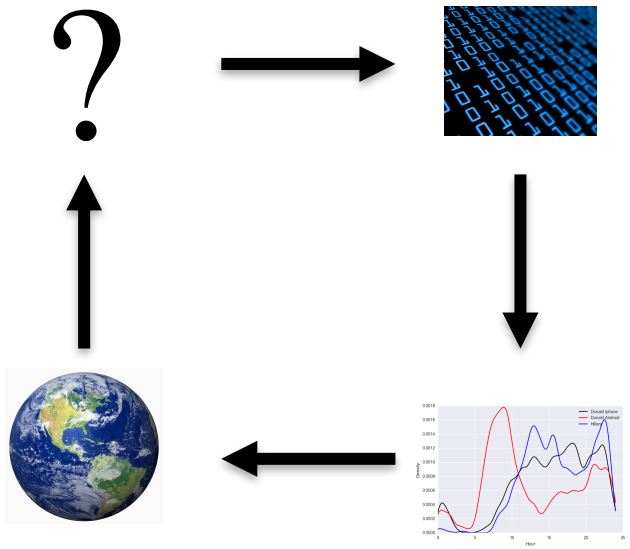
Data Science 100
Lec 17: Least Squares (LS)
as a Linear Prediction or Linear Fitting Method

DS100
Spring
2017

Slides by:
Bin Yu
binyu@stat.berkeley.edu
Joey Gonzalez
jegonzal@berkeley.edu
Thanks to Andrew Do for assistance on data analysis



Taxonomy
of Machine Learning/Statistics

Labeled Data

Indirect (reward)

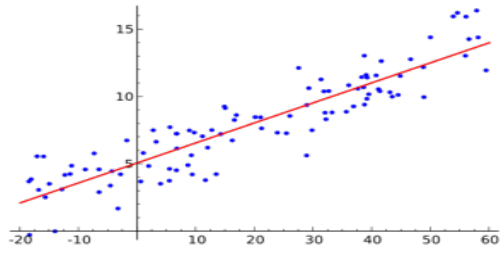
Unlabeled Data

Supervised Learning

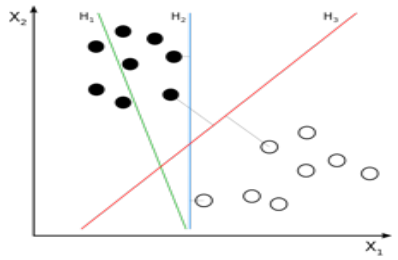
Reinforcement & Bandit Learning

Unsupervised Learning

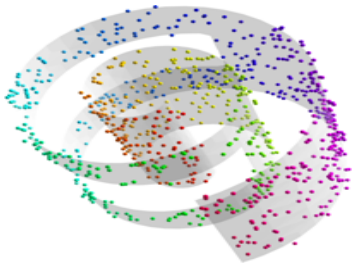
Regression



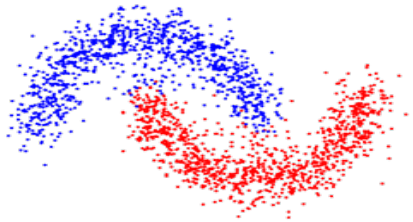
Classification



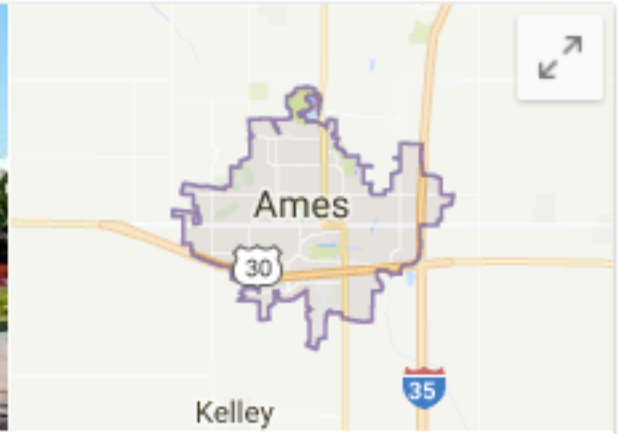
Dimensionality Reduction



Clustering



Q: how much does a house in Ames Iowa sell?



Ames

City in Iowa

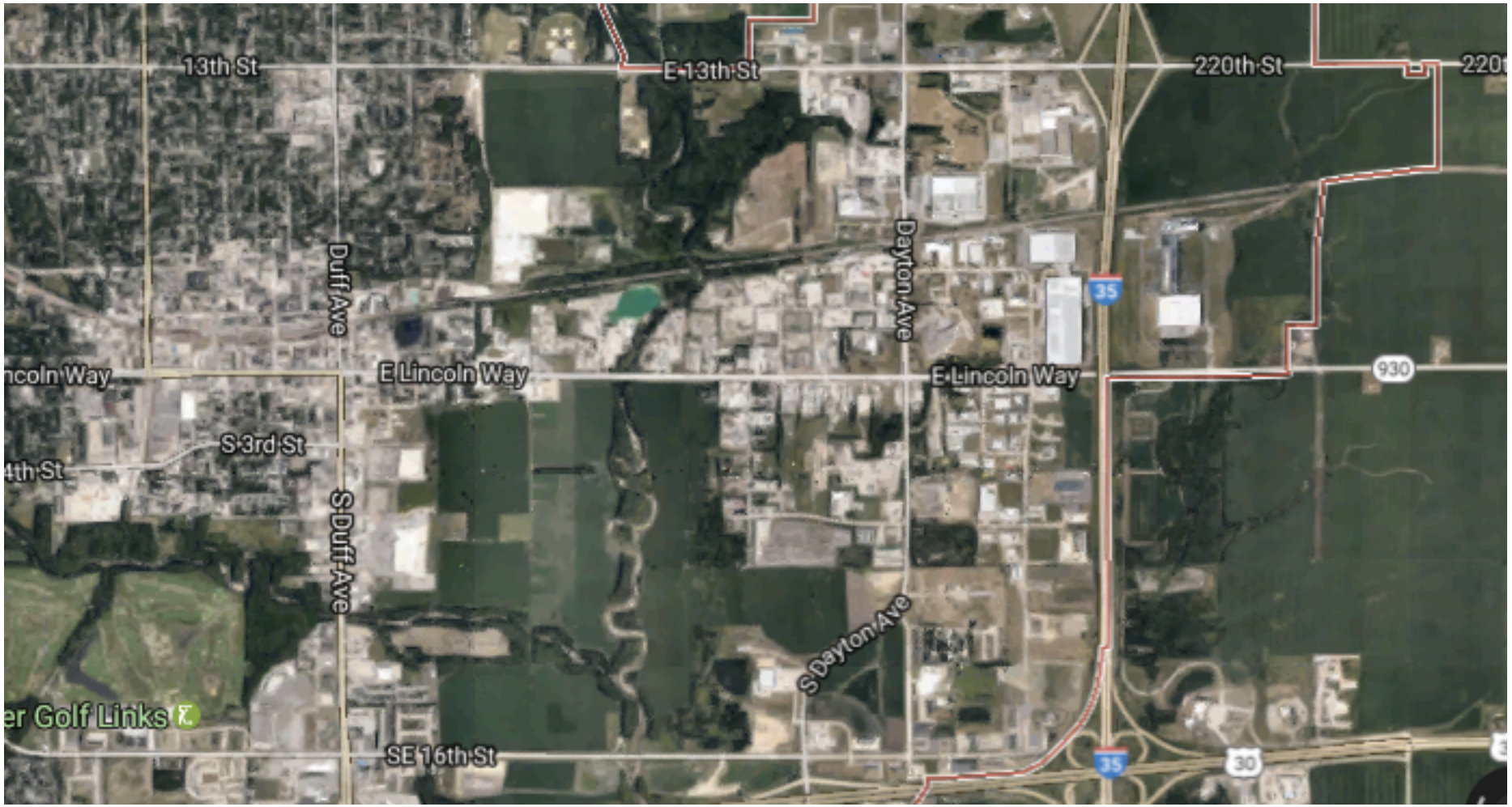
Ames is a city located in the central part of the U.S. state of Iowa in Story County. Lying approximately 30 miles north of Des Moines, it had a 2010 population of 58,965. Wikipedia

Weather: 24°F (-4°C), Wind N at 9 mph (14 km/h), 48% Humidity

Population: 61,792 (2013)

