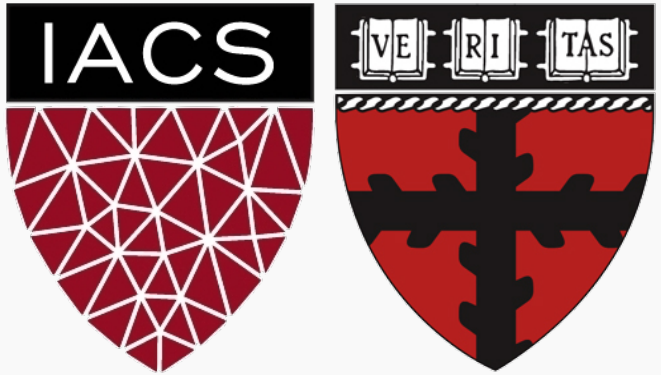
Lecture 1: Introduction to Regression
HUB AI

Pavlos Protopapas
Institute for Applied Computational Science
Harvard



Lecture Outline

Logistics

Class Structure

Data

Statistical Modeling I

k-Nearest Neighbors (kNN)

Model Fitness

How does the model perform predicting?

Comparison of Two Models

How do we choose from two different models?

Class Logistics

What

Online

Week	Date	Day	Session Type	Theme	Topic
1	5/3/2019	Online Friday 5/3 8:00am Boston Time	Lecture	Introduction to the Data Science	Introduction to the course. Brief review of prerequisites. Introduction to statistical learning and data science. Basic concepts. kNN Regression 1.
2	5/10/2019	Online Friday 5/10 8:00am Boston Time	Lecture	Regression	kNN Regression 2 with examples. Model estimation, error evaluation, model fitness and model comparison.
3	5/17/2019	Online Friday 5/17 8:00am Boston Time	Lecture	Regression	Linear Regression 1. Linear model and model fitting with an example.
4	5/24/2019	Online Friday 5/24 8:00am Boston Time	Lecture	Regression	Lenear Regression 2. Meaning of measurement error, significance of the predictors and bootstrapping.
5	5/31/2019	Online Friday 5/31 8:00am Boston Time	Lecture	Regression	Multiple Linear Regression, Polynomial Regression, Overfitting, Regularization and Model Selection

What

Offline, aka China

Week 1

6	7/15/2019	Mon	Lecture	Decision Trees and Ensembles	Logistic Regression. Overfitting, Regularization and Model Selection
			Lab		
6	7/16/2019	Tue	Hackathon/Projects	Towards milestone 1	Scraping Data
6	7/17/2019	Wed	Lecture	Decision Trees and Ensembles	Classification Trees, Regression Trees, CART
			Lab		Predict shopping trends with Tree Regressions using the Black Friday dataset
6	7/18/2019	Thu	Hackathon/Projects	Towards milestone 1	

What

Offline, aka China

Week 2

7	7/22/2019	Mon	Lecture	Decision Trees and Ensembles	Regression Trees, Bagging and Random Forest
			Lab		Implementing Classification in Python with Random Forests, Boosting and PCA using the Black Friday and UCI mHealth Datasets (PART A)
7	7/23/2019	Tue	Hackathon/Projects	Towards milestone 2	EDA
7	7/24/2019	Wed	Lecture	ANNs	Artificial Neural Networks 1 - Perceptron, Back Propagation
			Lab		MNIST digit classification and puppy classification using MLPs with Numpy and Keras, Regularization in MLPs
7	7/25/2019	Thu	Hackathon/Projects	Towards milestone 2	Additional Feeature Engineering

What

Offline, aka China

Week 3

8	7/29/2019	Mon	Lecture	ANNs	Artificial Neural Networks 2 - Anatomy of ANN, Regularization
			Lab		MNIST digit classification and puppy classification using MLPs with Numpy and Keras, Regularization in MLPs
8	7/30/2019	Tue	Hackathon/Projects	Towards milestone 3	Creating Model
8	7/31/2019	Wed	Lecture	CNNs	Artificial Neural Networks for Image Analysis, Introduction to Convolutional Neural Networks
			Lab	CNNs	CNNs for object classification with Keras
8	8/1/2019	Tue	Hackathon/Projects	Towards milestone 3	Model Comparison and Error Analysis and Reporting

Who

Pavlos Protopapas, Scientific Director of the Institute for Applied Computational Science (IACS). Teaches CS109a, CS109b and the Capstone course for the Data Science masters program. He has taught in GEC 3 courses already. His research is in astrostatistics and excited about the new telescopes coming online in the next few years.

Who: Teaching Fellow

Patrick Ohiomoba has a wealth of experience as teaching fellows for Data Science and Machine Learning courses at Harvard's School of Engineering and Applied Sciences. He is from Nigeria and he likes among other things eating good food, following sports (football) and keeping up with the latest ML and AI developments.

How

Lectures:

Fridays @8am [EST] (except today)

Labs:

Saturdays @8am [EST] Patrick Ohiomoba

Meeting ID: 943 758 711

https://zoom.us/j/943758711

Predicting Municipal Rental Prices

At some point in time, each of us would have went through the process of either renting or buying a house.

The renting of residential properties is an economic process that serves a really important role in many communities (especially urban communities) in the United States. First of all residential leasing is one of the key pathways for property owners and landlords to realize value from their property. In many cases (especially in communities where multi-family homes are common) renting is an essential part of the financial calculus that allows certain classes of people to fulfill their dreams of home/property ownership. A second factor is that residential rentals are an essential mechanism for people to live in a community without the often onerous burdens of purchasing properties outright. In fact for a number of the most iconic neighborhoods in the urban United States the essential character of the neighborhood (as well as the properties) would change drastically (in most cases for the worse) if residential renting was abolished.

Data Science

Why?

50 Best Jobs in America

Awards

Best Places to Work

Highest Rated CEOs

Best Places to Interview

Lists

Best Jobs

Best Cities for Jobs

Highest Paying Jobs

Oddball Interview Questions

Trends

Overview

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? Find out how.

United States | 2017

12k Shares

1 Data Scientist

4.8 / 5
Job Score

4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

4,184
Job Openings

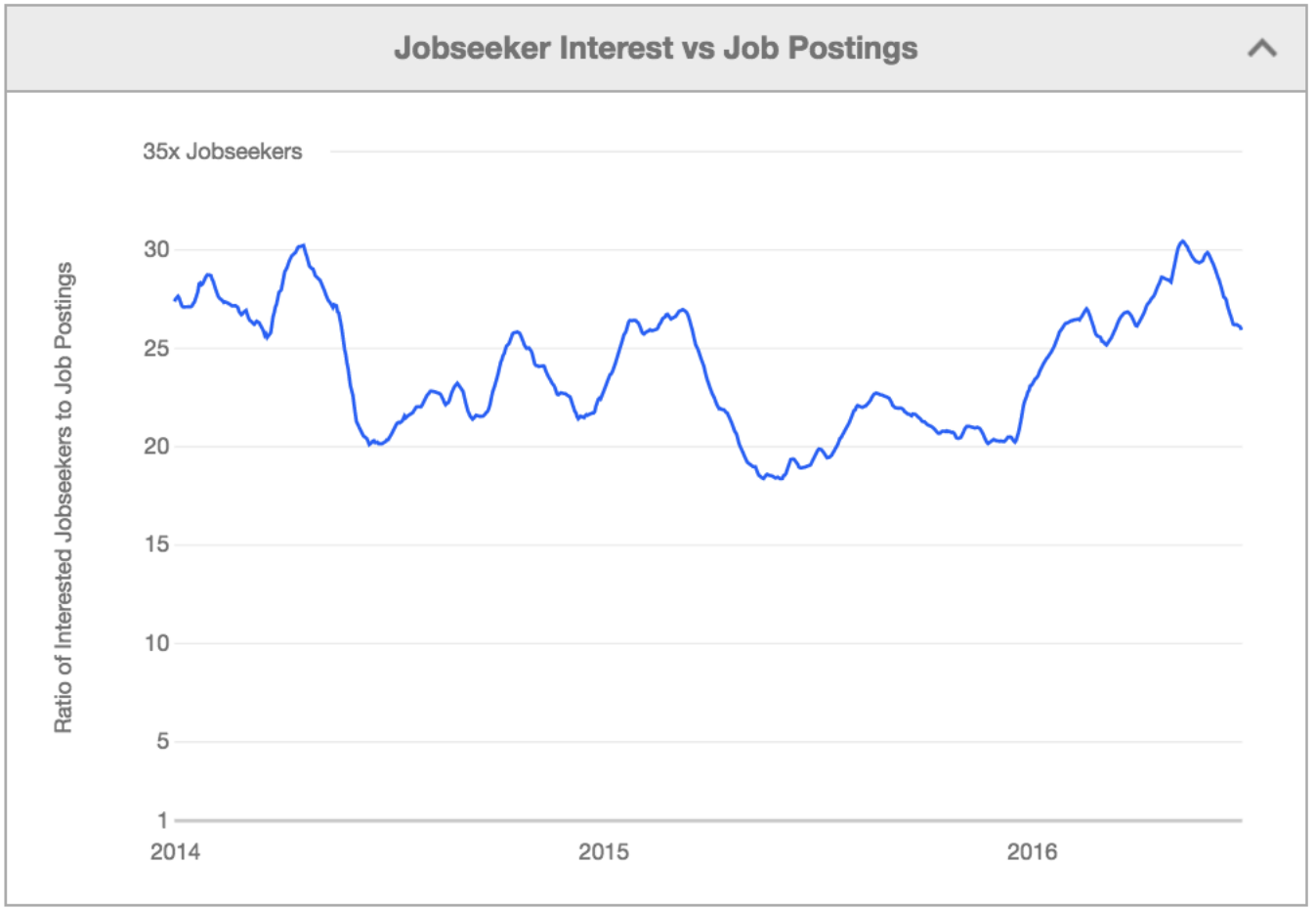
View Jobs

2 DevOps Engineer

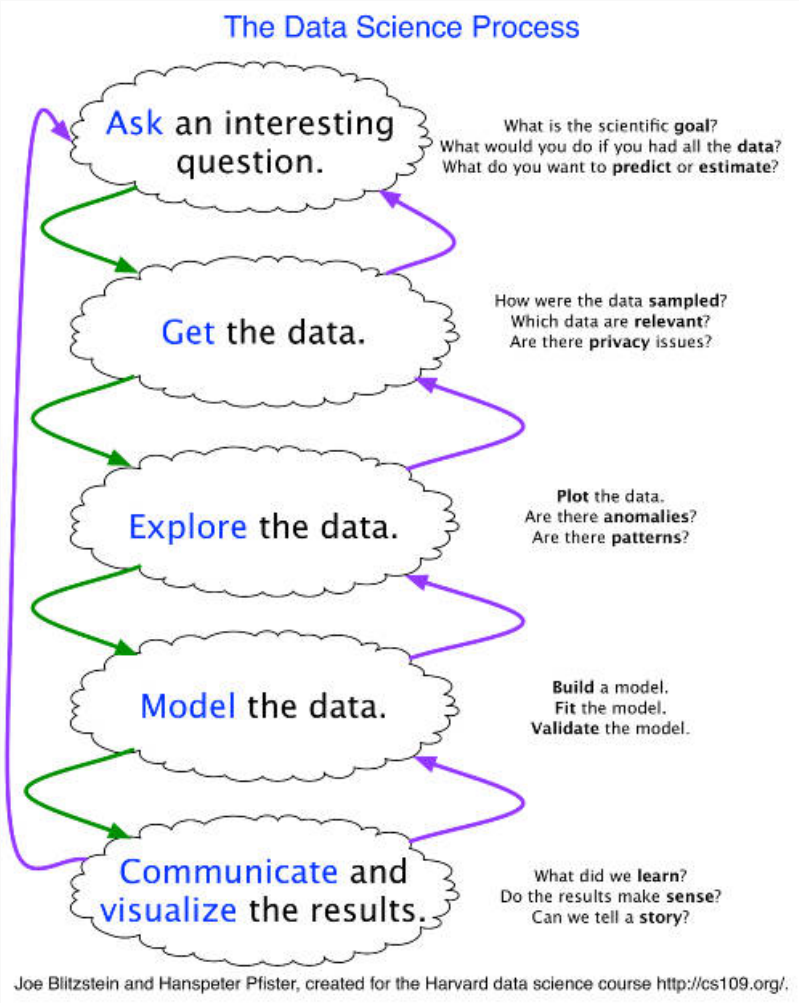
Why?


indeed
Find Jobs
Company Reviews
Find Salaries
Find Resumes
Employers / Post Job
Upload your resume
Sign in
Job Category Trends
Job Trends
"Data Scientist" Job Trends
"Data Scientist"
+ Add Term
Find Trends
Job Postings
Jun 16, 2017
"Data Scientist": 0.151%
Percentage of Matching Job Postings (%)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
2014
2015
2016
2017
https://plus.google.com/share?url=https%3A%2F%2Fwww.indeed.com%2Fjobtrends%2Fq-%2522Data-Scientist%2522.html

Why?


