CS 383: Machine Learning

Prof Adam Poliak

Fall 2024

09/17/2024

Lecture 05

Announcements

HW02 is due Sunday night

- **Reading quiz: Thursday**
  - Duame 7.6 (2+ pages)
  - ISL 59-63 (4+ pages)

- Midterm 1: Thursday October 3rd

Decision Trees: base cases summary

1) All examples have the same label ★
2) No more features remain to split on ★
3) Partition does not contain any examples
4) Maximum depth reached

Duame

5) (recommended) No features produce information gain

   i.e. all have same remaining features but there is still label heterogeneity

Decision Trees: implementation ideas

1) Make sure you can accommodate more than two children (i.e. not a binary tree)

2) Make sure your prediction/classification algorithm is recursive

3) You can parse the feature name to figure out continuous/discrete and how to classify

```
age<=44.5
```

Implementation Suggestions

- Start slow with entropy! Build up function by function
- Think back to trees in data structures
- Distinguish between data (X,y) and options for data (values for each feature, classes for y)

Outline

Continuous Features in Decision Trees

Learning problem so far + terminology

Bias-Variance Tradeoff

Linear regression

Continuous Features in Decision Tree

Temperature	Play Tennis?
80	Yes
48	Yes
60	Yes
48	Yes
40	No
48	No
90	No

Continuous Features in Decision Tree

1. Sort examples by feature values

2. Merge repeat values

3. Split when label changes

Outline

Continuous Features in Decision Trees

Learning problem so far + terminology

Bias-Variance Tradeoff

Linear regression

Learning Problem so far

Performance on training data <u>overestimates</u> accuracy

We must use a <u>held aside</u> test set to evaluate

Both training and testing data should be drawn from the same distribution

Training/test data should be drawn from the same distribution as seen in deployment (ideally)

Loss functions

- E.g., zero-one loss
  - Simple accuracy - is prediction right?
  - For binary or multi-class prediction

$$l(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

Loss functions

- E.g., zero-one loss
  - Simple accuracy - is prediction right?
  - For binary or multi-class prediction

$$l(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

- E.g., squared loss
  - For regression

$$l(y, \hat{y}) = (y - \hat{y})^2$$

Loss functions

- E.g., zero-one loss
  - Simple accuracy - is prediction right?
  - For binary or multi-class prediction

$$l(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{otherwise} \end{cases}$$

- E.g., squared loss
  - For regression

$$l(y, \hat{y}) = (y - \hat{y})^2$$

Absolute loss (also for regression)

$$\ell(y, \hat{y}) = |y - \hat{y}|$$

Formalize Learning Problem

Given:

- Loss function ℓ
- A sample of data D from an unknown distribution of all data $\boldsymbol{D}$
- A hypothesis space $H = \{h | h : X \rightarrow Y\}$

Find:

- A function $f(X) \rightarrow y$ that
- Minimizes error $\boldsymbol{D}$ with respect to ℓ

Inductive bias

class A

Training Data

class B

Testing Data

Inductive bias

class A

Training Data

class B

Testing Data

A A

B B

Inductive bias

class A

Training Data

class B

Testing Data

A

A

B

B

A: “fly”

B: “no fly”

Inductive bias

Training Data

Testing Data

A: “bird”

B: “mammal”

Why might learning fail?

Noise in the training data

- Typos in a restaurant review

Available features are insufficient

- x-ray does not capture the medical issue

“Correct” prediction is up to interpretation

- Parental controls on web content

Learning algorithm cannot cope with the data

Adapted from Duame Chap 2

Hyperparameters

- Difficult to define precisely, but typically a parameter that controls other parameters
- What is one hyperparameter in decision trees?
- We can't choose hyperparameters via test data (breaks cardinal rule!)
- But we can use *validation data*

General Training Approach

1. Split your data into 70% training data, 10% development data and 20% test data.
2. For each possible setting of your hyperparameters:
   (a) Train a model using that setting of hyperparameters on the training data.
   (b) Compute this model's error rate on the development data.
3. From the above collection of models, choose the one that achieved the lowest error rate on development data.
4. Evaluate that model on the test data to estimate future test performance.