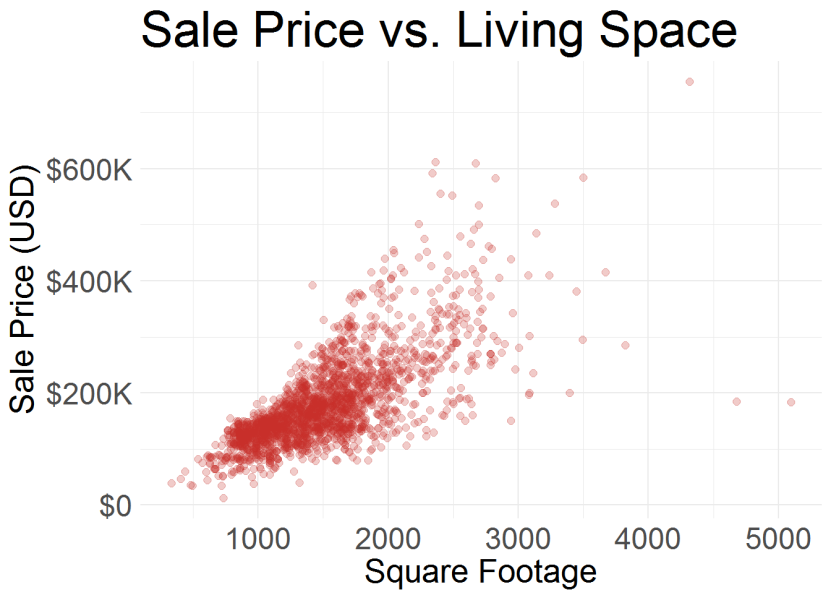
Local time: Saturday 10:54 PM

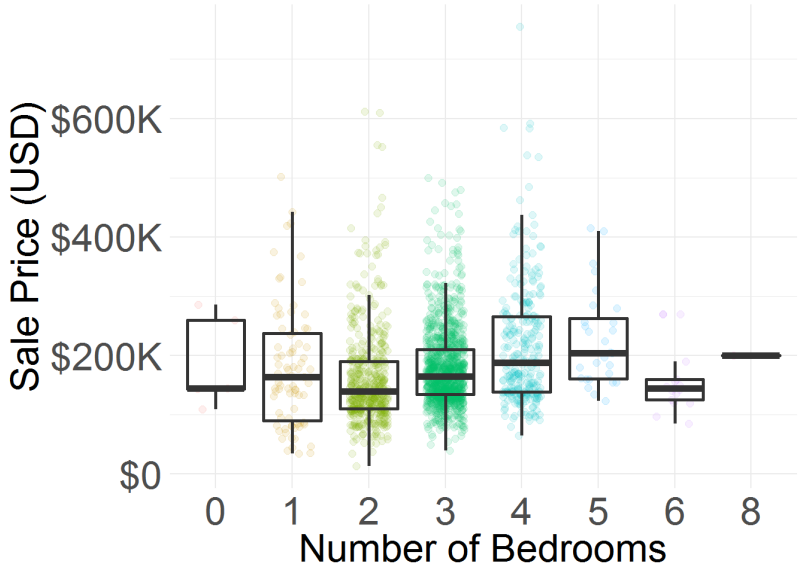
P: all houses in Ames, Iowa

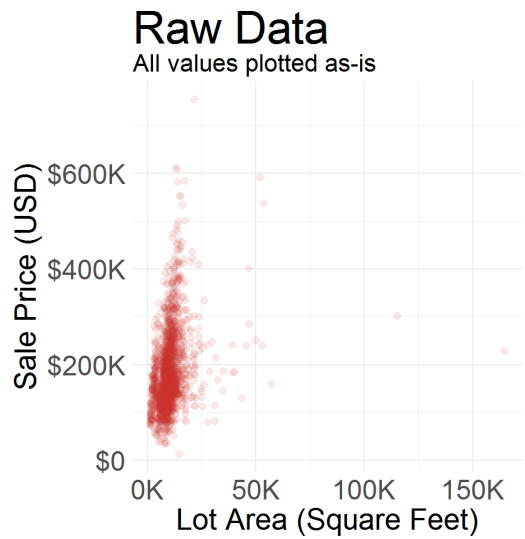


Visualizing data

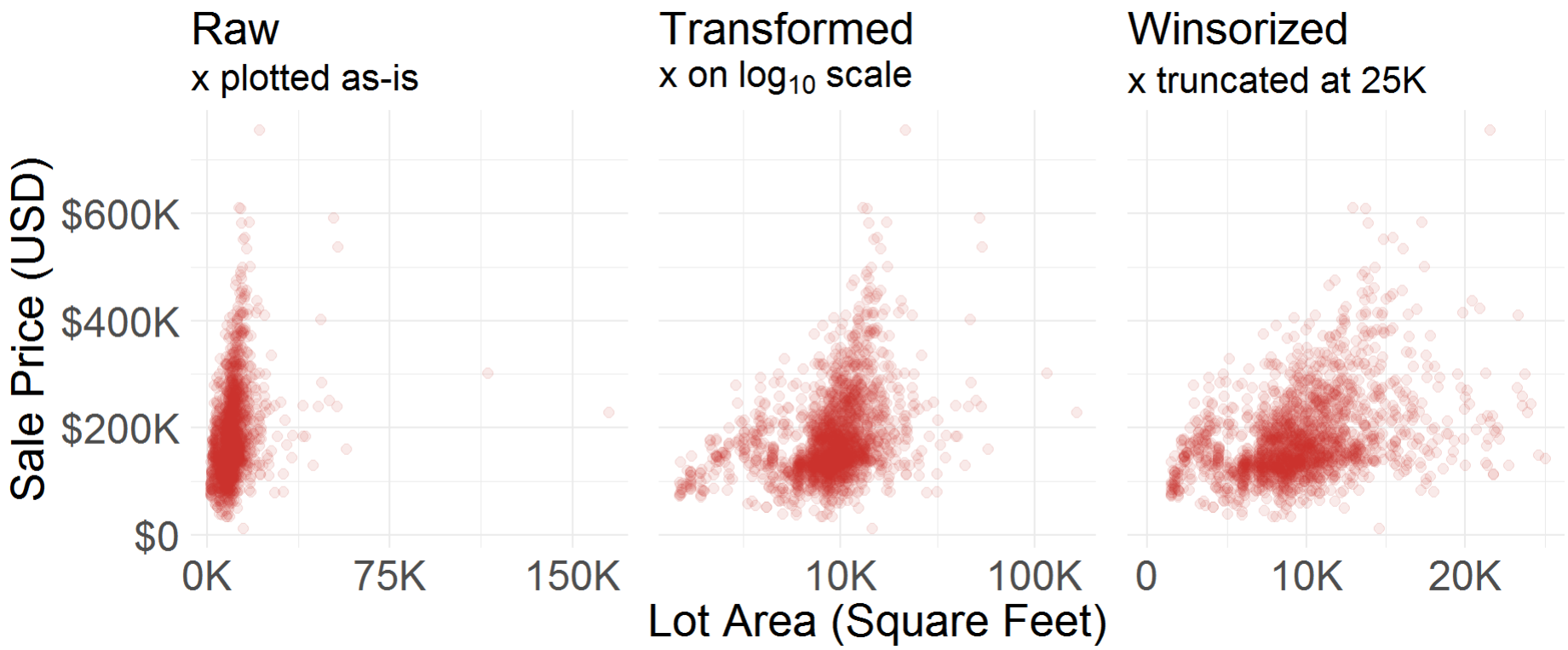
- 3 pairwise scatter plots of Y=house price against x: sq. ft.; number of bedrooms; lot size, which are what I think the most important predictors







Three ways to look at y=house price and x=lot size



Least Squares, a powerful linear prediction method

210 years ago…

Ceres story:

Piazzi, Gauss, Kepler, …

In 1801, Giuseppe Piazzi observed Ceres…

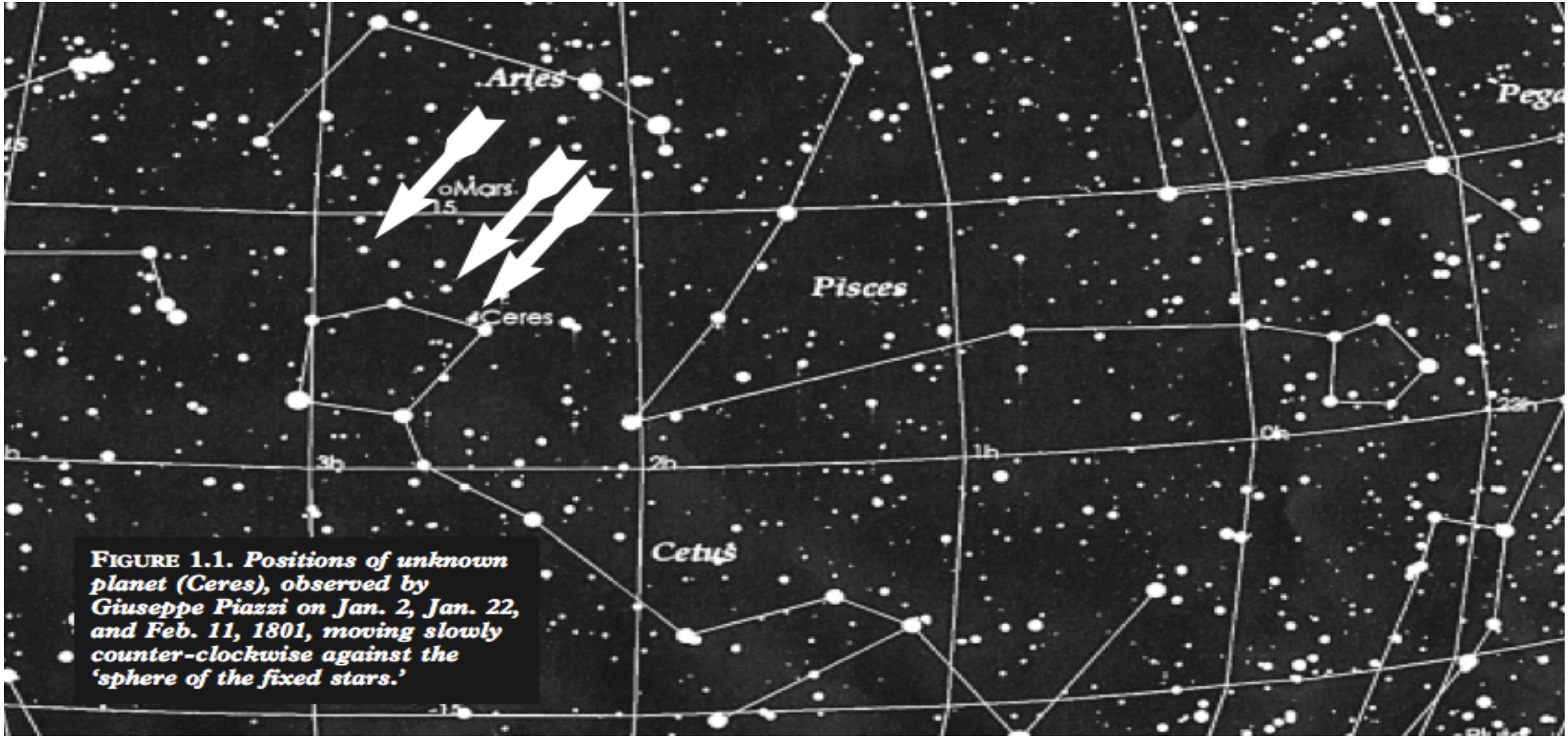


FIGURE 1.1. *Positions of unknown planet (Ceres), observed by Giuseppe Piazzi on Jan. 2, Jan. 22, and Feb. 11, 1801, moving slowly counter-clockwise against the 'sphere of the fixed stars.'*

Tennenbaum and Director, 1997

Giuseppe Piazzi

(July 16, 1746 – July 22, 1826)

Piazzi was an Italian Catholic priest of the Theatine order, mathematician, and astronomer. He supervised the compilation of the Palermo Catalogue of stars, containing 7,646 star entries with unprecedented precision. Piazzi discovered Ceres, today known as the largest member of the asteroid belt. -- Wikepedia

Carl Friedrich Gauss

(April 30, 1777 – February 23, 1855)

Gauss was a German mathematician and physical scientist who contributed significantly to many fields, including number theory, algebra, statistics, analysis, differential geometry, geodesy, geophysics, electrostatics, astronomy and optics. -- Wikipedia

Johannes Kepler

(December 27, 1571 – November 15, 1630)

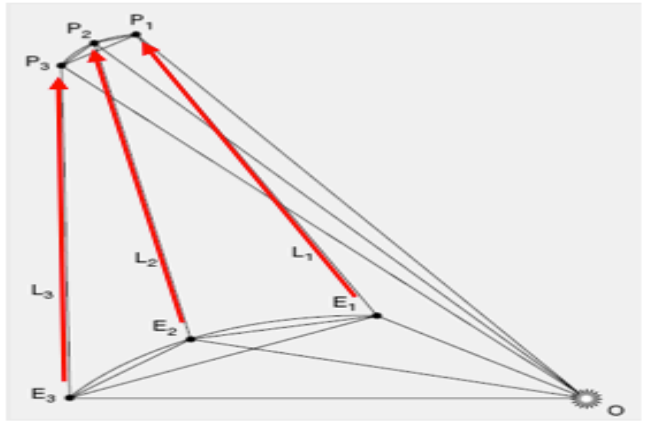
Kepler was a German mathematician, astronomer and astrologer.