Jobseeker Interest vs Job Postings
35x Jobseekers
30
25
20
15
10
5
1
Ratio of Interested Jobseekers to Job Postings
2014
2015
2016
Find "Data Scientist" jobs

Data Science Process



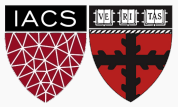
Let’s start

Predicting a Variable

Let's image a scenario where we'd like to predict one variable using another (or a set of other) variables.

Examples:

- Predicting the amount of view a YouTube video will get next week based on video length, the date it was posted, previous number of views, etc.
- Predicting which movies a Netflix user will rate highly based on their previous movie ratings, demographic data etc.



Data

What are Data?

"A datum is a single measurement of something on a scale that is understandable to both the recorder and the reader. Data are multiple such measurements." Claim: everything is (can be) data!

Where do data come from?

- **Internal sources:** already collected by or is part of the overall data collection of you organization. For example: business-centric data that is available in the organization data base to record day to day operations; scientific or experimental data

- **Existing External Sources:** available in ready to read format from an outside source for free or for a fee. For example: public government databases, stock market data, Yelp reviews, [your favorite sport]-reference

- **External Sources Requiring Collection Efforts**: available from external source but acquisition requires special processing. For example: data appearing only in print form, or data on websites

Ways to gather online data

How to get data generated, published or hosted online:

- **API (Application Programming Interface):** using a prebuilt set of functions developed by a company to access their services. Often pay to use. For example: Google Map API, Facebook API, Twitter API

- **RSS (Rich Site Summary):** summarizes frequently updated online content in standard format. Free to read if the site has one. For example: news-related sites, blogs

- **Web scraping:** using software, scripts or by-hand extracting data from what is displayed on a page or what is contained in the HTML file.

Web scraping

- Why do it? Older government or smaller news sites might not have APIs for accessing data, or publish RSS feeds or have databases for download. Or, you don't want to pay to use the API or the database.
- How do you do it? *See lab session.*
- Should you do it?
  - You just want to explore: Are you violating their terms of service? Privacy concerns for website and their clients?
  - You want to publish your analysis or product: Do they have an API or fee that you are bypassing? Are they willing to share this data? Are you violating their terms of service? Are there privacy concerns?

Types of data

What kind of values are in your data (data types)?

Simple or atomic:

- **Numeric**: integers, floats
- **Boolean**: binary or true false values
- **Strings**: sequence of symbols

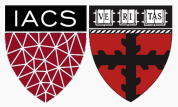
Data Types

What kind of values are in your data (data types)? Compound, composed of a bunch of atomic types:

- **Date and time**: compound value with a specific structure
- **Lists:** a list is a sequence of values
- **Dictionaries:** A dictionary is a collection of key-value pairs, a pair of values x : y where x is usually a string called the key representing the "name" of the entry, and y is a value of any type.

Example: Student record: what are x and y?

- First: Pavlos
- Last: Ohiomoba
- Classes: [CS-109A]



Data storage

How is your data represented and stored (data format)?

- **Tabular Data:** a dataset that is a two-dimensional table, where each row typically represents a single data record, and each column represents one type of measurement (csv, dat, xlsx, etc.).

- **Structured Data:** each data record is presented in a form of a [possibly complex and multi-tiered] dictionary (json, xml, etc.)

- **Semistructured Data:** not all records are represented by the same set of keys or some data records are not represented using the key-value pair structure.

Data format

How is your data represented and stored (data format)?

- Textual Data
- Temporal Data
- Geolocation Data

Tabular Data

In tabular data, we expect each record or observation to represent a set of measurements of a single object or event.

First Look At The Data

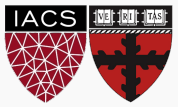
```
In [27]: hubway_data = pd.read_csv('hubway_trips.csv', low_memory=False)
         hubway_data.head()
```

Out[27]:

	seq_id	hubway_id	status	duration	start_date	strt_statn	end_date	end_statn	bike_nr	subsc_type	zip_code	birth_d
0	1	8	Closed	9	7/28/2011 10:12:00	23.0	7/28/2011 10:12:00	23.0	B00468	Registered	'97217	1976.0
1	2	9	Closed	220	7/28/2011 10:21:00	23.0	7/28/2011 10:25:00	23.0	B00554	Registered	'02215	1966.0
2	3	10	Closed	56	7/28/2011 10:33:00	23.0	7/28/2011 10:34:00	23.0	B00456	Registered	'02108	1943.0
3	4	11	Closed	64	7/28/2011 10:35:00	23.0	7/28/2011 10:36:00	23.0	B00554	Registered	'02116	1981.0
4	5	12	Closed	12	7/28/2011 10:37:00	23.0	7/28/2011 10:37:00	23.0	B00554	Registered	'97214	1983.0

Each type of measurement is called a **variable** or an **attribute** or a **predictor** of the data (e.g. `seq_id`, `status` and `duration` are variables or attributes). The number of attributes is called the **dimension**. These are often called **features**.

We expect each table to contain a set of **records** or **observations** of the same kind of object or event (e.g. our table above contains observations of rides/checkouts).



Types of Data

We'll see that it's important to distinguish between classes of variables or attributes based on the type of values they can take on.

- **Quantitative variable:** is numerical and can be either:
    - discrete - a finite number of values are possible in any bounded interval. For example: "Number of siblings" is a discrete variable
    - continuous - an infinite number of values are possible in any bounded interval. For example: "Height" is a continuous variable
- **Categorical variable:** no inherent order among the values For example: "What kind of pet you have" is a categorical variable

Common Issues

Common issues with data:

- **Missing values:** how do we fill in?
- **Wrong values:** how can we detect and correct?
- **Messy format**
- **Not usable:** the data cannot answer the question posed

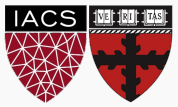
Messy Data

The following is a table accounting for the number of produce deliveries over a weekend.

What are the variables in this dataset? What object or event are we measuring?

	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

What's the issue? How do we fix it?



Messy Data

We're measuring individual deliveries; the variables are Time, Day, Number of Produce.

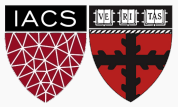
	Friday	Saturday	Sunday
Morning	15	158	10
Afternoon	2	90	20
Evening	55	12	45

Problem: each column header represents a single value rather than a variable. Row headers are "hiding" the Day variable. The values of the variable, "Number of Produce", is not recorded in a single column.

Fixing Messy Data

We need to reorganize the information to make explicit the event we're observing and the variables associated to this event.

ID	Time	Day	Number
1	Morning	Friday	15
2	Morning	Saturday	158
3	Morning	Sunday	10
4	Afternoon	Friday	2
5	Afternoon	Saturday	9
6	Afternoon	Sunday	20
7	Evening	Friday	55
8	Evening	Saturday	12
9	Evening	Sunday	45



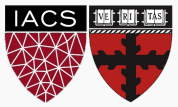
Tabular = Happy Pavlos ☺

Common causes of messiness are:

- Column headers are values, not variable names
- Variables are stored in both rows and columns
- Multiple variables are stored in one column/entry
- Multiple types of experimental units stored in same table

In general, we want each file to correspond to a dataset, each column to represent a single variable and each row to represent a single observation. We want to tabularize the data.

This makes Python happy.



Data Exploration: Descriptive Statistics

Basics of Sampling

Population versus sample:

- A **population** is the entire set of objects or events under study. Population can be hypothetical "all students" or all students in this class.
- A **sample** is a "representative" subset of the objects or events under study. Needed because it's impossible or intractable to obtain or compute with population data.

Biases in samples:

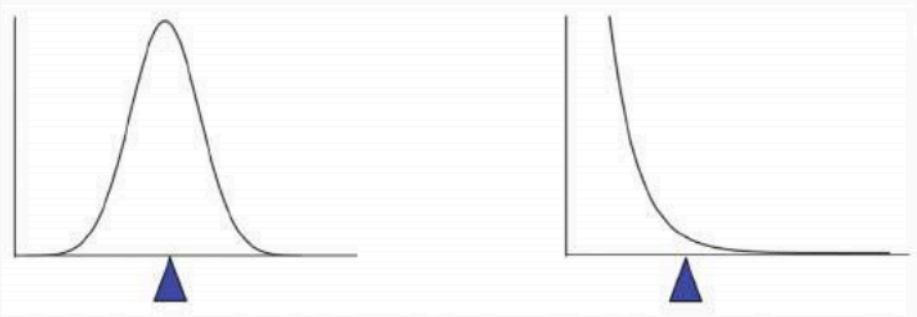
- **Selection bias**: some subjects or records are more likely to be selected
- Volunteer/**nonresponse bias**: subjects or records who are not easily available are not represented

Examples?

Sample mean

The **mean** of a set of *n* observations of a variable is denoted $\bar{x}$ and is defined as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



The mean describes what a "typical" sample value looks like, or where is the "center" of the distribution of the data.

Key theme: there is always uncertainty involved when calculating a sample mean to estimate a population mean.

Sample median

The **median** of a set of n number of observations in a sample, ordered by value, of a variable is is defined by

$$\text{Median} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2} + x_{(n+1)/2}}{2} & \text{if } n \text{ is even} \end{cases}$$

Example (already in order):

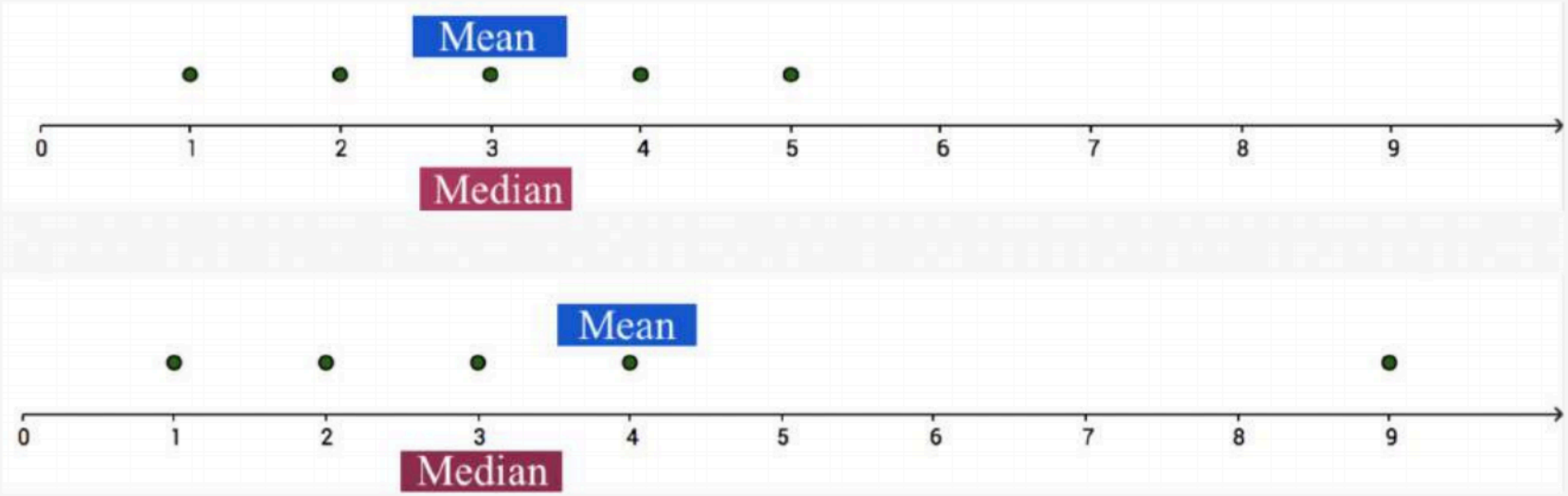
Ages: 17, 19, 21, 22, 23, 23, 23, 38

Median = (22+23)/2 = 22.5

The median also describes what a typical observation looks like, or where is the center of the distribution of the sample of observations.