Duame Chap 2

Outline

Continuous Features in Decision Trees

Learning problem so far + terminology

Bias-Variance Tradeoff

Linear regression

Regression Setup

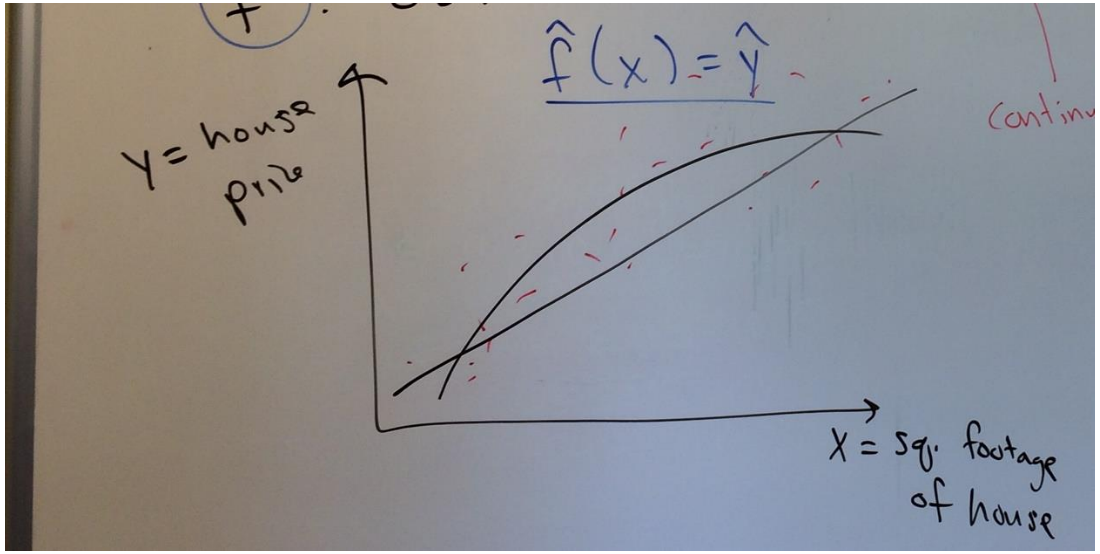
What we observe

Error term
Mean 0
(independent of x)

Model: $y = f(x) + \varepsilon$

We choose $\hat{f}$ - our estimate of f

$$\widehat{f(x)} = \hat{y}$$



Loss Function

$\ell\left(y, \hat{f}(x)\right)$ quantifies how far our prediction is from the true value

We want to minimize Expected Loss:

$$\mathbb{E}_{(x,y)}\left[\ell\left(y, \hat{f}(x)\right)\right]$$

Expected Values

Weight

$$\mathbb{E}[X] = \sum_{v \in X} p(X = v)\, v$$

Weighted die example

Expected Value Rules

Additivity of Expectation

$$\mathbb{E}[X+Y] =$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

Linearity of Expectation

$$\mathbb{E}[\alpha X] =$$

$$= \alpha\mathbb{E}[X]$$

https://prob140.org/textbook/content/Chapter_08/04_Additivity.html

Expected Values

Quick Check

Let X and Y be random variables on the same space, with $E(X) = 5$ and $E(Y) = 3$.

(a) Find $E(X - Y)$.

(b) Find $E(2X - 8Y + 7)$.