A key figure in the 17th century scientific revolution, he is best known for his laws of planetary motion... -- Wikipedia

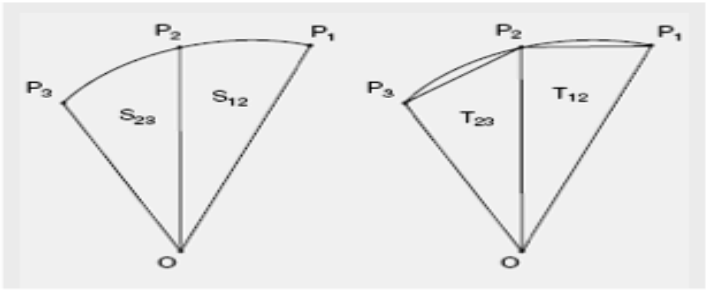
Gauss predicted Ceres' trajectory or future locations accurately

He used Kepler's second law of planetary motion, geometric relationships, approximations, and corrections,…

Piazzi's data: lines of sight L_1, L_2, L_3 and elapsed times between observations



Sectoral areas swept out by orbit are proportional to elasped times



$$\frac{S_{12}}{S_{23}} = \frac{t_2 - t_1}{t_3 - t_2} = 0.94952\,,$$

$$\frac{S_{12}}{S_{13}} = \frac{t_2 - t_1}{t_3 - t_1} = 0.48705\,,$$

$$\frac{S_{23}}{S_{13}} = \frac{t_3 - t_2}{t_3 - t_1} = 0.51295\,.$$

Approximate sectoral areas with triangular areas

$$\frac{T_{23}}{T_{13}} = \text{(approximately)}\ \frac{S_{23}}{S_{13}} = 0.513\,, \qquad \text{="c"}$$

$$\frac{T_{12}}{T_{23}} = \text{(approximately)}\ \frac{S_{12}}{S_{23}} = 0.487\,. \qquad \text{="d"}$$

Tennenbaum and Director, 1997; Le Corvec, Donatelli and Hunt, 2007 Math 221 project, UCB

Piazzi-Gauss-Kepler:
early data science research

Piazzi: data collection (3 sightings of Ceres and elapse times)

Kepler: domain knowledge (second law of planetary motion)

Gauss: geometry and many new math/stat methods to solve a real problem (predicting trajectory of Ceres)

Kepler, Piazzi, and Gauss were all mathematicians and astronomers since astronomy was very much part of mathematics then.

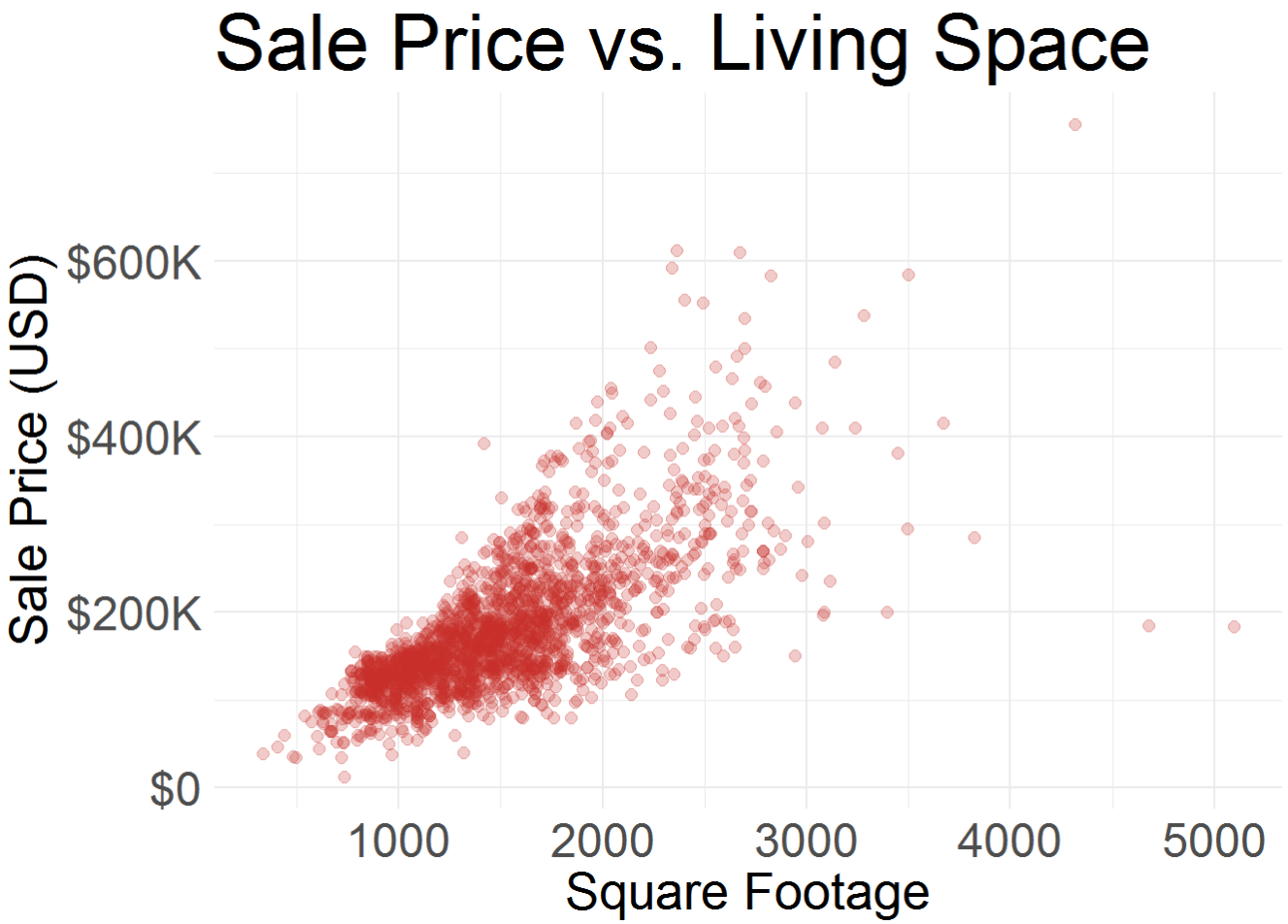
Least Squares (LS) Method

Gauss' method to find the trajectory of Ceres can be viewed as a ingenious iterative and approximate method to solve a Least Squares (LS) problem, based on "domain knowledge".

Gauss used squared loss function to measure how close the prediction is to the observed position of Ceres

For the sake of visualization, let us look at the LS function in 2-dim and in modern notations

Linear prediction with one predictor or feature



Least Squares (LS) Method

Given n training data units $(x_{i1}, x_{i2}, y_i)_{i=1}^{n}$, the Least Squares function $L(\theta)$ in 2-dim with $\theta \in R^2$

For the ith house in the data set, two predictors x, one response variable y:

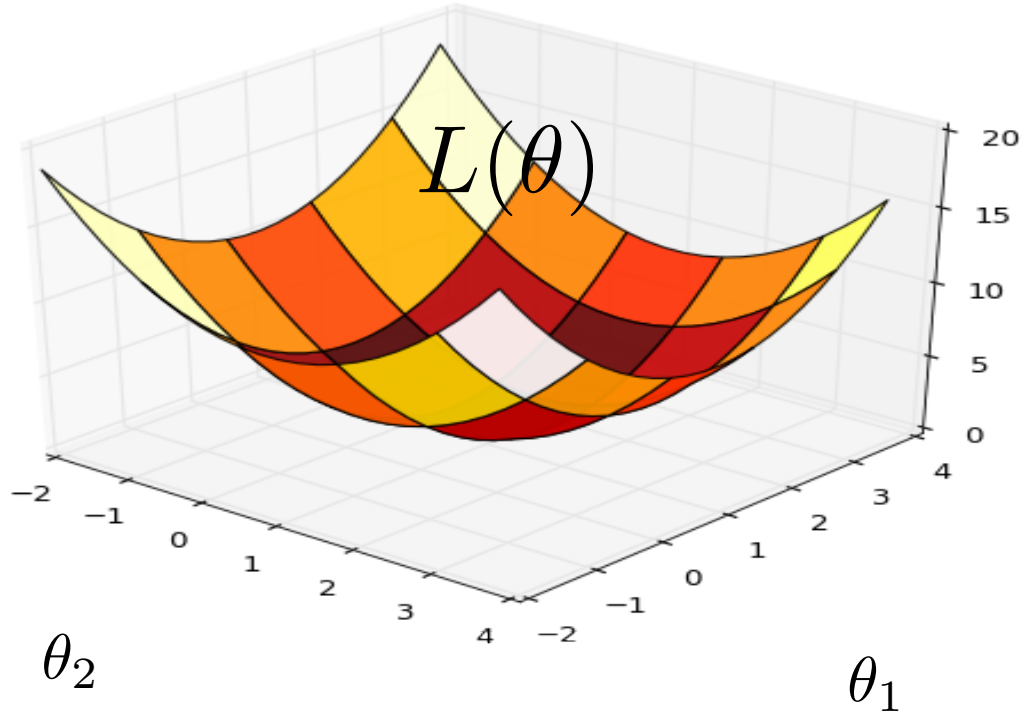
x_{i1} = 1; x_{i2} = sq. ft. y_i = house price

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_1 x_{i1} - \theta_2 x_{i2})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$

$$\theta^T x_i = \theta_1 x_{i1} + \theta_2 x_{i2}$$ Inner combination of two 2-dim vectors