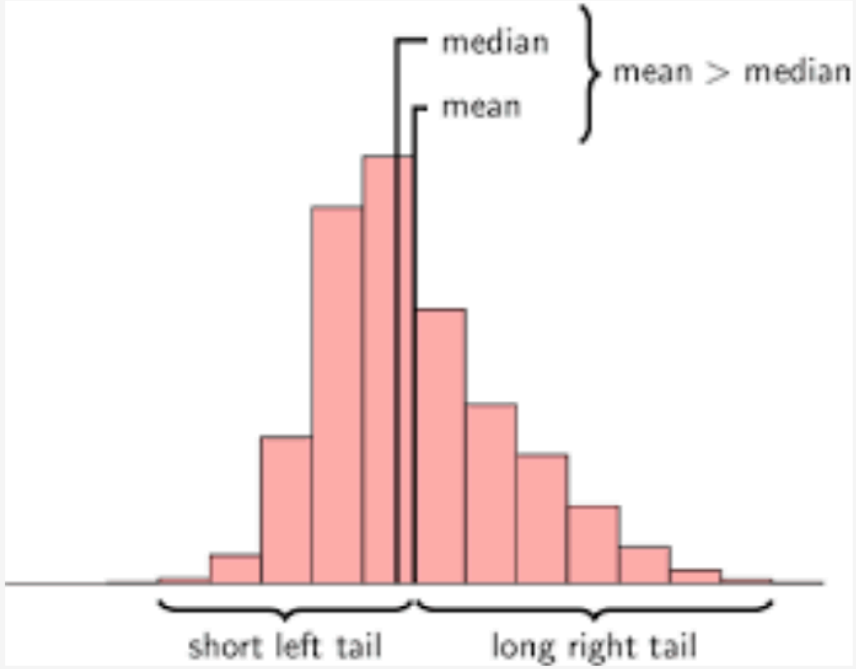
Mean vs. Median

The mean is sensitive to extreme values (**outliers**)



Mean, median, and skewness

The mean is sensitive to outliers\.



The above distribution is called **right-skewed** since the mean is greater than the median. Note: **skewness** often “follows the longer tail”.

Computational time

How hard (in terms of algorithmic complexity) is it to calculate

- the mean?

  at most $O(n)$

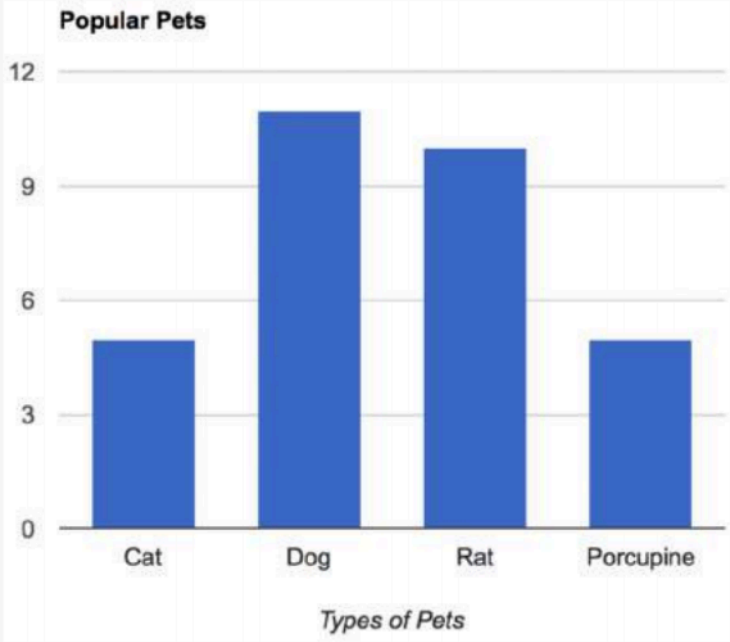
- the median?

  at most $O(n)$

Note: Practicality of implementation should be considered!

Regarding Categorical Variables…

For categorical variables, neither mean or median make sense. Why?



The mode might be a better way to find the most “representative” value

Measures of Spread: Range

The spread of a sample of observations measures how well the mean or median describes the sample.

One way to measure spread of a sample of observations is via the **range**.

Range = Maximum Value - Minimum Value

Measures of Spread: Variance

The (sample) **variance**, denoted s^2, measures how much on average the sample values deviate from the mean:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2$$

Note: the term $|x_i - \bar{x}|$ measures the amount by which each x_i deviates from the mean $\bar{x}$. Squaring these deviations means that s^2 is sensitive to extreme values (outliers).

Note: s^2 doesn't have the same units as the x_i :(

What does a variance of 1,008 mean? Or 0.0001?

Measures of Spread: Standard Deviation

The (sample) **standard deviation**, denoted s, is the square root of the variance

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} |x_i - \bar{x}|^2}$$

Note: s does have the same units as the x_i. Phew!

Modeling

Data

The **Advertising data set** consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. Everything is given in units of $1000.

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani "

Response vs. Predictor Variables

There is an asymmetry in many of these problems:

The variable we'd like to predict may be more difficult to measure, is more important than the other(s), or may be directly or indirectly influenced by the values of the other variable(s).

Thus, we'd like to define two categories of variables:

- variables whose value we want to predict
- variables whose values we use to make our prediction.

Response vs. Predictor Variables

X
predictors
features
covariates

Y
outcome
response variable
dependent variable

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

n observations

p predictors

Response vs. Predictor Variables

$X = X_1, \dots, X_p$

$X_j = x_{1j}, \dots, x_{ij}, \dots, x_{nj}$

predictors

features

covariates

$Y = y_1, \dots, y_n$

outcome

response variable

dependent variable

TV	radio	newspaper	sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9

n observations

p predictors

Definition

We are observing $p + 1$ number variables and we are making n sets of observations. We call:

- the variable we'd like to predict the **outcome** or **response variable**; typically, we denote this variable by Y and the individual measurements y_i.
- the variables we use in making the predictions the **features** or **predictor** variables; typically, we denote these variables by $X = X_1, \dots, X_p$ and the individual measurements $x_{i,j}$.

Note: i indexes the observation ($i = 1, \dots, n$) and j indexes the value of the j-th predictor variable (j $= 1, \dots, p$).

Statistical Model

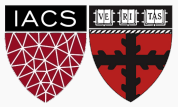
True vs. Statistical Model

We will assume that the response variable, Y, relates to the predictors, X, through some unknown function expressed generally as:

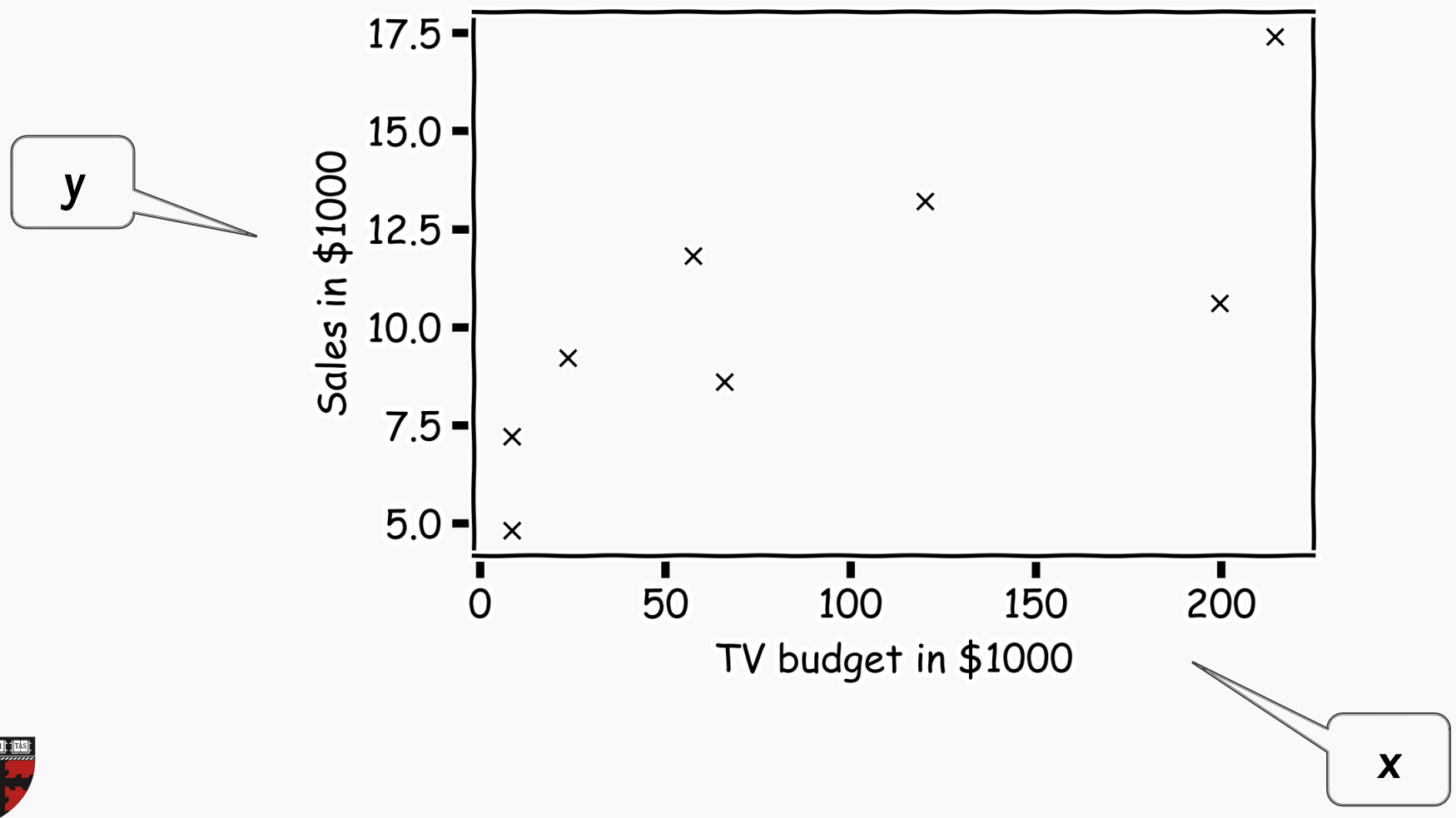
$$Y = f(X) + \varepsilon$$

Here, f is the unknown function expressing an underlying rule for relating Y to X, ε is the random amount (unrelated to X) that Y differs from the rule $f(X)$.

A ***statistical model*** is any algorithm that estimates f. We denote the estimated function as $\widehat{f}$.