https://prob140.org/textbook/content/Chapter_08/04_Additivity.html

Expected loss

Read it as (x,y) has distribution $\mathcal{D}$

$(x, y) \sim \mathcal{D}$

Probability of (x,y) occurring

loss

Expected Loss

$$\sum_{x,y \in D} D(x, y)\ \ell\left(y, \hat{f}(x)\right)$$

$$\frac{1}{n}\sum_{i}^{n} \ell\left(y_i, \hat{f}(x_i)\right)$$

Mean Squared Error

Squared error:
$\ell(y, \hat{y}) = (y - \hat{y})^2$

Mean Square Error (MSE):
$$\frac{1}{n}\sum_{i}^{n}(y_i - \widehat{y_i})^2$$

$$\mathbb{E}[MSE] = \frac{1}{n}\sum_{i}^{n}\mathbb{E}[(y_i - \widehat{y_i})^2]$$

Mean Squared Error

$$\mathbb{E}[(y-\hat{y})^2]$$

$$= \mathbb{E}\left[\left(y-\hat{f}\right)^2\right]$$

error ε

model issue

$$= \mathbb{E}\left[\left(y-f+f\ -\ \hat{f}\right)^2\right]$$

$$= Var(\varepsilon) + \ \mathbb{E}\left[\left(f\ -\ \hat{f}\right)^2\right]$$

Mean Squared Error

$$\mathbb{E}[(y-\hat{y})^2] = Var(\varepsilon) + \mathbb{E}\left[\left(f - \hat{f}\right)^2\right]$$

$$\mathbb{E}\left[\left(f - \hat{f}\right)^2\right] = \mathbb{E}\left[\left(f - \mathbb{E}[\hat{f}] + \mathbb{E}[\hat{f}] - \hat{f}\right)^2\right]$$

$$= bias\left(\hat{f}\right)^2 + var\left(\hat{f}(x)\right)$$

$$\mathbb{E}[MSE] = bias\left(\hat{f}\right)^2 + var\left(\hat{f}\right) + var(\varepsilon)$$

Mean Squared Error

$$\mathbb{E}[MSE] \; = bias\left(\hat{f}\right)^2 + var\left(\hat{f}\right) + var(\varepsilon)$$

