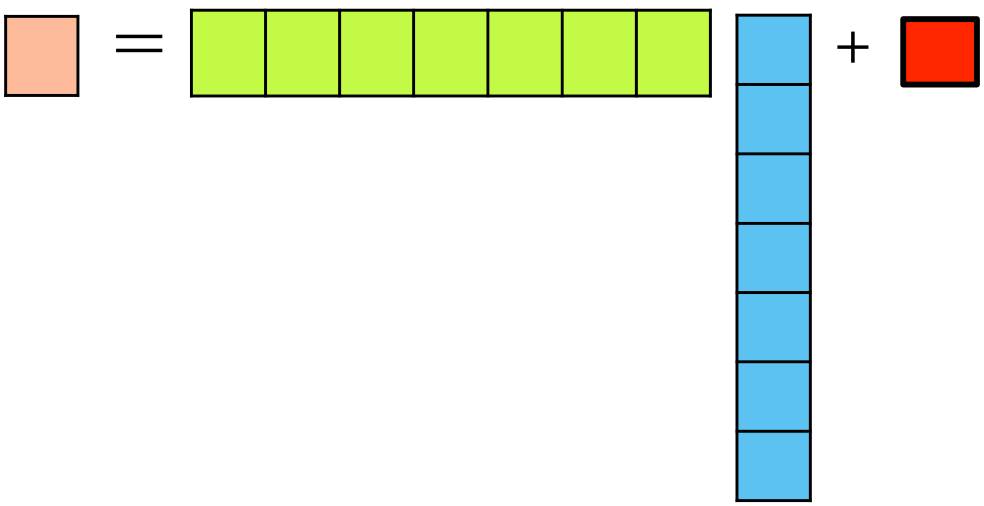
Data:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

d : # of features/size of the input

n : # of examples/datapoints

$$y_1 = x_1^T w + \epsilon_1$$



$$y_1 = x_1^T w + \epsilon_1$$

$$y_2 = x_2^T w + \epsilon_2$$

The regression problem in matrix notation

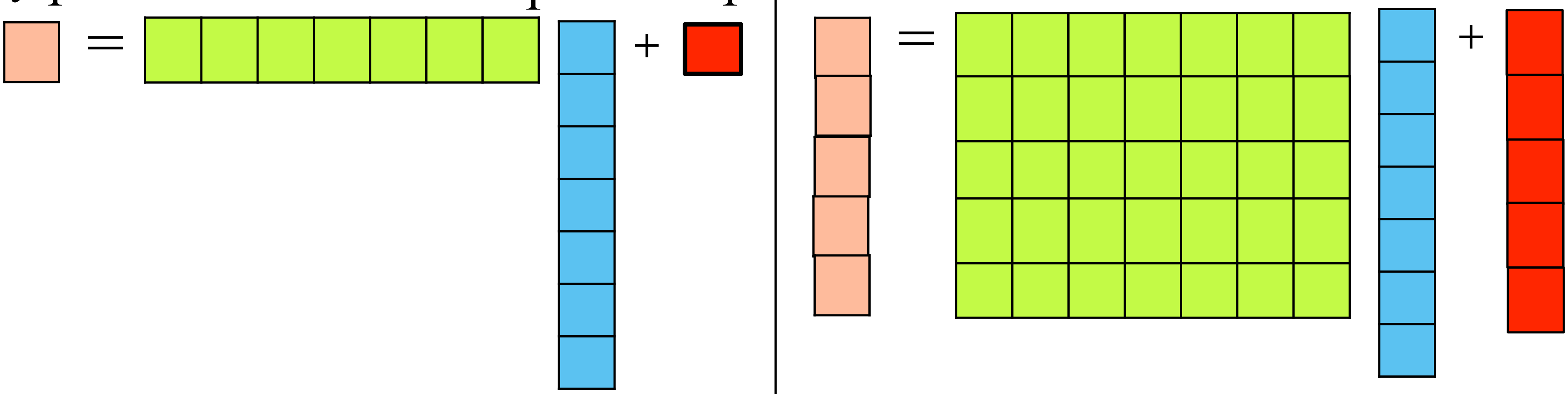
Data:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