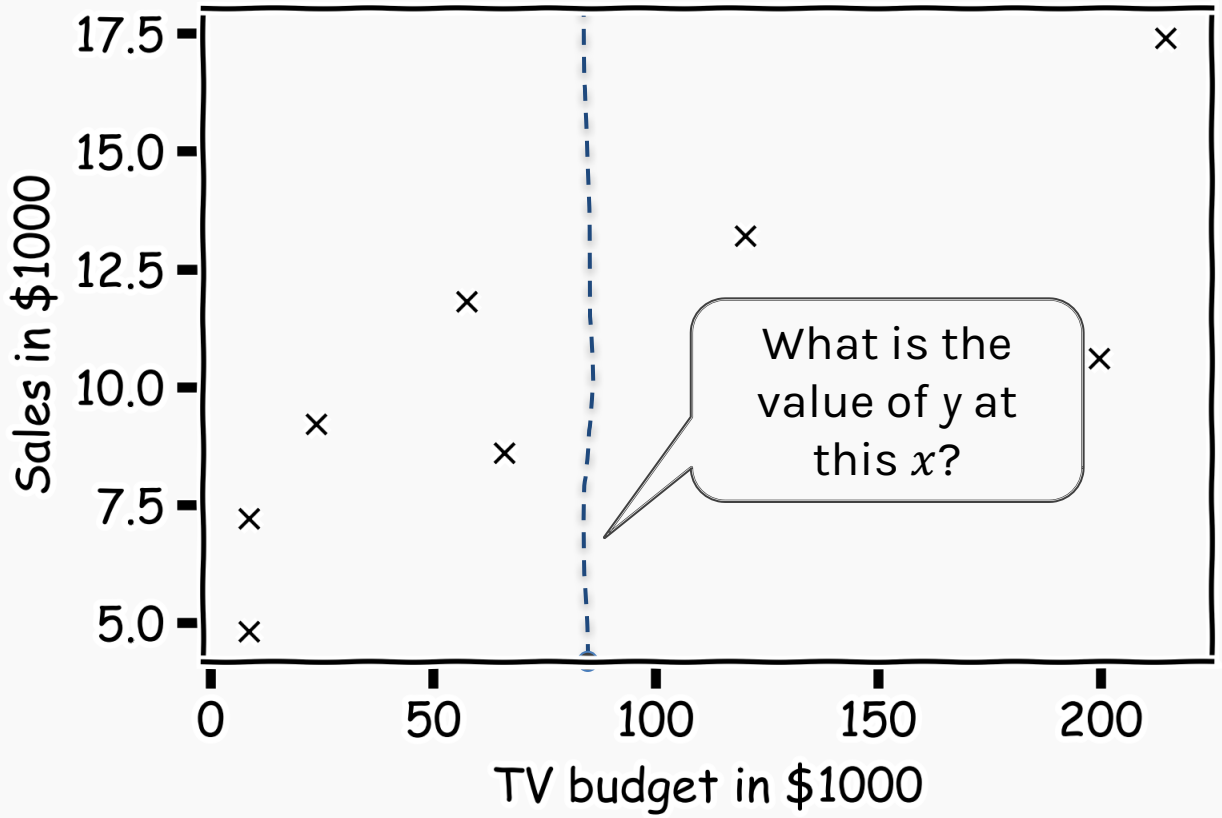


Statistical Model



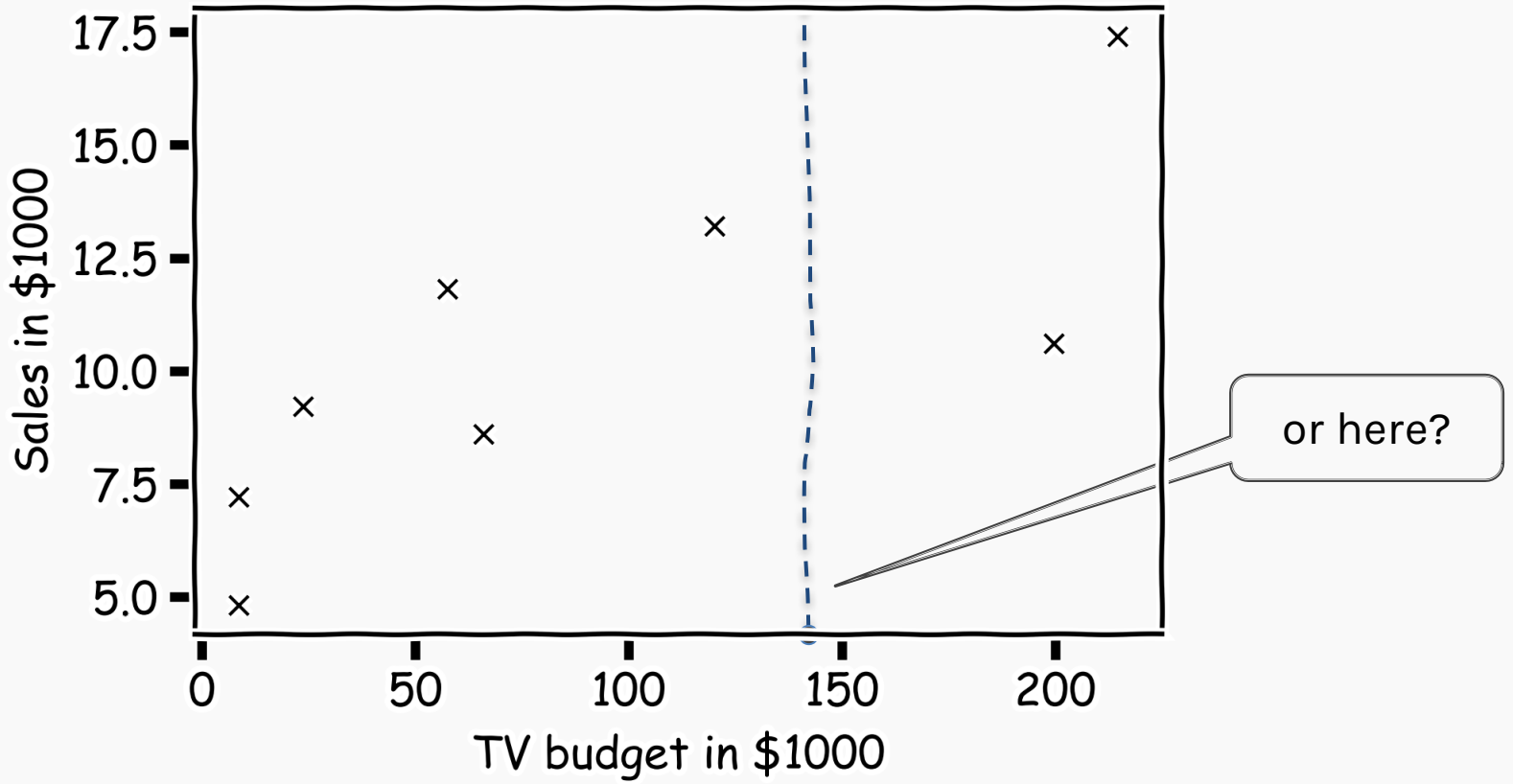
Statistical Model

How do we find $\hat{f}(x)$?



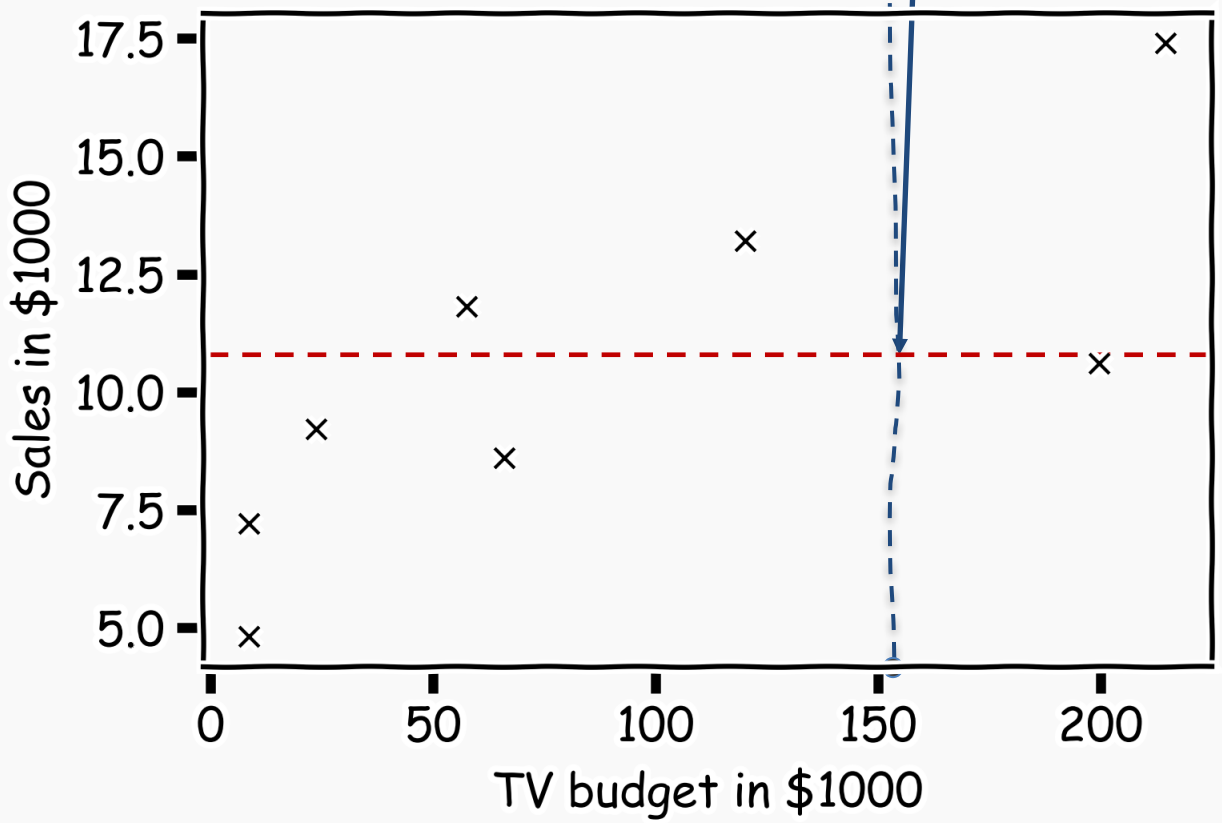
Statistical Model

How do we find $\hat{f}(x)$?



Statistical Model

Simple idea is to take the mean of all y's, $\hat{f}(x) = \frac{1}{n}\sum_1^n y_i$



Prediction vs. Estimation

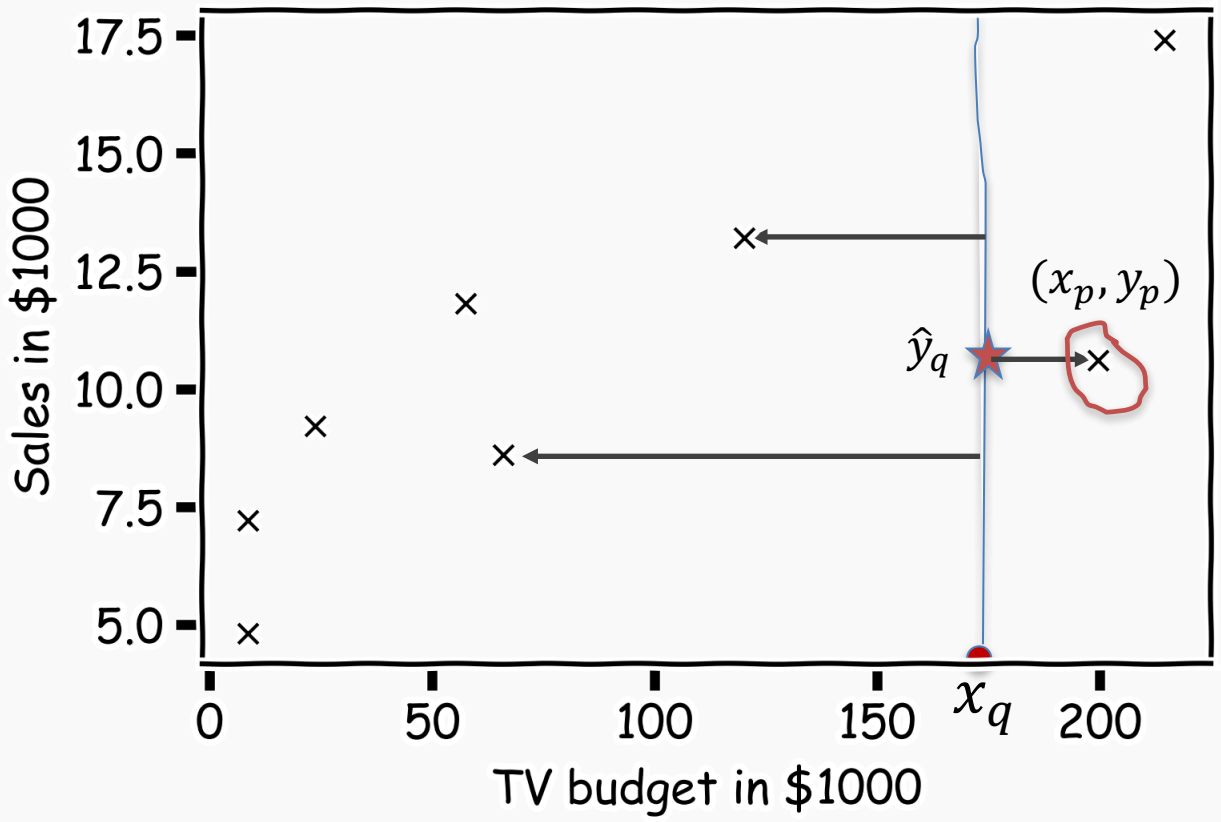
For some problems, what's important is obtaining $\hat{f}$, our estimate of f. These are called ***inference*** problems.

When we use a set of measurements, $(x_{i,1}, \dots, x_{i,p})$ to predict a value for the response variable, we denote the ***predicted*** value by:

$$\hat{y}_i = \hat{f}(x_{i,1}, \dots, x_{i,p}).$$

For some problems, we don't care about the specific form of $\hat{f}$, we just want to make our prediction $\hat{y}$ as close to the observed value y as possible. These are called ***prediction problems.***

Simple Prediction Model



What is $\hat{y}_q$ at some x_q ?