Bias: approximation error

error introduced by approximated a real-life problem

Variance: estimation error

amount $\hat{f}$ would change if we trained on different data

Bias Variance Tradeoff

Assessing Model Accuracy

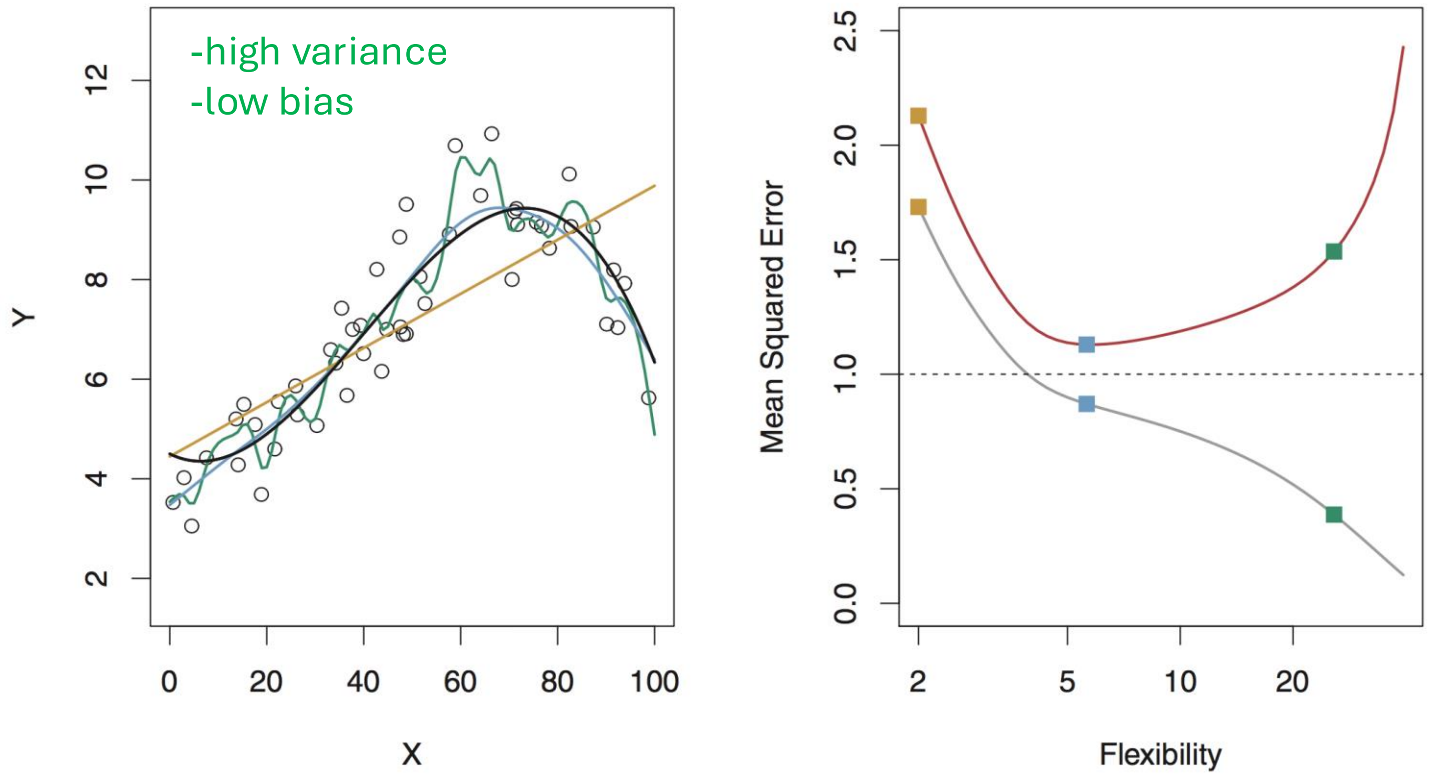


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

ISL, Fig 2.9

Outline

Continuous Features in Decision Trees

Learning problem so far + terminology

Bias-Variance Tradeoff

Linear regression

Goals of Inference

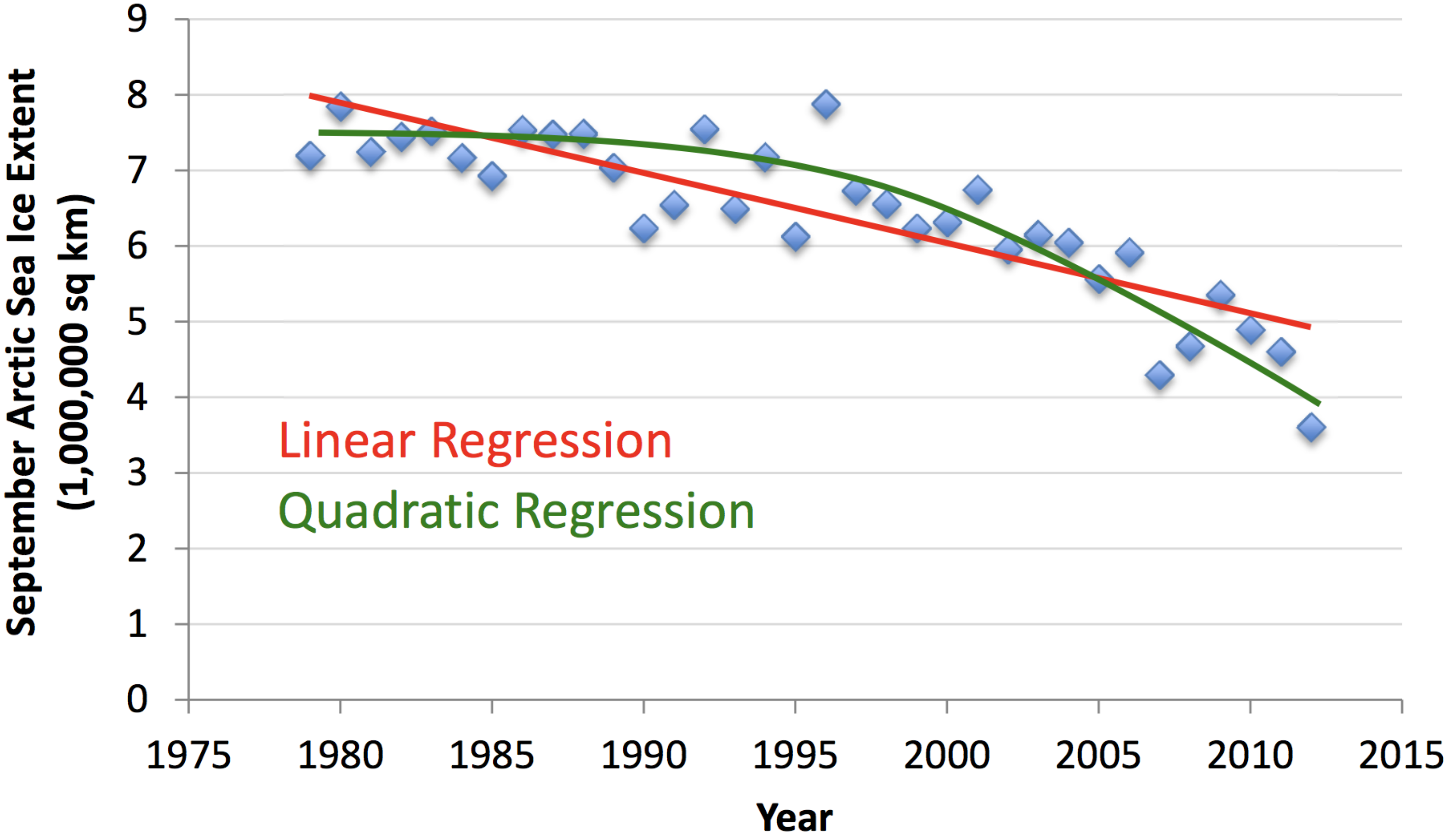
1) Which of the features/explanatory variables/predictors (x) are associated with the response variable (y)?