$\theta_1, \ \theta_2$ linear "**weights**" or linear coefficients

LS loss function with when p=2



Least Squares (LS) is Powerful

- One of the most widely used techniques
- Fundamental to many larger models
- Easy to interpret
  - e.g., the weights or coefficients tell us something about the features/predictors
    - Positive or negative relationships …
- Efficient to solve
  - Fast numerical methods
  - Closed form solutions

LS: Finding the Best Parameters

Prediction rule: $$f_\theta(x) := \theta^T x$$

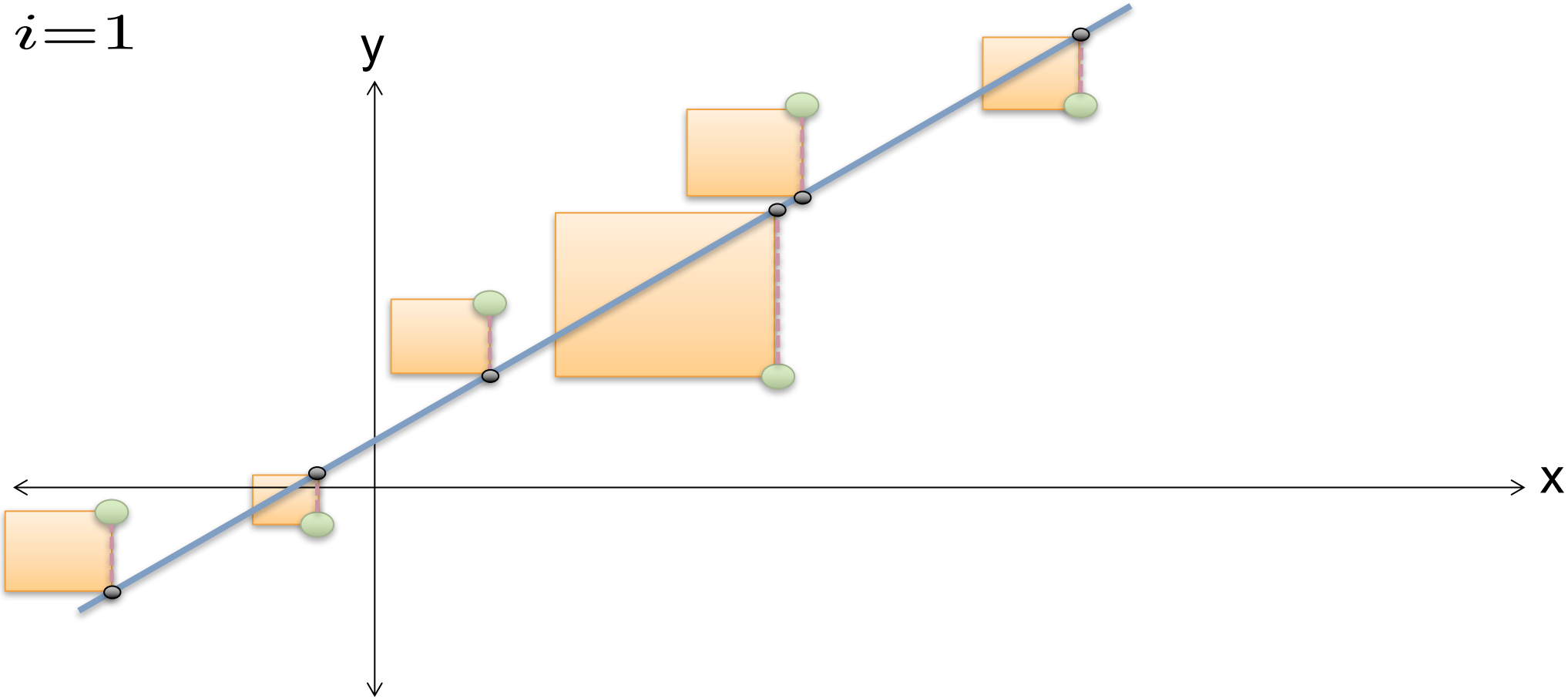
Step 1: define a **Loss Function:** *Average Prediction Error given* θ

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_\theta(x_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \theta^T x_i\right)^2$$

That is, Average Squared Differences between **observed (y_i)** and **predicted $f_\theta(x_i)$** values

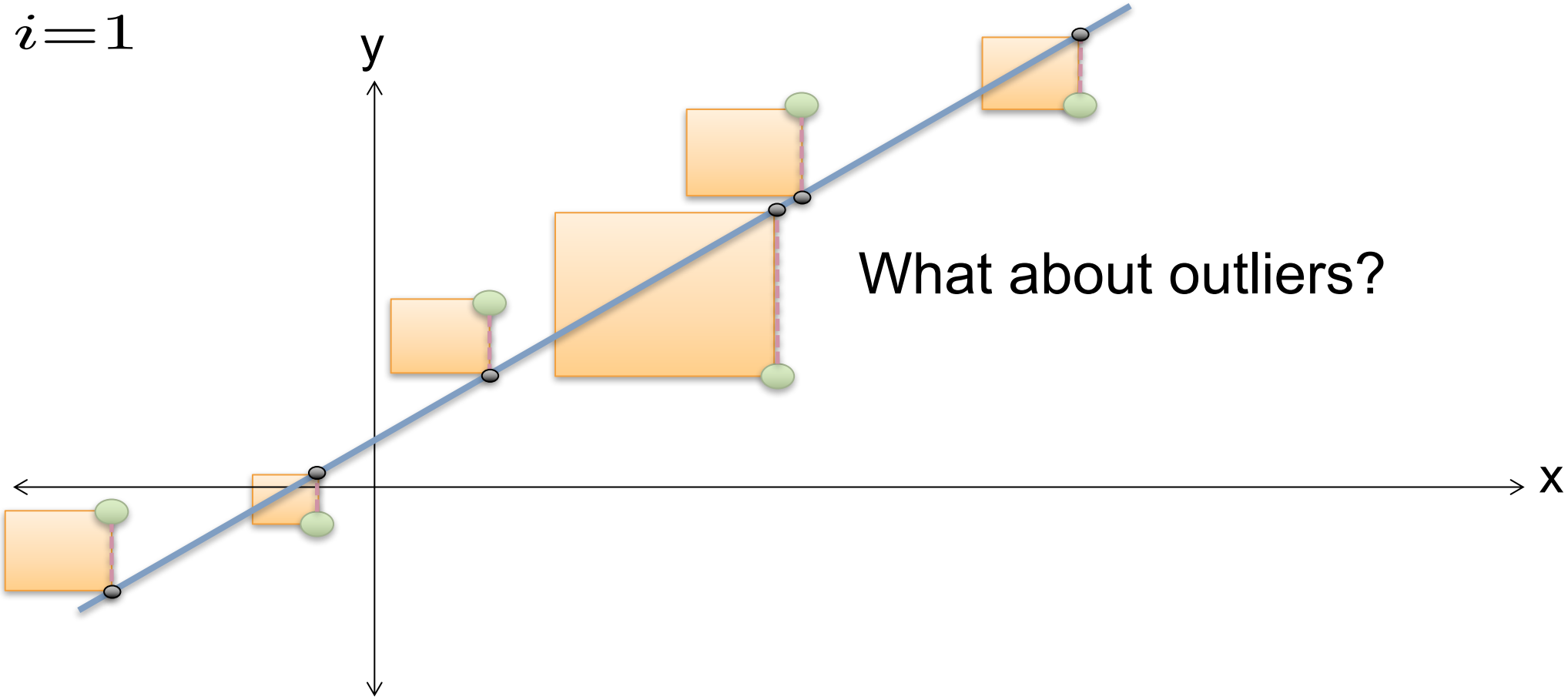
The meaning of Squared Loss (Error)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$



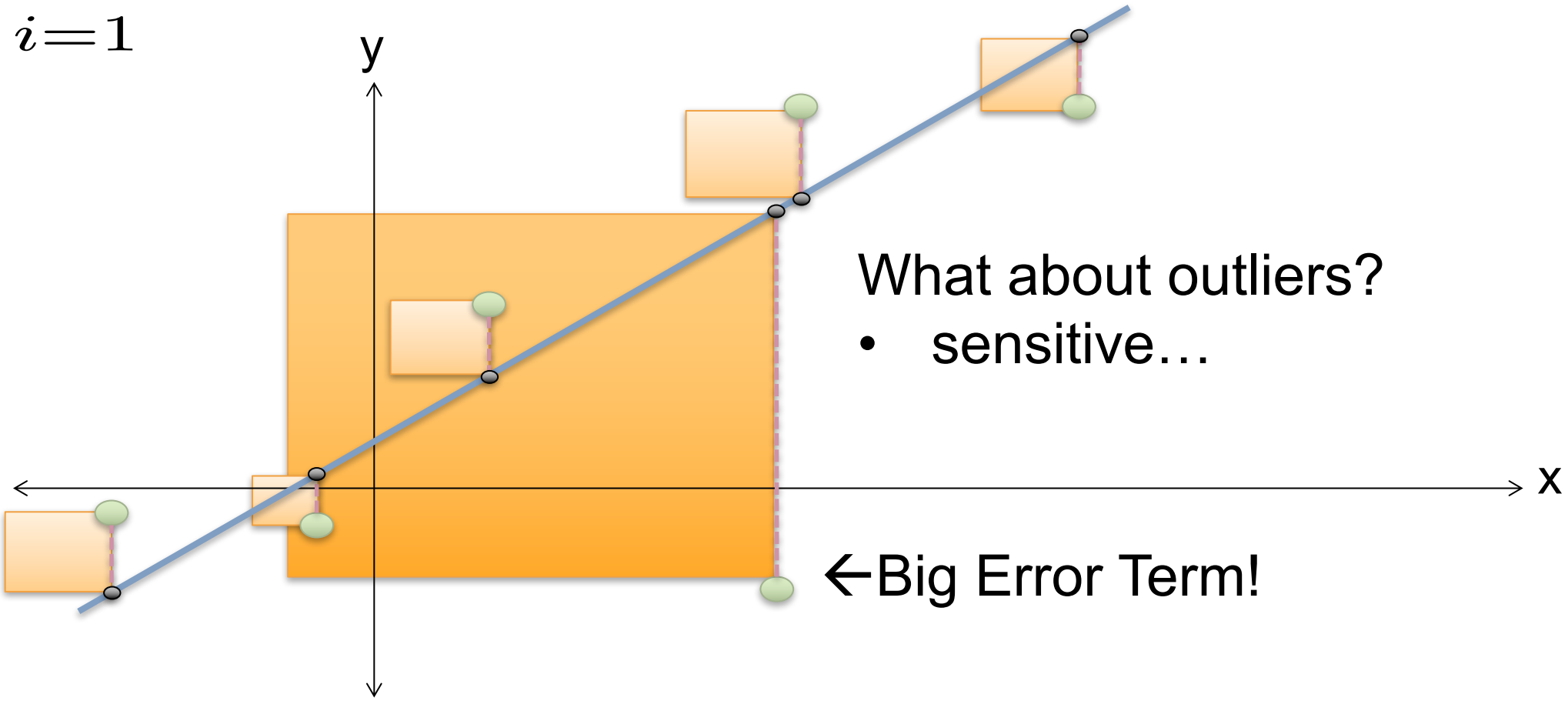
The meaning of Squared Loss (Error)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$



The meaning of Squared Loss (Error)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$



What about outliers?