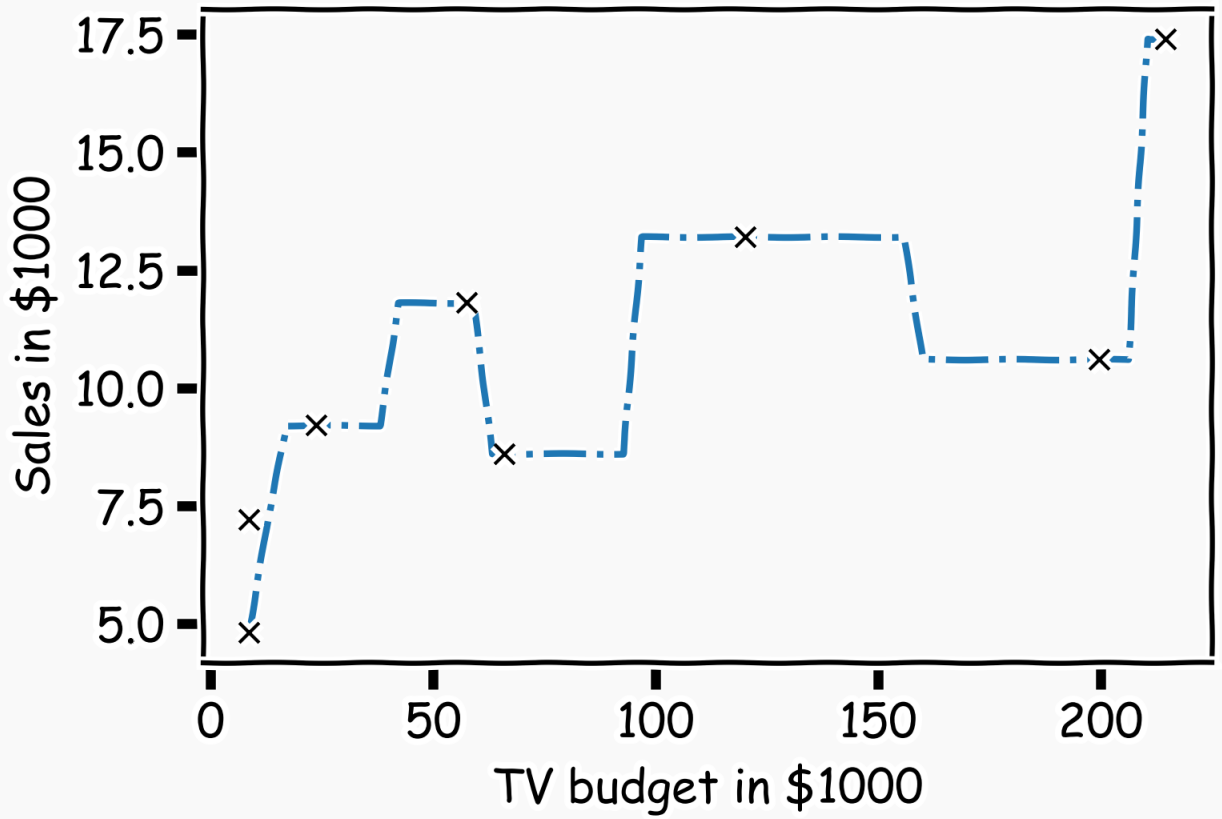
Find distances to all other points $D(x_q, x_i)$

Find the nearest neighbor, (x_p, y_p)

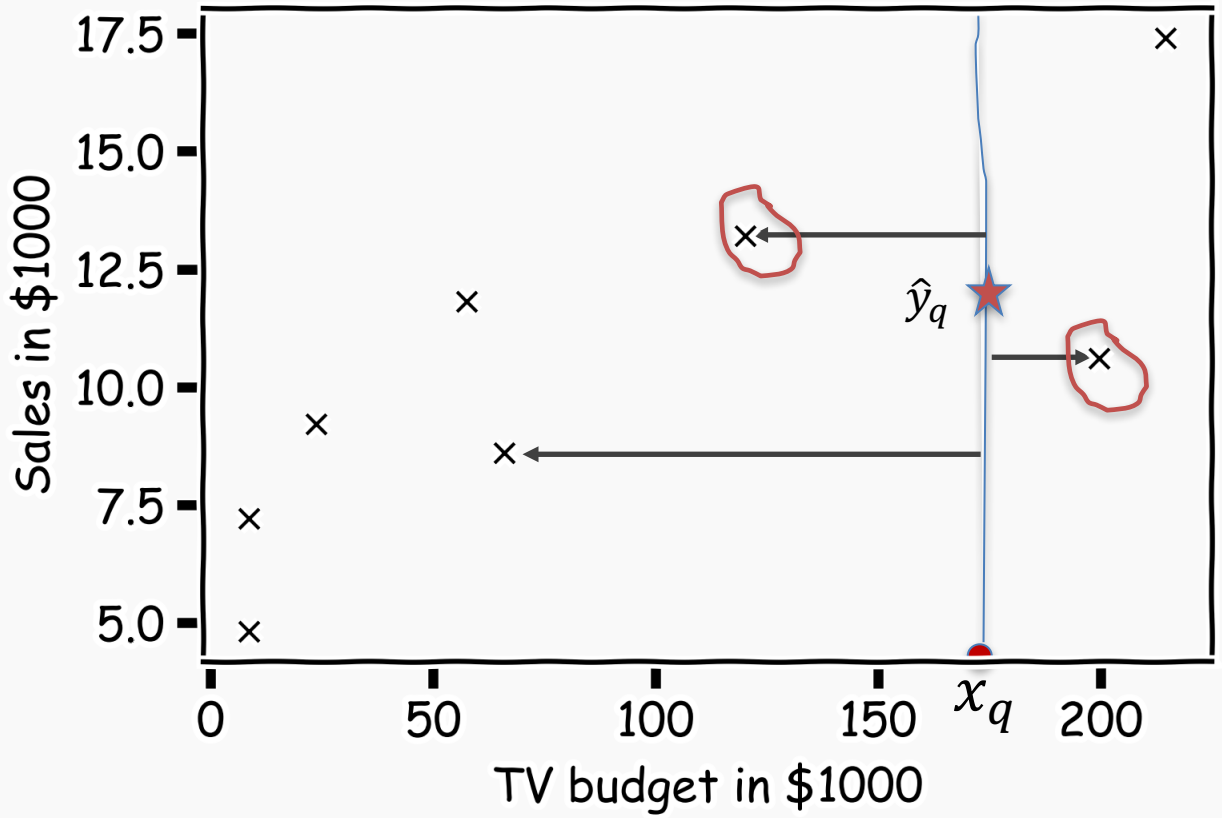
Predict $\hat{y}_q = y_p$

Simple Prediction Model

Do the same for "all" $x's$



Extend the Prediction Model



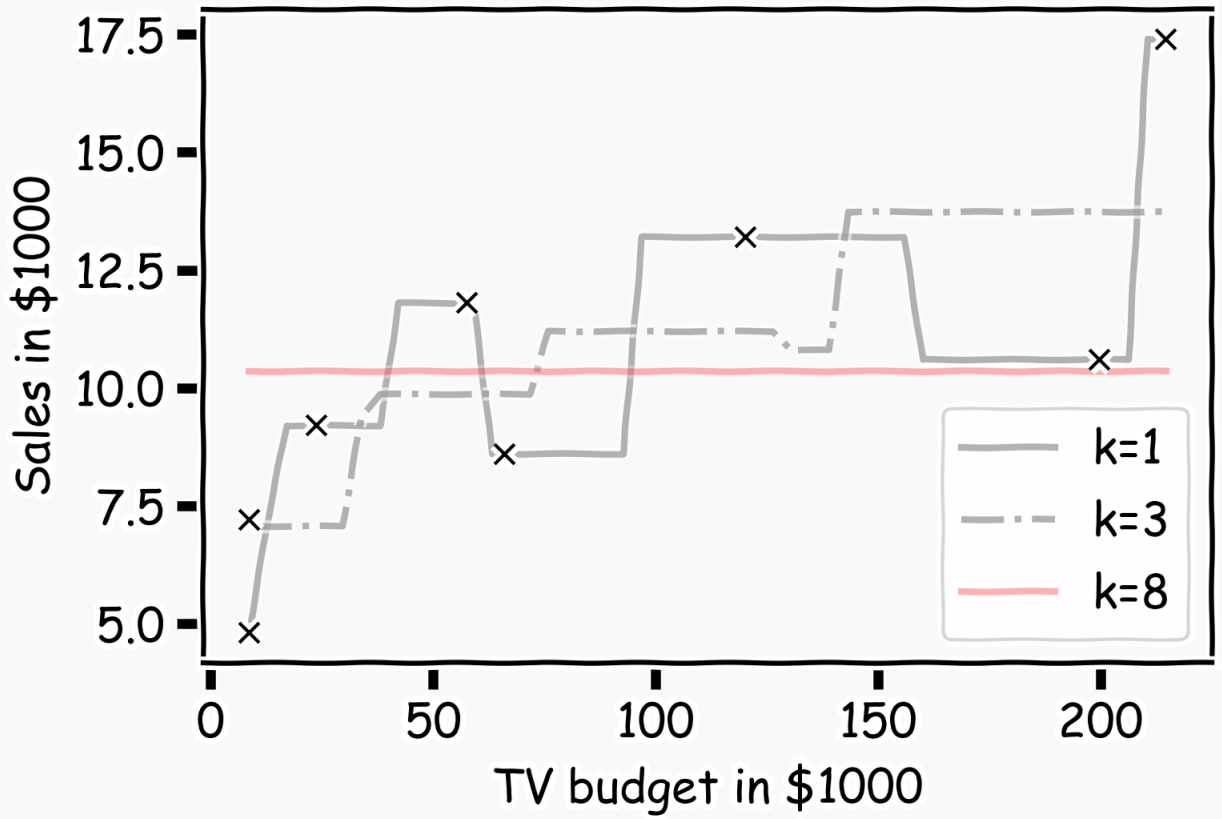
What is $\hat{y}_q$ at some x_q ?

Find distances to all other points $D(x_q, x_i)$

Find the k-nearest neighbors, $x_{q_1}, \ldots, x_{q_k}$

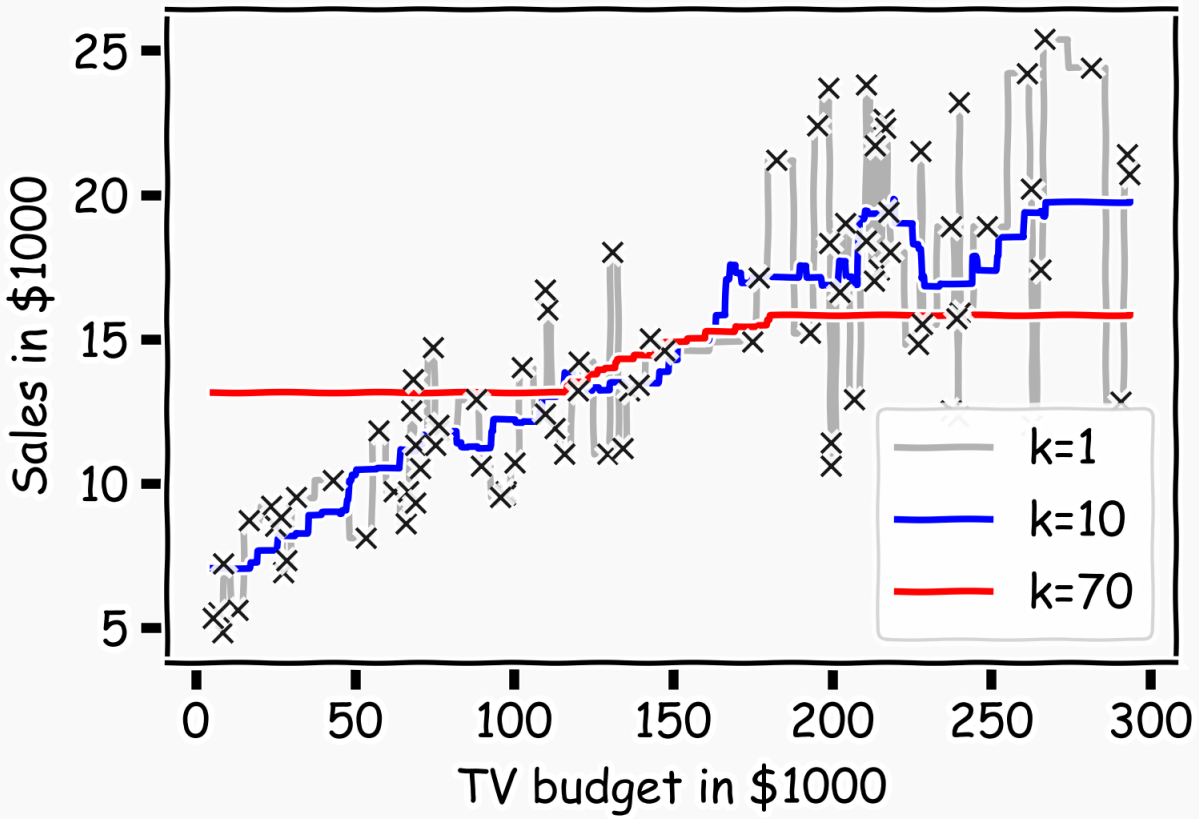
Predict $\widehat{y_q} = \frac{1}{k}\sum_i^k y_{q_i}$