2) What is the relationship between x and y?

3) Is a linear model enough?

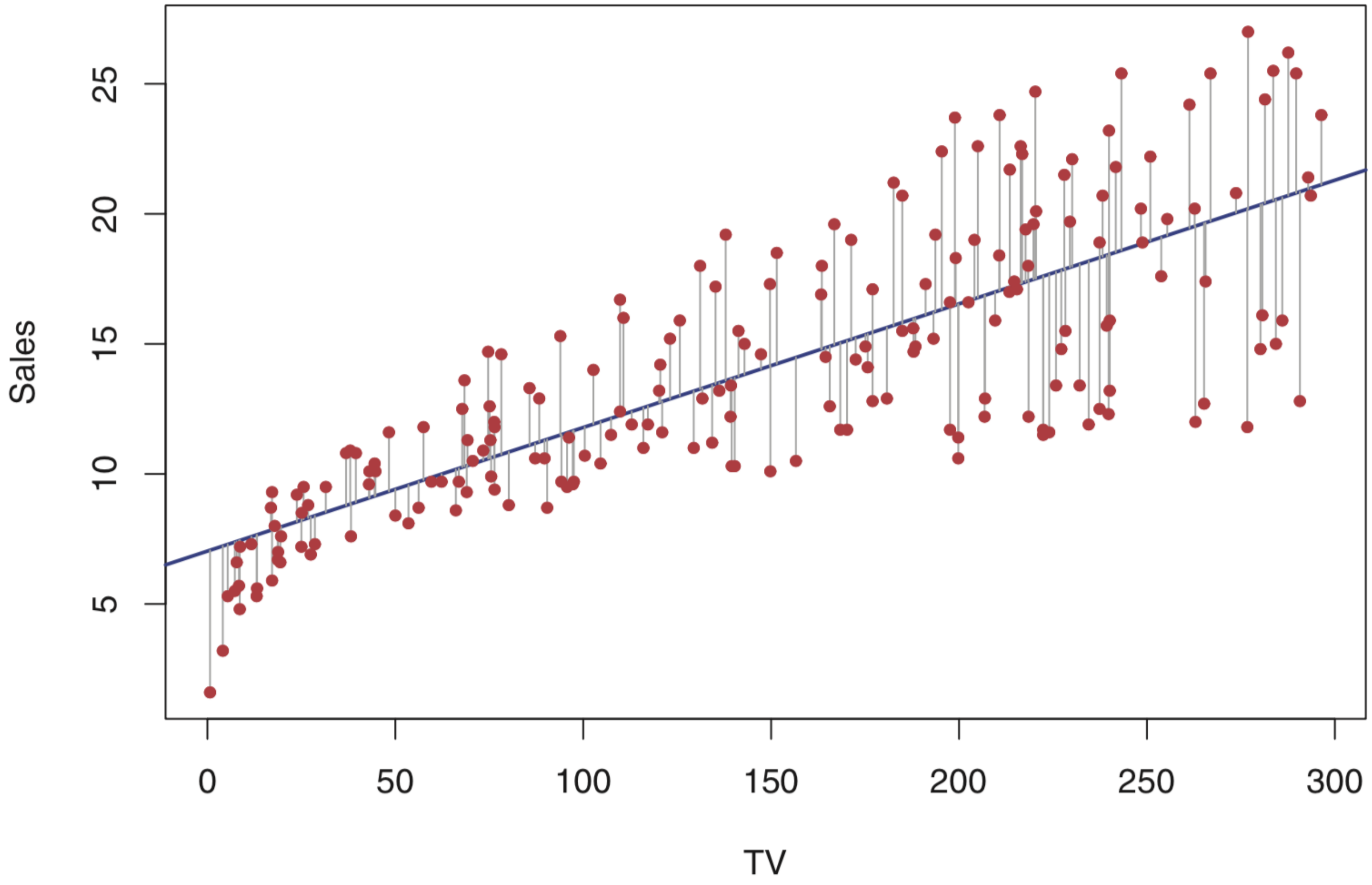
4) Can we predict y given a new x?