- sensitive…

←Big Error Term!

Finding the Best Parameters

Model: $f_\theta(x) := \theta^T x$

Step 1: define LS **Loss Function:**

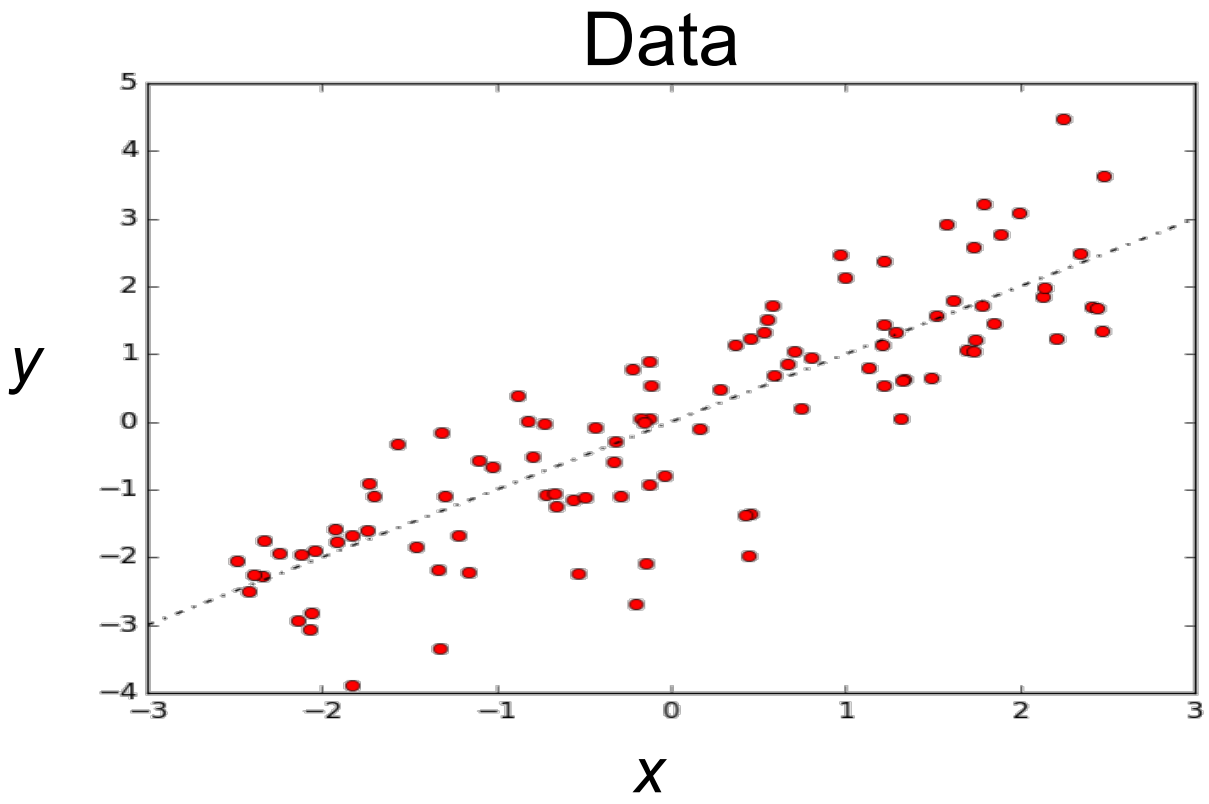
$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$$

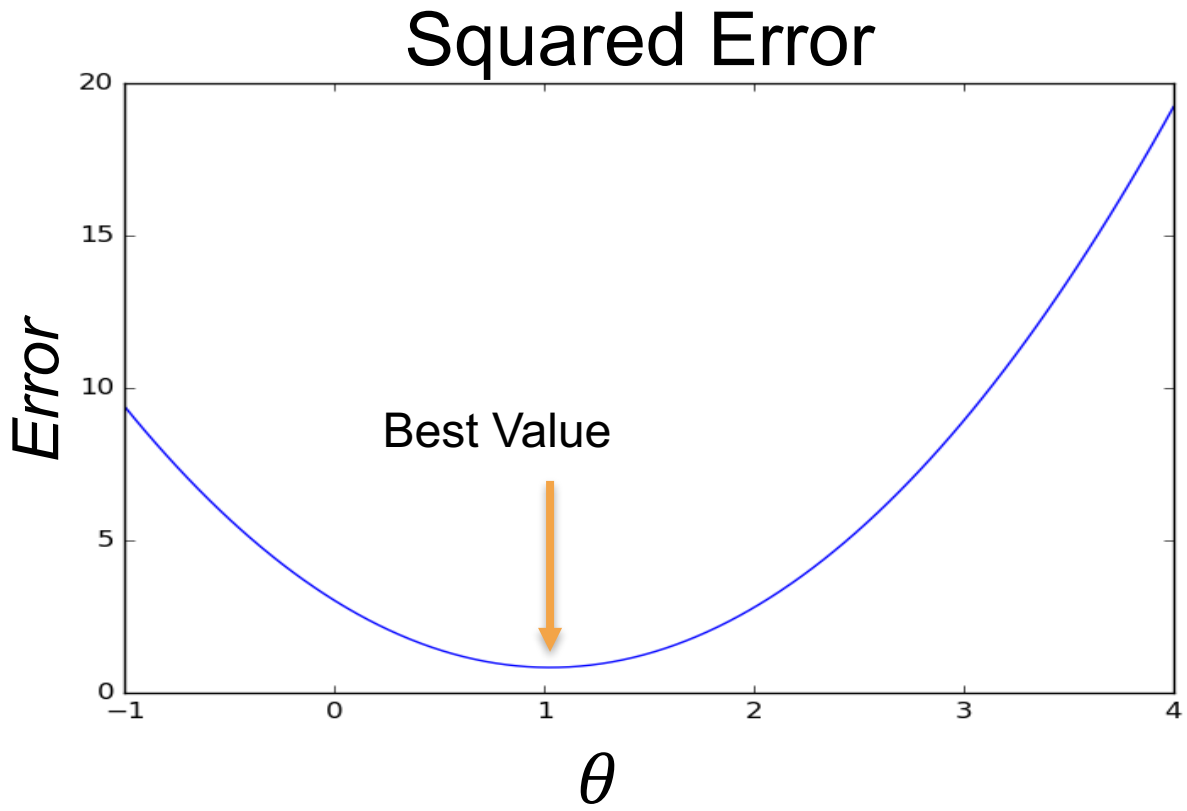
Step 2: Search for best model parameters or LS weight vector