Simple Prediction Models



Simple Prediction Models

Same models on more data



k-Nearest Neighbors

The ***k-Nearest Neighbor (kNN) model*** is an intuitive way to predict a quantitative response variable:

> *to predict a response for a set of observed predictor values, we use the responses of other observations most similar to it*

Note: this strategy can also be applied in classification to predict a categorical variable. We will encounter kNN again later in the course in the context of classification.

k-Nearest Neighbors - kNN

For a fixed a value of *k*, the predicted response for the i-th observation is the average of the observed response of the *k*-closest observations:

$$\hat{y}_n = \frac{1}{k} \sum_{i=1}^{k} y_{n_i}$$

where $\{x_{n1}, \ldots, x_{nk}\}$ are the *k* observations most similar to x_i (*similar* refers to a notion of distance between predictors).

Things to Consider

Model Fitness

How does the model perform predicting?

Comparison of Two Models

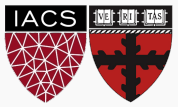
How do we choose from two different models?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

How well we know $\hat{f}$

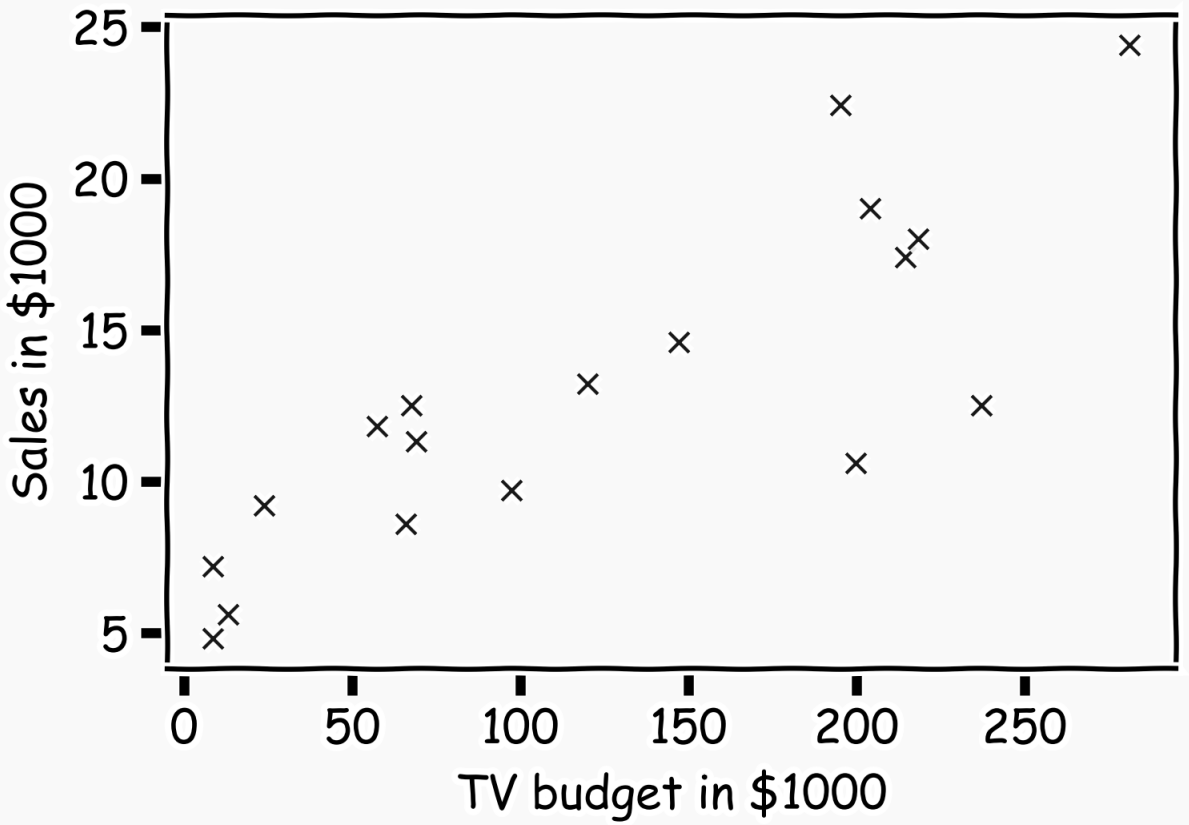
The confidence intervals of our $\hat{f}$



Error Evaluation

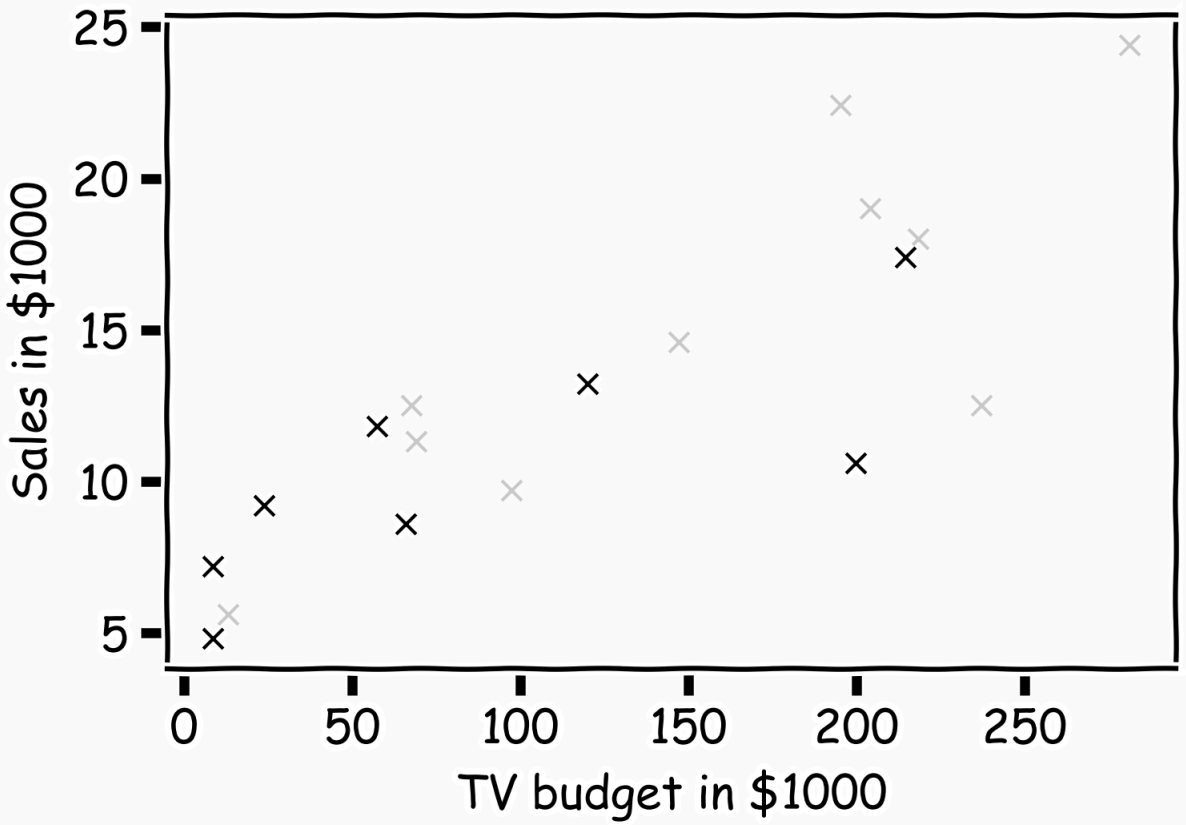
Error Evaluation

Start with some data.



Error Evaluation

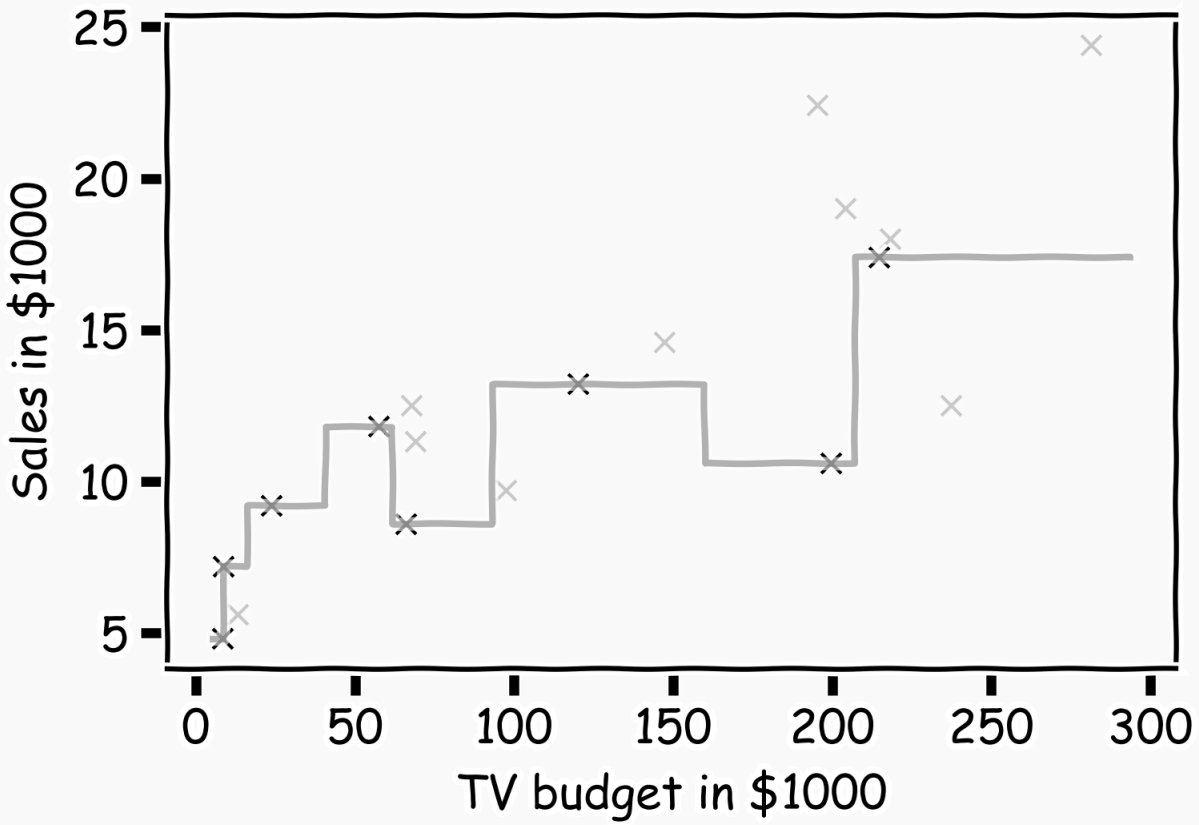
Hide some of the data from the model. This is called **train-test** split.



We use the train set to estimate $\hat{y}$, and the test set to evaluate the model.