Regression Example



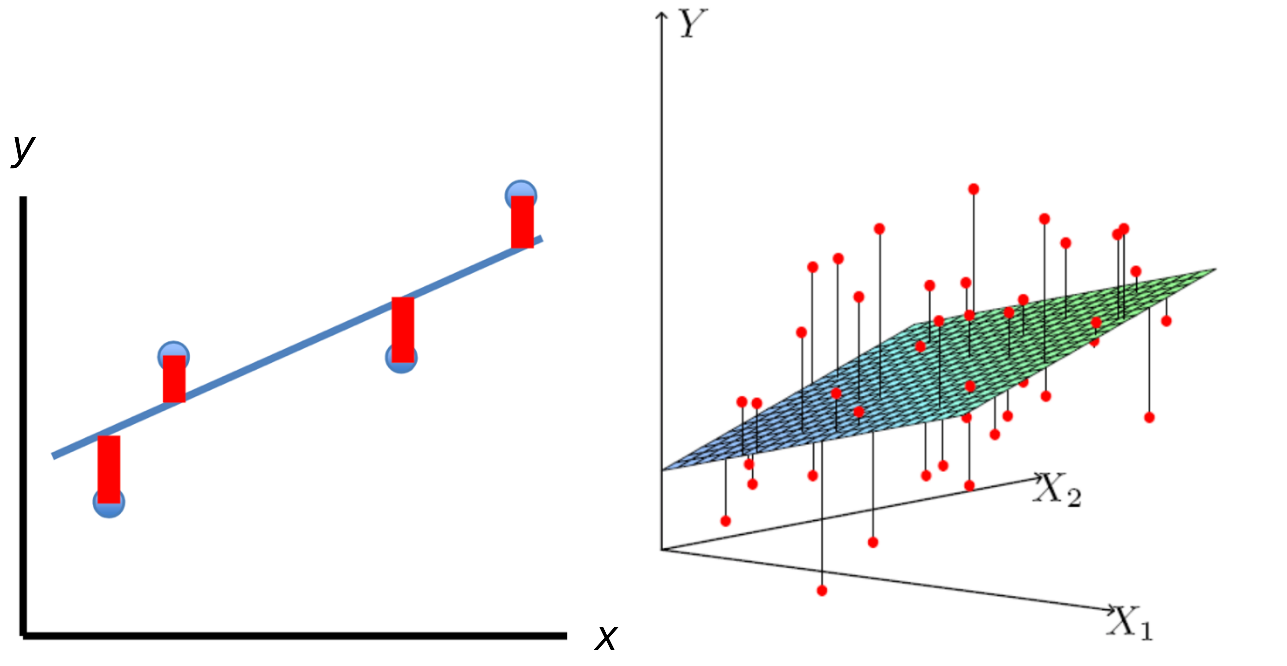
[Data from G. Witt. Journal of Statistics Education, Volume 21, Number 1 (2013)]

Example: predict sales from TV advertising budget



ISL: Figure 3.1

Cost Function: sum of squared errors



p=1 *p*=2