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$$

Minimizing the Squared Error to get LS weights

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

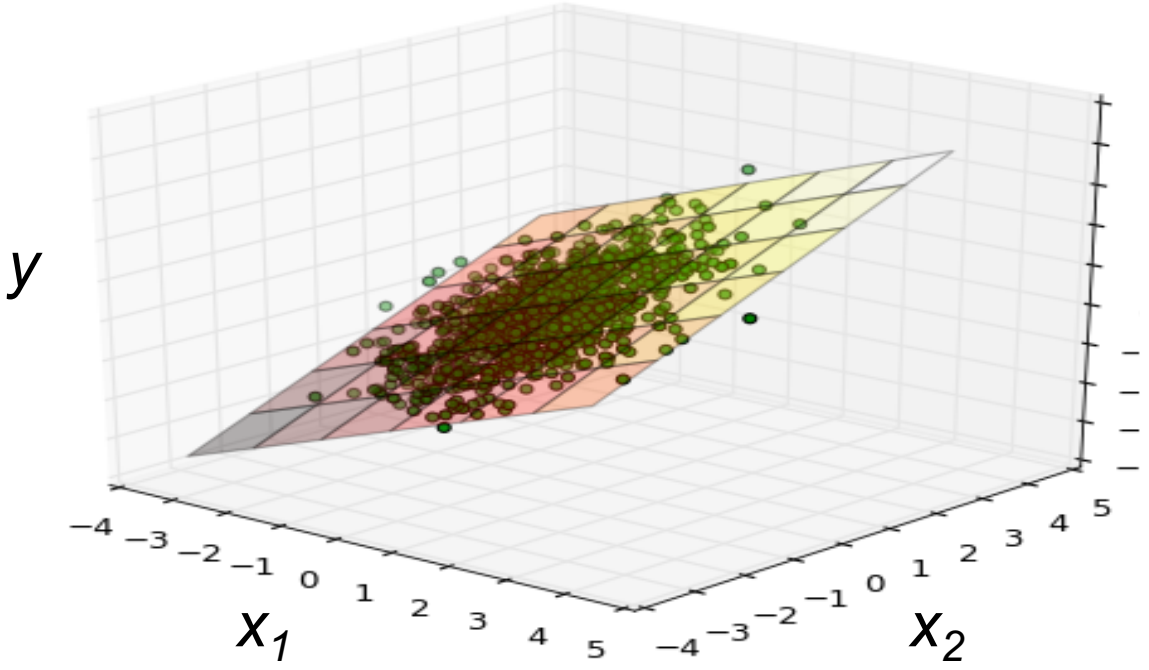




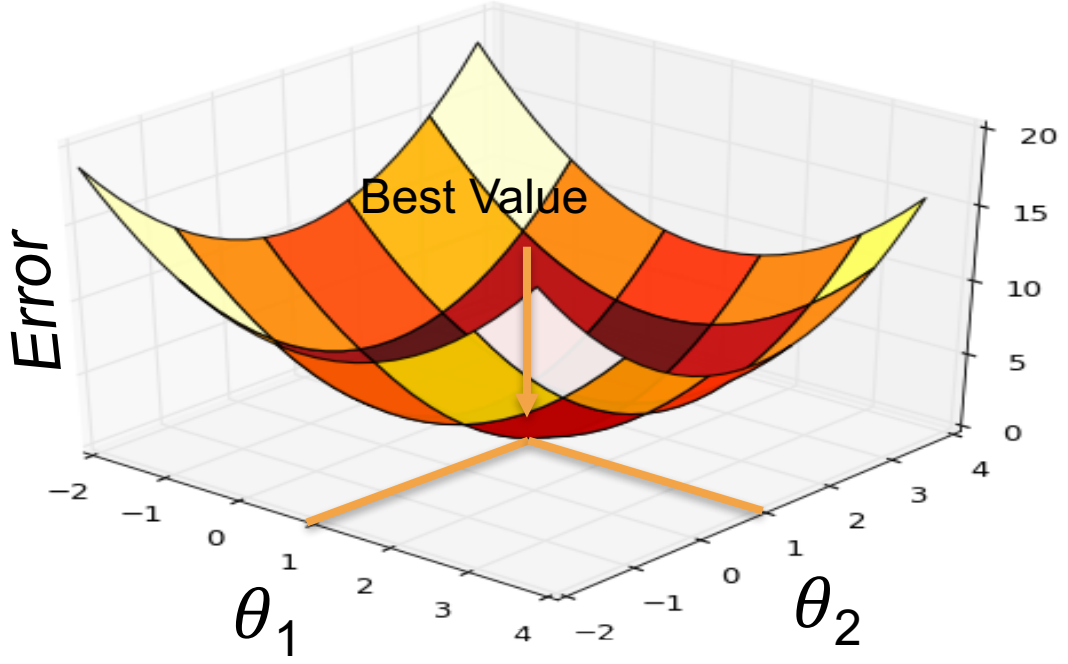
Minimizing the Squared Error

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

Data

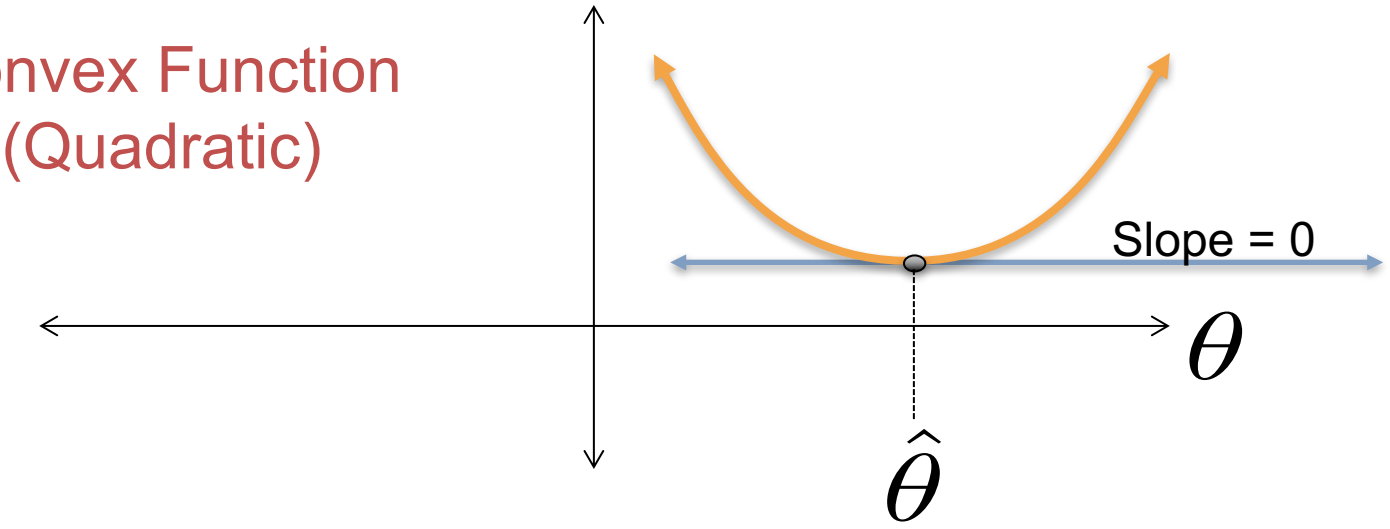


Squared Error



Minimizing the Squared Error

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$



- Take the gradient and set it equal to zero

Minimizing the Squared Error

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

- Taking the gradient

Chain Rule

$$\nabla_\theta \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2 = -2 \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i) x_i$$

$$= -2 \frac{1}{n} \sum_{i=1}^{n} y_i x_i + 2 \frac{1}{n} \sum_{i=1}^{n} (\theta^T x_i) x_i$$

- Setting equal to zero and solving for θ (sys. Linear eq.)

$$\sum_{i=1}^{n} (\theta^T x_i) x_{ij} = \sum_{i=1}^{n} y_i x_{ij} \qquad \forall j \in \{1, \ldots, d\}$$

Easier in matrix form …

Writing the data in Matrix form

- Represent data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ as:

Covariate (Design) Matrix

$$X = \underbrace{\begin{bmatrix} — & x_1 & — \\ — & x_2 & — \\ & \cdots & \\ — & x_n & — \end{bmatrix}}_{p} \in \mathbb{R}^{np}$$

n

Response Vector

$$Y = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{1} \in \mathbb{R}^{n}$$

n

Minimizing the Squared Error

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - \theta^T x_i)^2$$

- Setting gradient equal to zero to obtain the Normal Equation

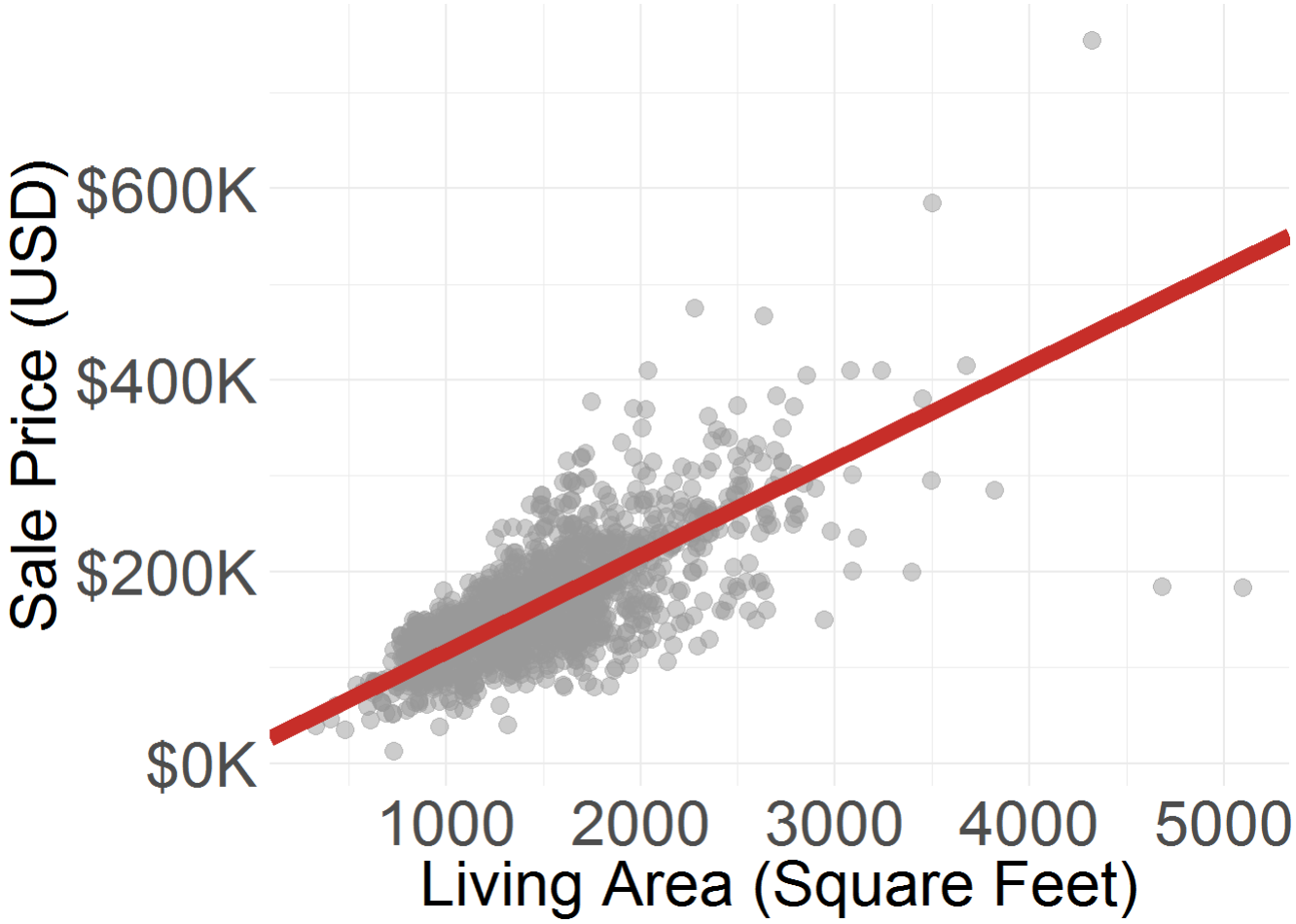
$$\sum_{i=1}^{n} (\theta^T x_i) x_i = \sum_{i=1}^{n} y_i x_i \Rightarrow X^T X \theta = X^T y$$

- If X is not singular, solving Normal Equation for θ gives the LS weight vector

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

- Solved using any standard linear algebra library

LS fit with one predictor and intercept (p=2): living area size = sq. ft.



y = 15910.14+100.21 x ($)