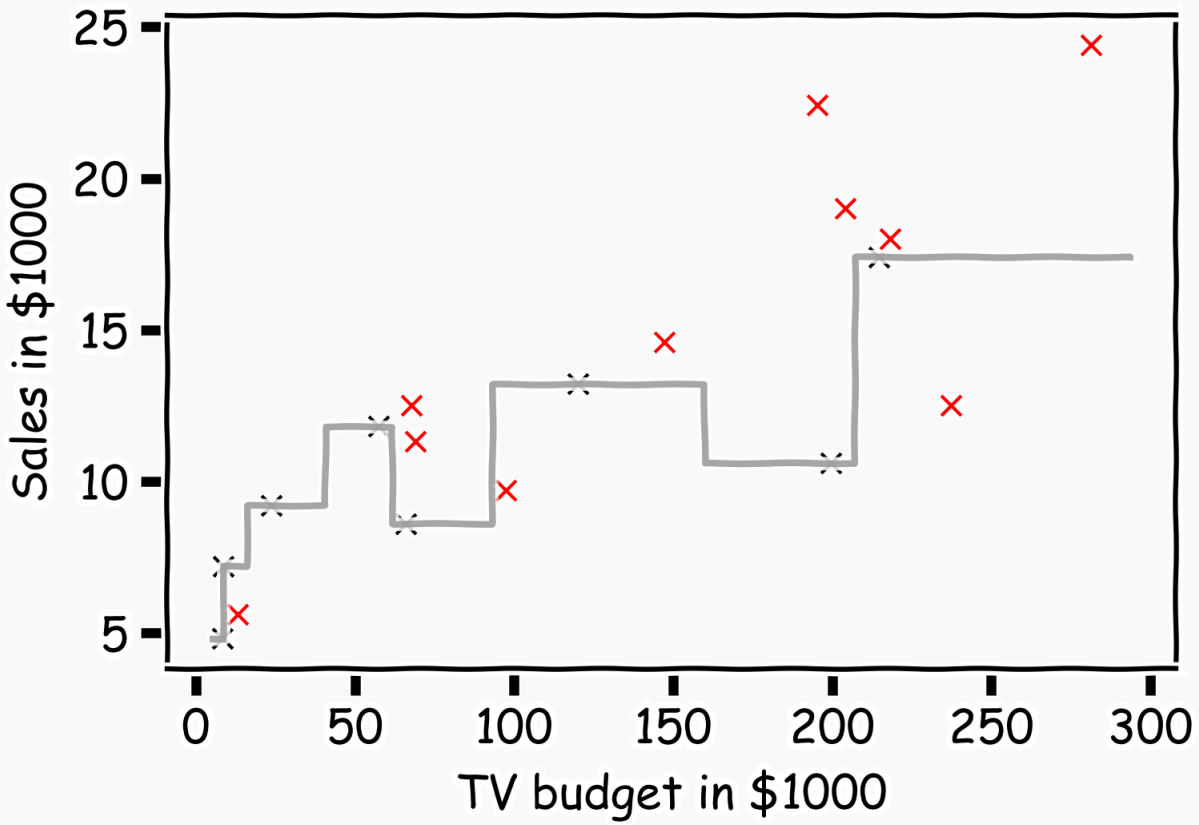
Error Evaluation

Estimate $\hat{y}$ for $k=1$.



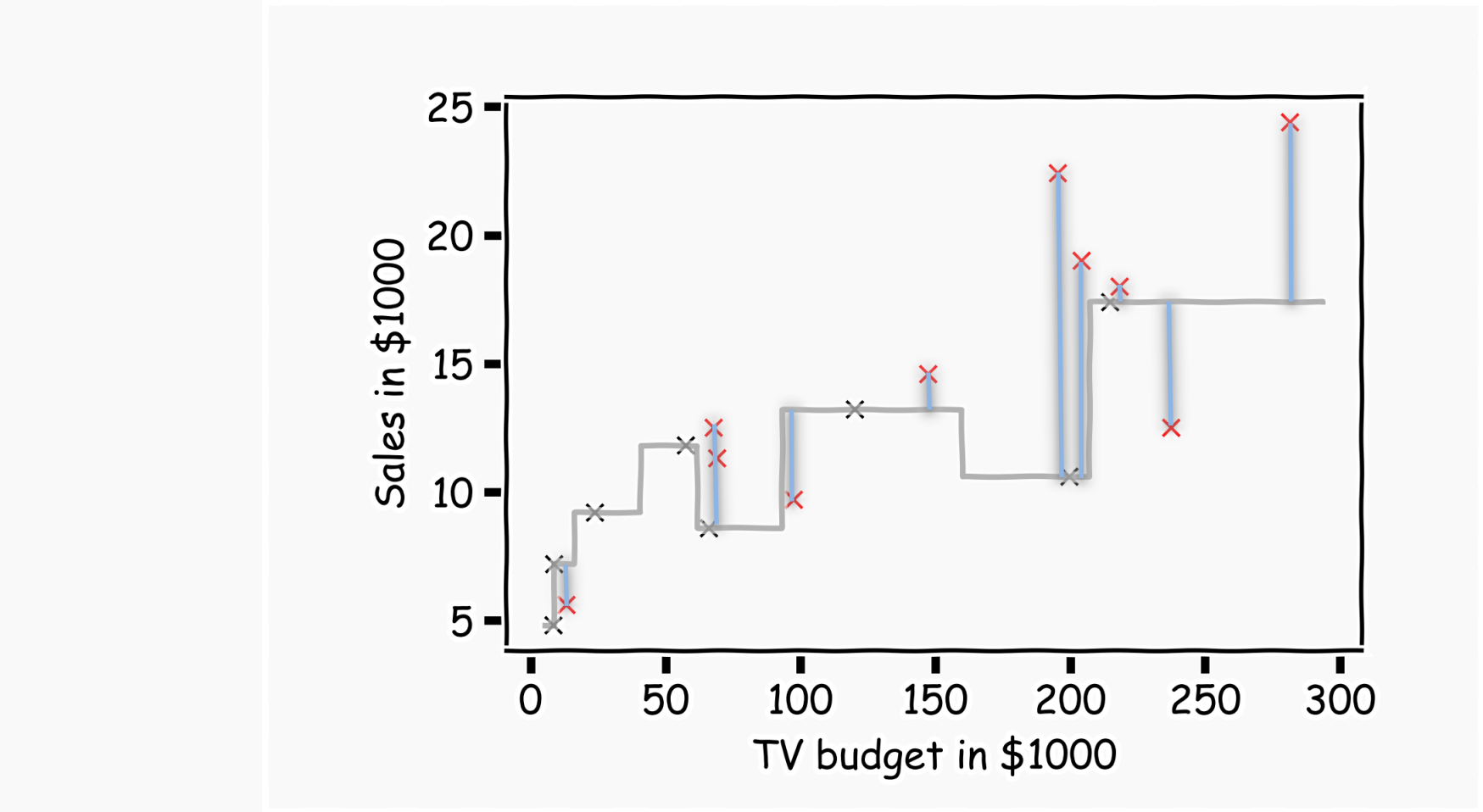
Error Evaluation

Now, we look at the data we have not used, the **test data** (red crosses).



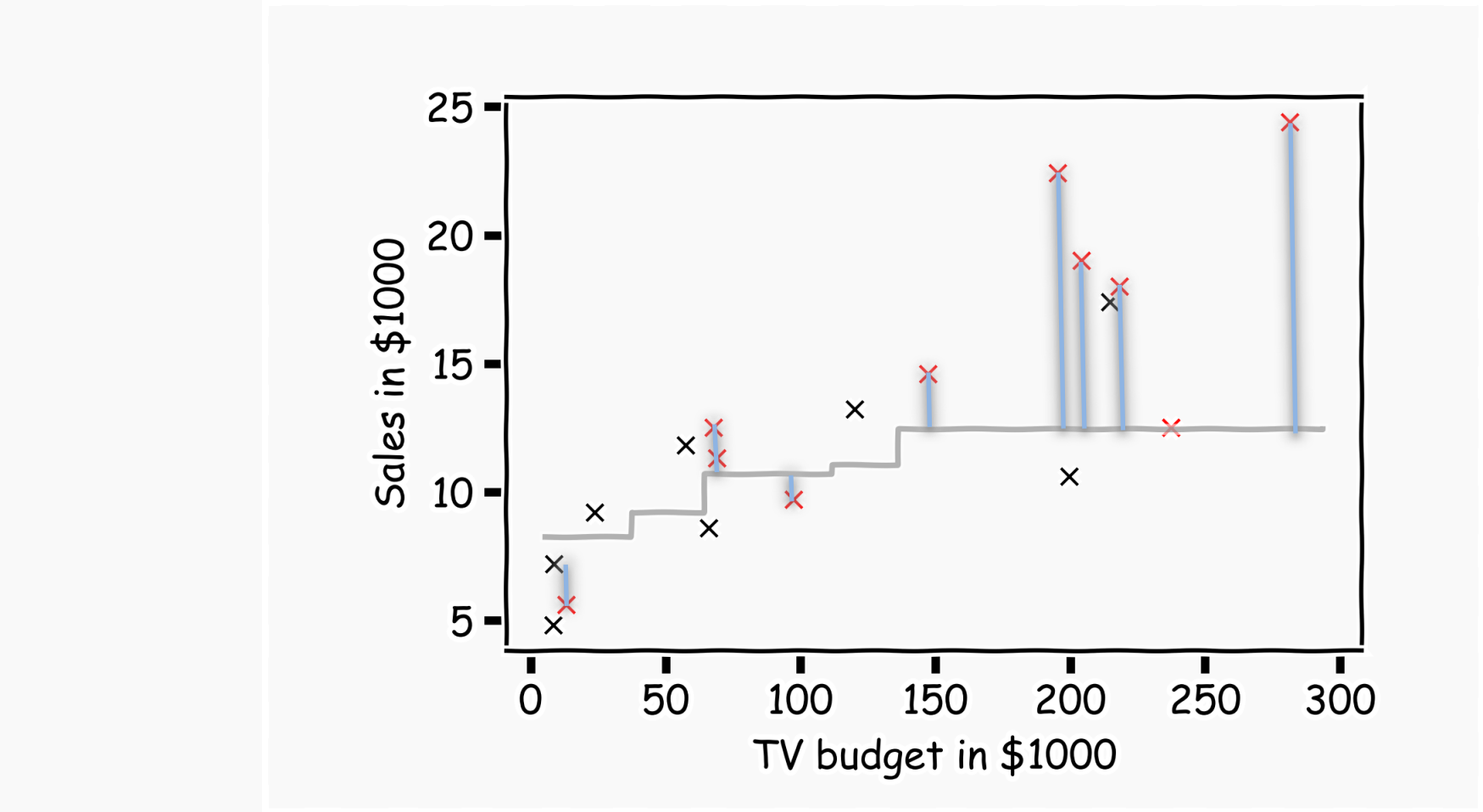
Error Evaluation

Calculate the **residuals** $(y_i - \hat{y}_i)$.



Error Evaluation

Do the same for k=3.



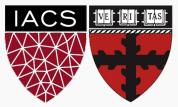
Error Evaluation

In order to quantify how well a model performs, we define a ***loss*** or ***error function.***

A common loss function for quantitative outcomes is the **Mean Squared Error (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$$

The quantity $y_i - \hat{y}_i$ is called a ***residual*** and measures the error at the *i*-th prediction.



Error Evaluation

Caution: The MSE is by no means the only valid (or the best) loss function!

Question: What would be an intuitive loss function for predicting categorical outcomes?

Note: The square **R**oot of the **M**ean of the **S**quared **E**rrors (RMSE) is also commonly used.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

Things to Consider

Comparison of Two Models

How do we choose from two different models?