d : # of features/size of the input

n : # of examples/datapoints

$$y_1 = x_1^T w + \epsilon_1 \qquad y = Xw + \epsilon$$



The regression problem in matrix notation

$$\widehat{w}_{MLE} = \arg\min_w \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

d : # of features

n : # of examples/datapoints

ℓ_2 norm: $\|z\|_2 = \sqrt{\sum_{i=1}^{n} z_i^2} = \sqrt{z^\top z}$

$$\widehat{w}_{LS} = \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2$$
$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w)$$

[related to HW0 questions A6 and A7]

The regression problem in matrix notation

$$= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w)$$

The regression problem in matrix notation

$$\widehat{w}_{MLE} = \arg\min_w \sum_{i=1}^{n} (y_i - x_i^\top w)^2$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

d : # of features

n : # of examples/datapoints

ℓ_2 norm: $\|z\|_2 = \sqrt{\sum_{i=1}^{n} z_i^2} = \sqrt{z^\top z}$

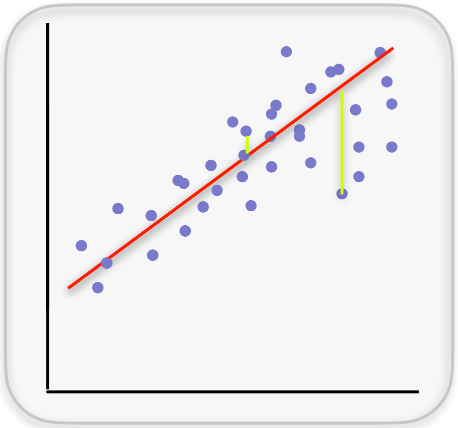
$$\begin{aligned} \widehat{w}_{LS} &= \arg\min_w \|\mathbf{y} - \mathbf{X}w\|_2^2 \\ &= \arg\min_w (\mathbf{y} - \mathbf{X}w)^T(\mathbf{y} - \mathbf{X}w) \end{aligned}$$

$$\hat{w}_{LS} = \hat{w}_{MLE} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

The regression problem in matrix notation

Recall that we start with a linear model with no offset

$$y_i = x_i^T w + \epsilon_1$$



$$\begin{aligned}\widehat{w}_{LS} &= \arg\min_w ||\mathbf{y} - \mathbf{X}w||_2^2 \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

We can add the offset to the linear model, with a new parameter b

$$y_i = x_i^T w + b + \epsilon_1$$

$$\begin{aligned}\widehat{w}_{LS}, \widehat{b}_{LS} &= \arg\min_{w,b} \sum_{i=1}^n \left(y_i - (x_i^T w + b)\right)^2 \\ &= \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2\end{aligned}$$

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$

$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

Dealing with an offset

$$\widehat{w}_{LS}, \widehat{b}_{LS} = \arg\min_{w,b} ||\mathbf{y} - (\mathbf{X}w + \mathbf{1}b)||_2^2$$

$$\mathbf{X}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{X}^T\mathbf{1} = \mathbf{X}^T\mathbf{y}$$
$$\mathbf{1}^T\mathbf{X}\widehat{w}_{LS} + \widehat{b}_{LS}\mathbf{1}^T\mathbf{1} = \mathbf{1}^T\mathbf{y}$$

If $X^T\mathbf{1} = 0$, i.e., if each feature is mean-zero or we pre-processed the data have zero-mean, then

$$\widehat{w}_{LS} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\widehat{b}_{LS} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Make Predictions

$$\widehat{w}_{LS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

$$\widehat{b}_{LS} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

A new house is about to be listed. What should it sell for?

$$\hat{y}_{\text{new}} = x_{\text{new}}^T \hat{w}_{LS} + \hat{b}_{LS}$$

Questions?