RMS = Root Mean Square
= $ 41.6K

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\theta}^T x_i)^2}{n-p}}$$

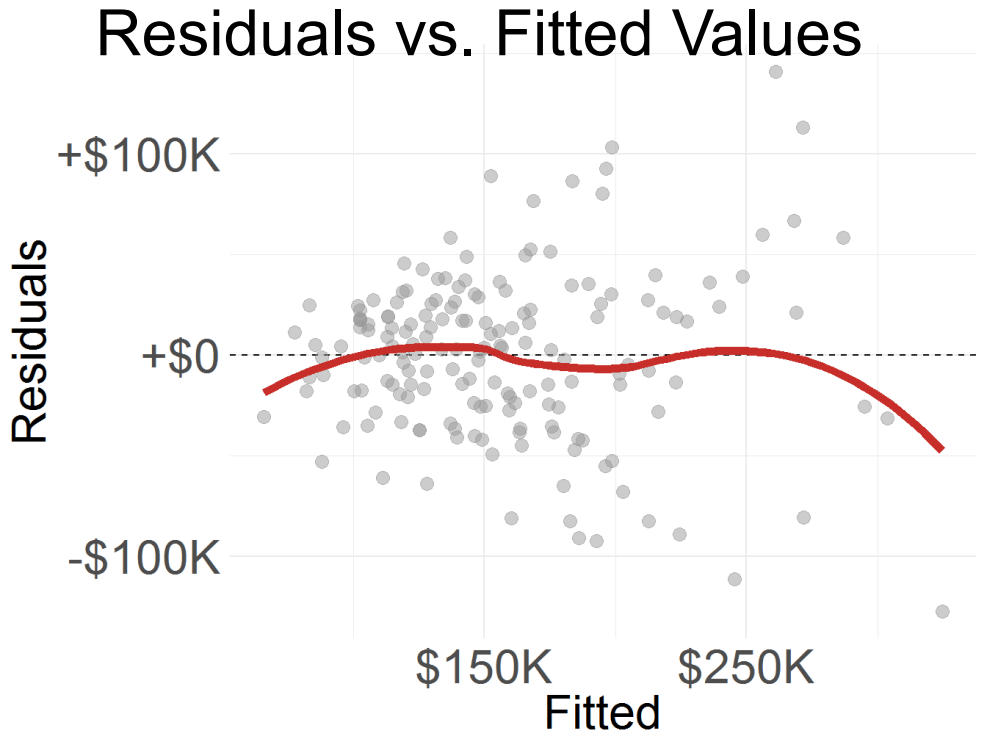
n=172 training data size
p=2

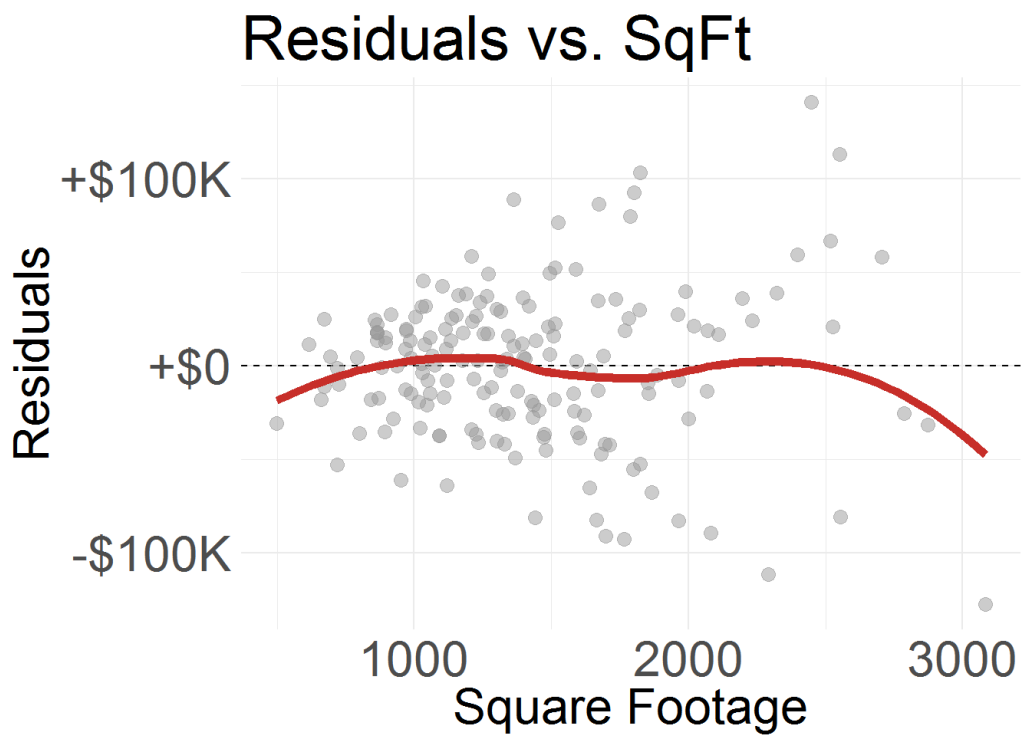
V: 1. residual plots

Fitted values are the y-values on the LS line $\hat{y}_i = \hat{\theta}^T x_i$ $\hat{\theta}$ = LS weights

Residual $e_i = y_i - \hat{y}_i$

Residual Sum of Squares (RSS) $RSS = \sum_{i=1}^{n} e_i^2$





V: 2. Interpreting the weights

- Positive weight for sq. ft. = larger houses are on average **associated** with higher prices

  It makes sense based on “domain knowledge” or our understanding of the housing market…

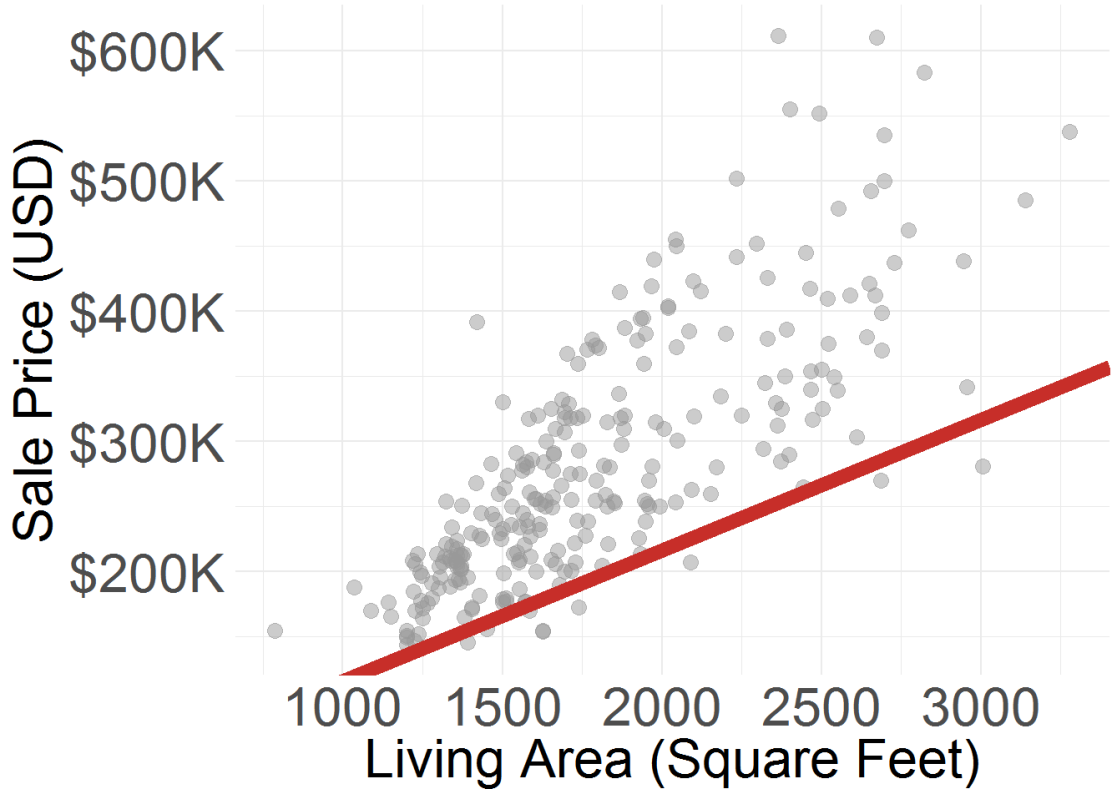
- Is it the same as “the bigger the house, the higher the price”?

- Is it the same as “if you build a big house, it will have a high price”? which is a causal statement…

V: 3. Prediction on test data

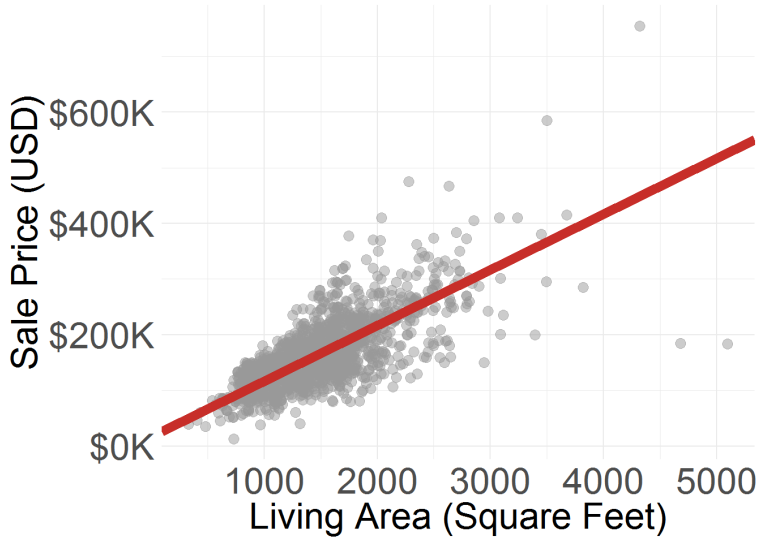
$$RMSPE = \sqrt{\frac{\sum_{i=1}^{N}(y_i^* - \hat{\theta}^T x_i^*)^2}{N}}$$ = $109.6K (N=261 test data size)

(RMSPE= Root Mean Square Prediction Error)



Training data

RMS = Root Root Mean Square
= $ 43.5K



The best data scientist is like Sherlock Holmes

There is nothing like first-hand evidence.'

Sherlock Holmes Quote

-A Study in Scarlet

First hand experience for DS is data collection itself – we should get close to it as much as we can…

We did not think about “R” until now…

- Is training representative? Is test data more like “replication” or “extrapolation”?

  Seeking “first-hand” evidence…

- What consisted of training data?

- What consisted of test data?

As it turns out…

- Training – a random sample from houses except those in the 5 most expensive neighborhoods in terms of price per sq. ft.
- Test -- Green Hills, Greens, Stone Brook, NRH, and Somerset houses
- What is Green Hills? I had to play Sherlock so I emailed a friend who lives there. He replied, twice

Greenhill in Ames ?

Is it a retirement housing for olderly ?