Model Fitness

How does the model perform predicting?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

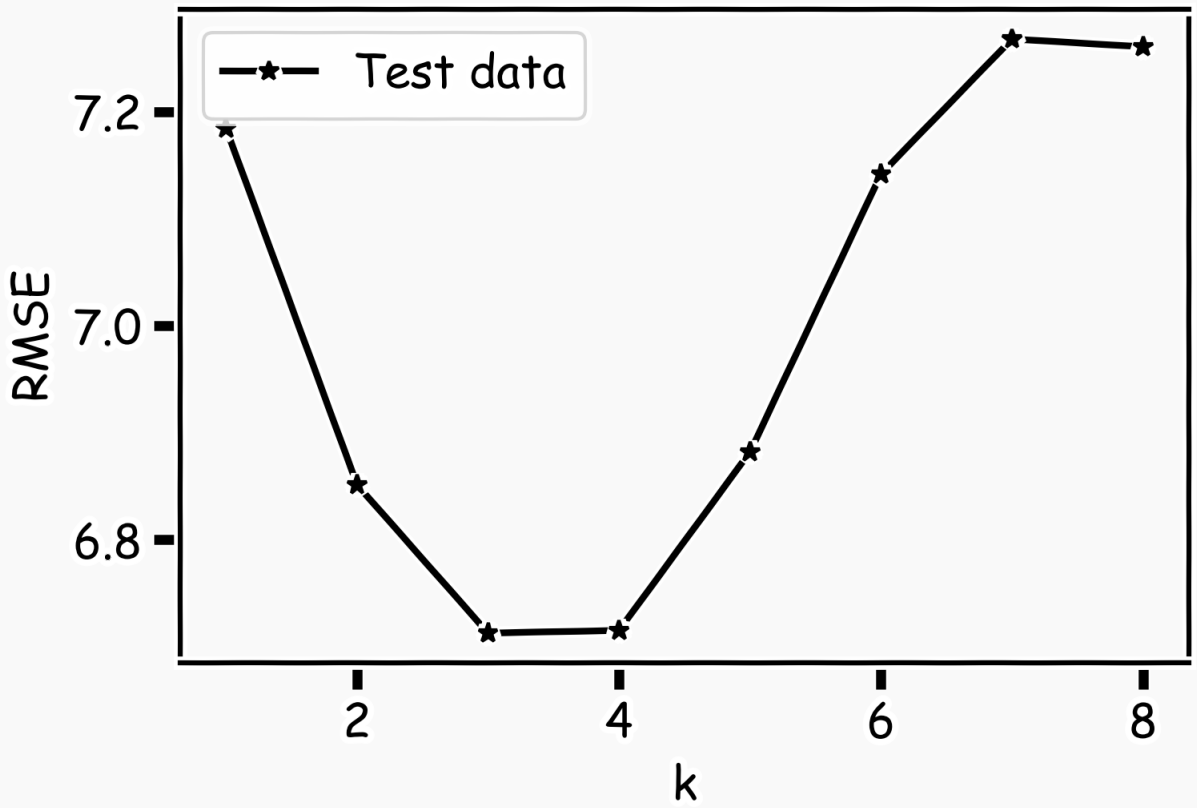
How well do we know $\hat{f}$

The confidence intervals of our $\hat{f}$

Model Comparison

Model Comparison

Do the same for all *k*'s and compare the RMSEs. *k*=3 seems to be the best model.



Things to Consider

Comparison of Two Models

How do we choose from two different models?

Model Fitness

How does the model perform predicting?

Evaluating Significance of Predictors

Does the outcome depend on the predictors?

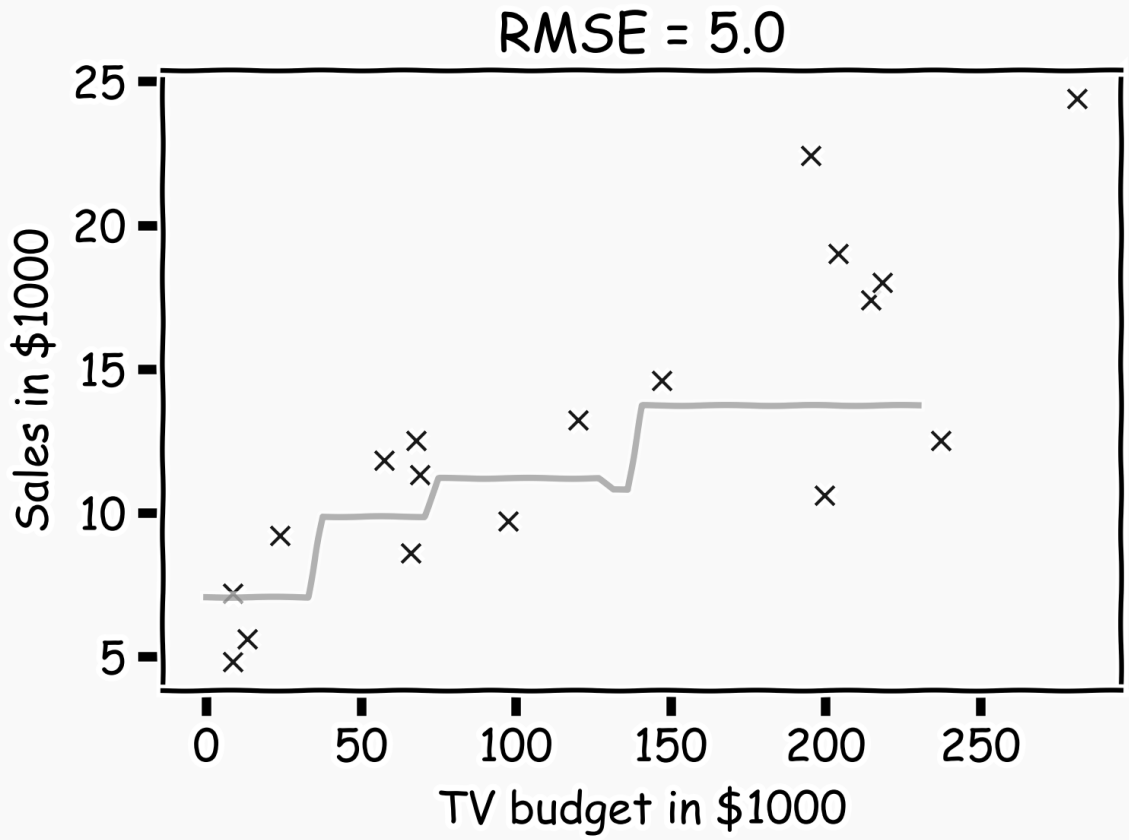
How well do we know $\hat{f}$

The confidence intervals of our $\hat{f}$

Model Fitness

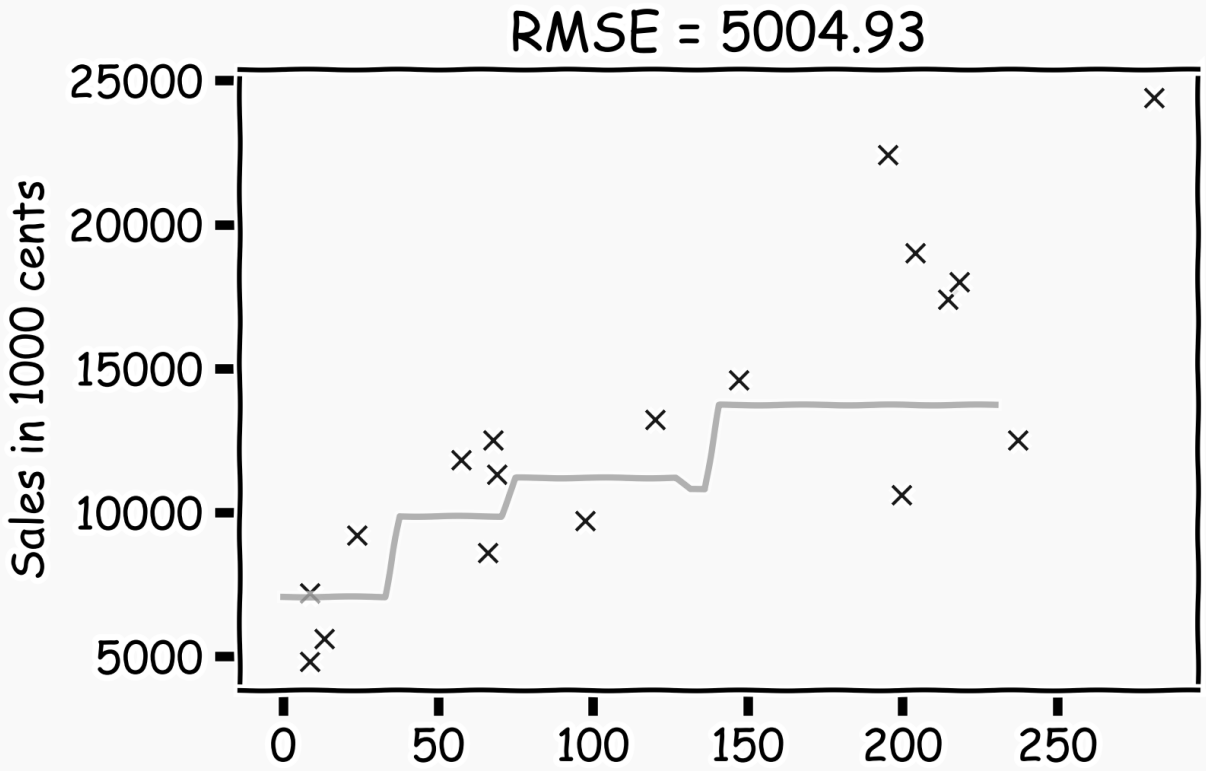
Model fitness

For a subset of the data, calculate the RMSE for k=3. Is RMSE=5.0 good enough?



Model fitness

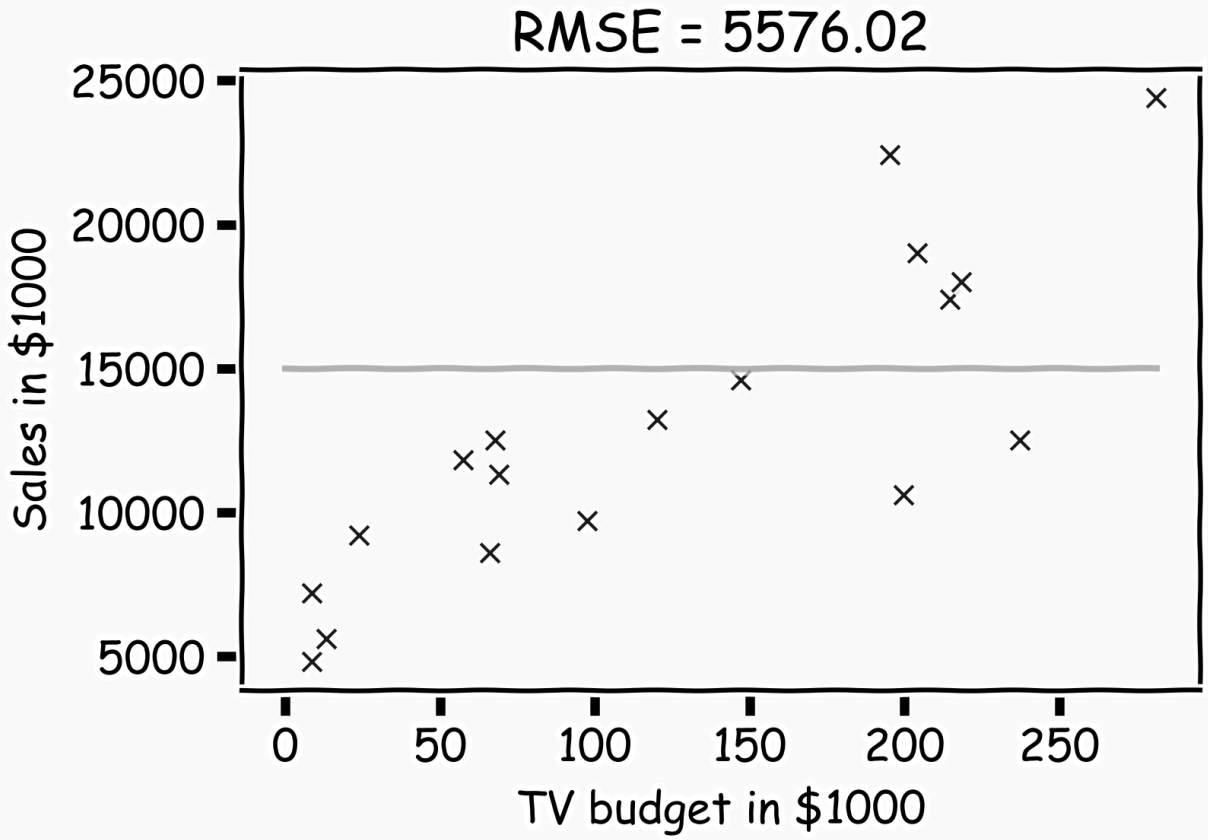
What if we measure the Sales in cents instead of dollars?



RMSE is now 5004.93.
Is that good?

Model fitness

It is better if we compare it to something.



We will use the simplest model:

$$\hat{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

R-squared

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2}$$

- If our model is as good as the mean value, $\bar{y}$, then $R^2 = 0$
- If our model is perfect then $R^2 = 1$
- R^2 can be negative if the model is worst than the average. This can happen when we evaluate the model in the test set.