Yes, HA David was there toward the end. It is not a normal housing

Stone Brook is an old housing division just east of North Ridge Height. NRH is a much newer and up market, mpre pricy and bigger houses

Have not heard Greens

I plan to call a realtor from Ames…

R: circling back to the representativeness question

- Training data is not Representative of the population: all the house in Ames

- Since test data is from expensive neighborhoods, it is an extrapolation prediction problem –

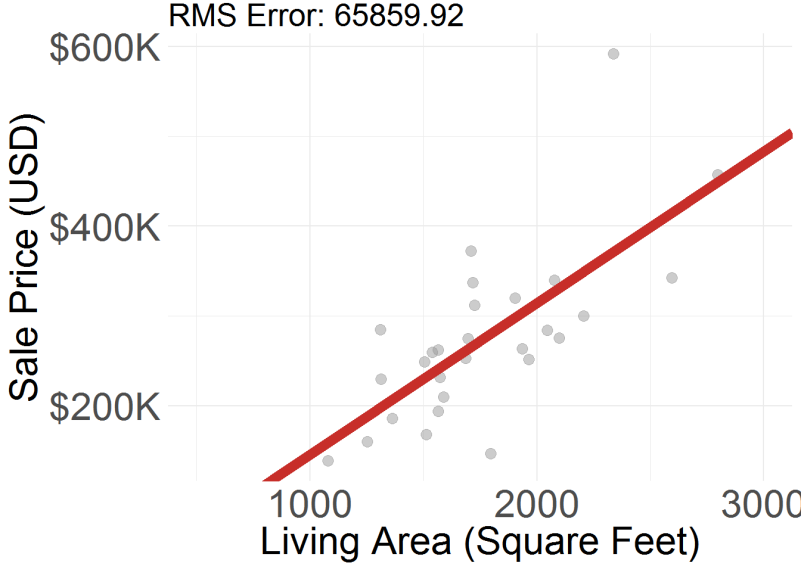
  Nursing home Green Hills is different from the rest of the expensive neighborhoods as well - very different living arrangement in nature

Using only the expensive neighborhoods for training and test (p=2): replication

Training. RMS=65.9K

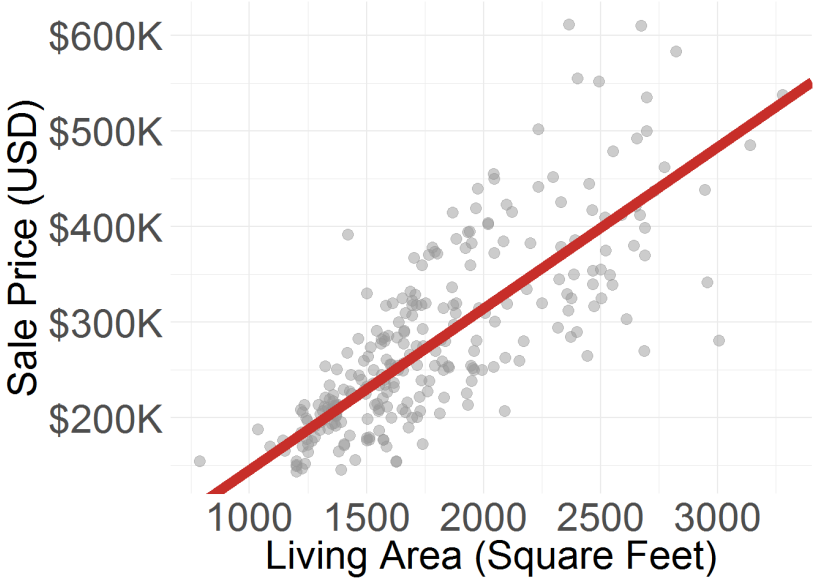
n-=28

RMS Error: 65859.92



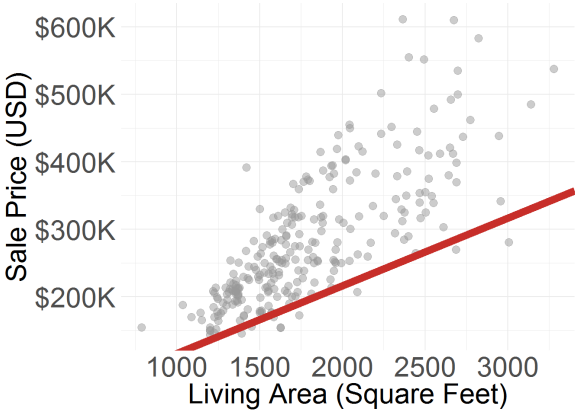
Test: RMSPE= 60.5K

N=261

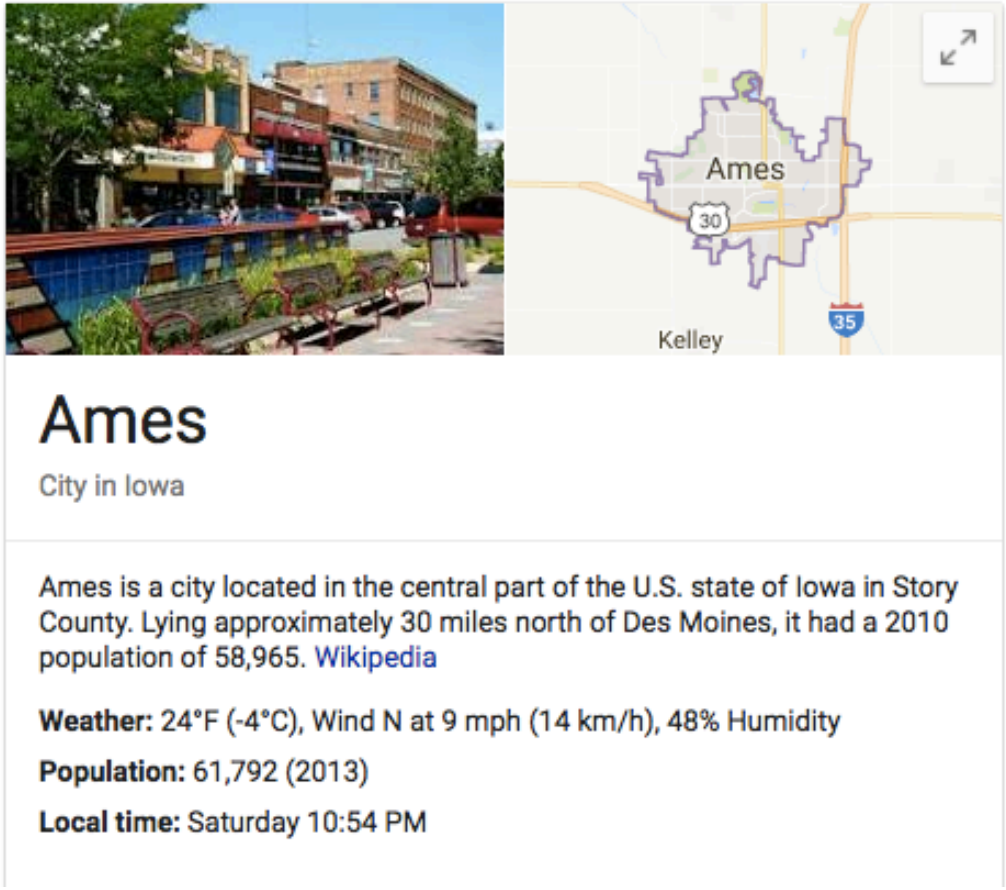


RMSPE=$109.6K

Extrapolation to test data



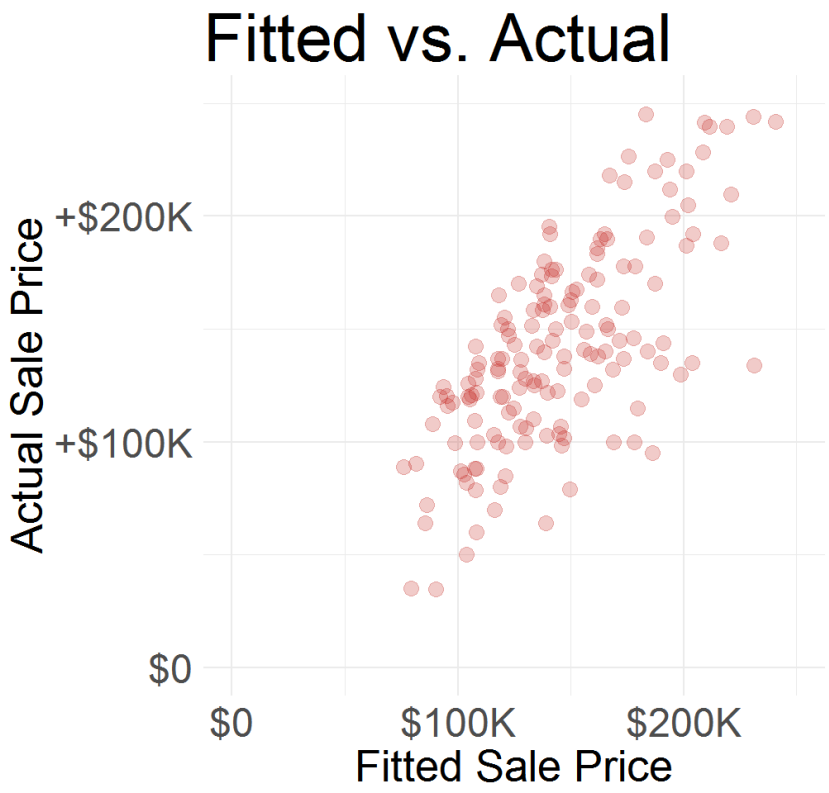
Back to Q: how much does a house in Ames Iowa sell?



Suppose you live in Ames now and want to sell your house, how would you go about estimating the sale price? You will be paid $1000 if your estimate is within 30K of the actual sale price.

LS fit with three predictors and intercept (p=4)

House price = 25636.55 + 106.29 sq. ft. – 12375.78 number of bedrooms + 1.62 lot size ($)



RMS = Root Mean Square
= $ 43.1K

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\theta}^T x_i)^2}{n-p}}$$

n=172 training data size
P=4

Recall RMS=43.5K for
Sq.ft + intercept fit

Summary

- QPRV an effective conceptual framework for prediction problems
  - Data collection process is key
  - pay attention to replication vs extrapolation

- LS Formulation, sensitive to “outliers”, which can be the most informative

- LS weight/coeff. formula

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

- More predictors always help reduce training error (RSS); they can help prediction error as well, but not always