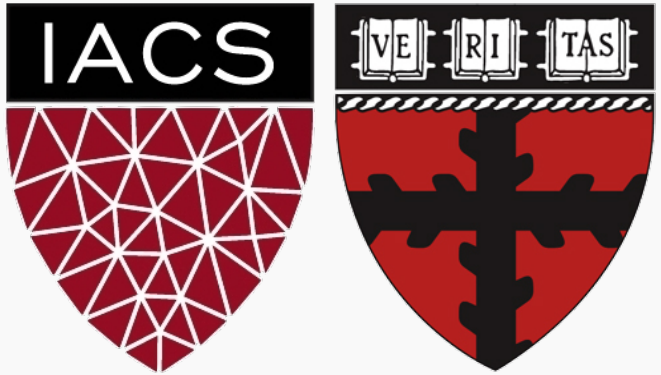
Lecture 4: Introduction to Regression

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner



Background

Roadmap:

Lecture 1	What is Data Science?
Lecture 2	Data: types, formats, issues, etc, and briefly visualization
Lecture 3 and Lab2	How to quickly prepare data and scrape the web
This lecture	**How to model data and evaluate model fitness.**
Next 3 lectures	Linear regression, confidence intervals, model selection cross validation, regularization

Lecture Outline

Statistical Modeling

k-Nearest Neighbors (kNN)

Model Fitness

How does the model perform predicting?

Comparison of Two Models

How do we choose from two different models?

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

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

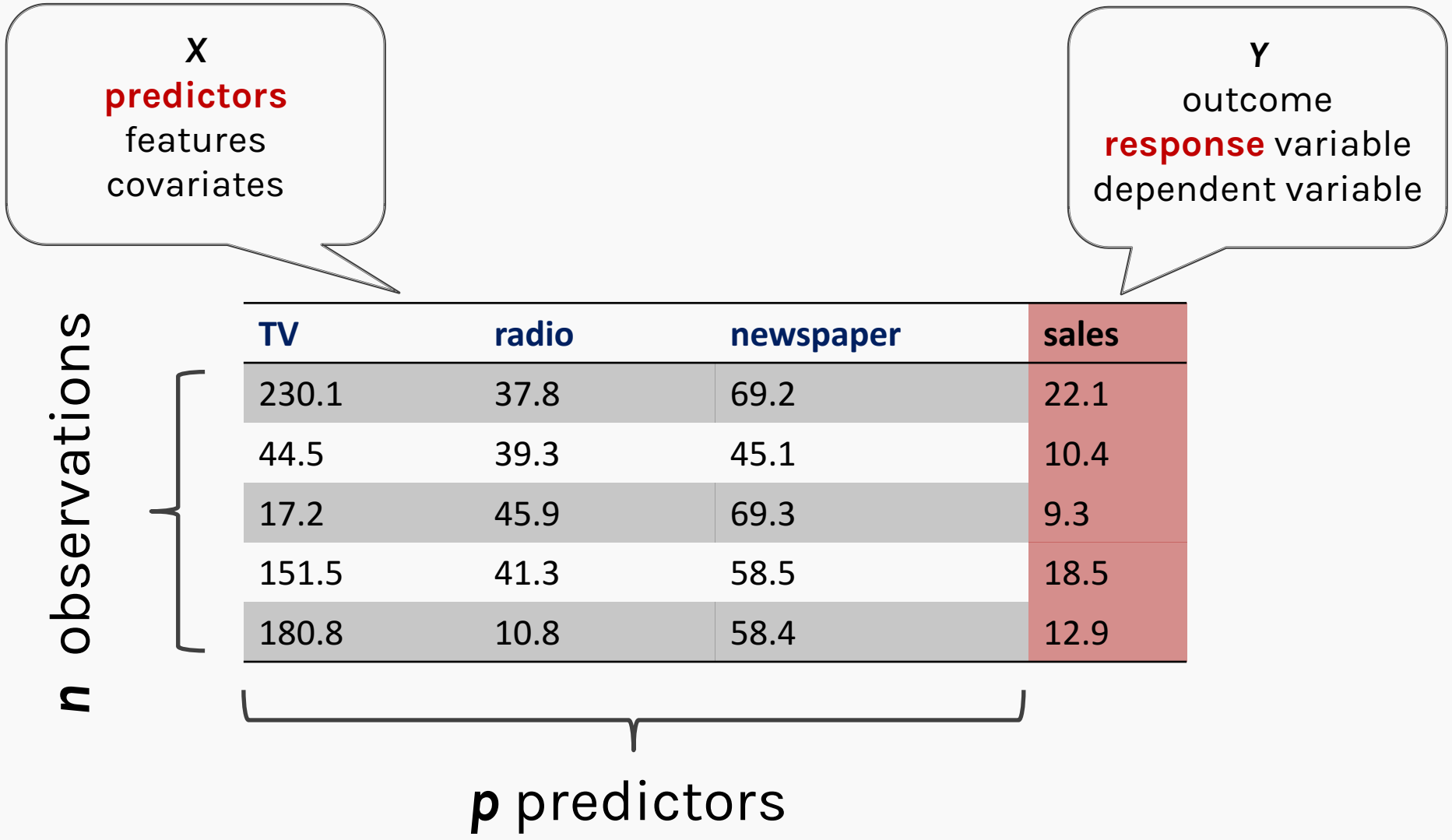
There is an asymmetry in many of these problems:

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



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

Response vs. Predictor Variables

$X = X_1, \ldots, X_p$

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

predictors

features

covariates

$Y = y_1, \ldots, y_n$

outcome

response variable

dependent variable

n observations

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

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 $X = X_1, \ldots, X_p$ and the individual measurements $x_{i,j}$.

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

Statistical Model

True vs. Statistical Model

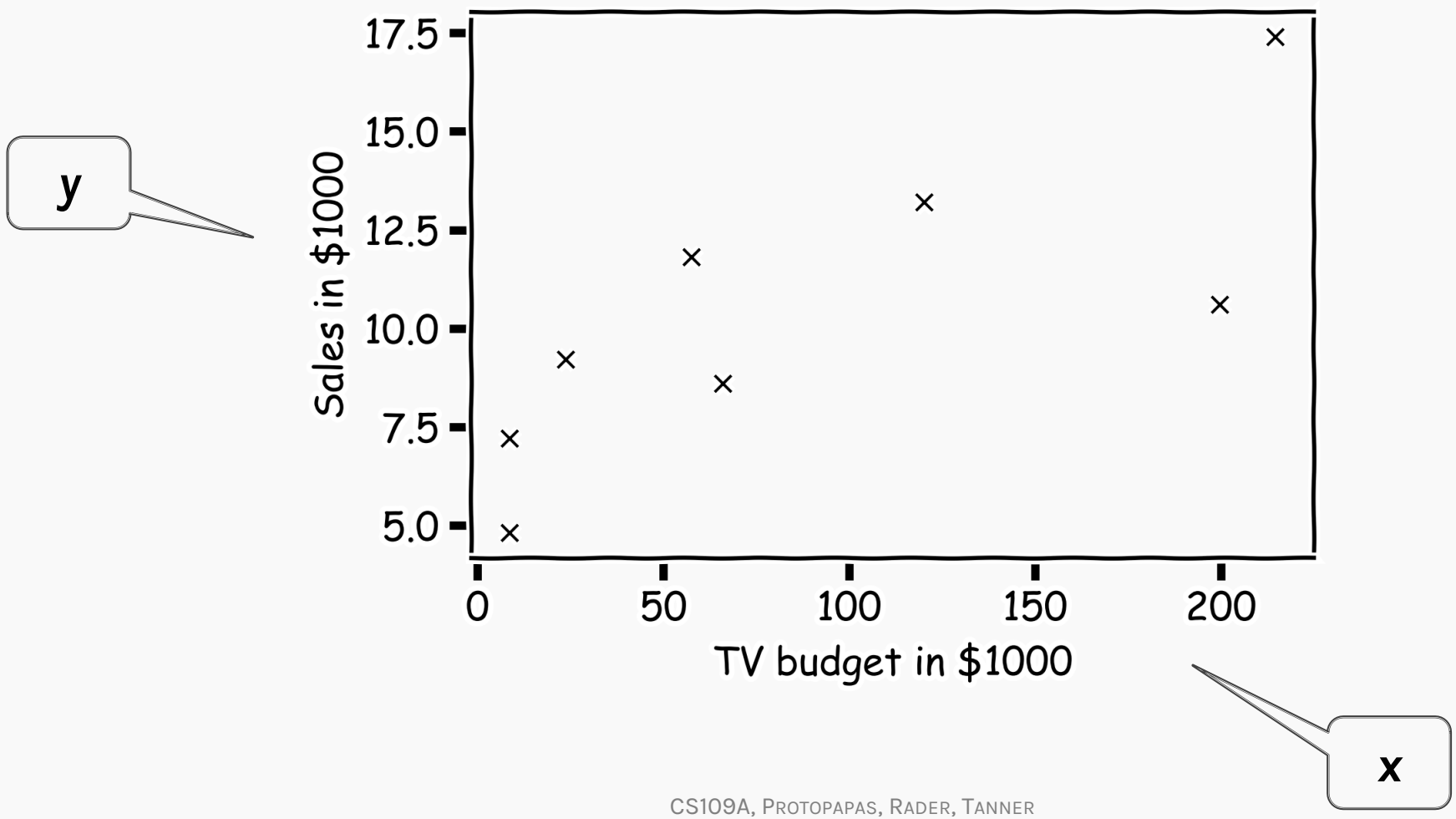
We will assume that the response variable, Y, relates to the predictors, X, through some unknown function expressed generally as:

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

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 $f(X)$.

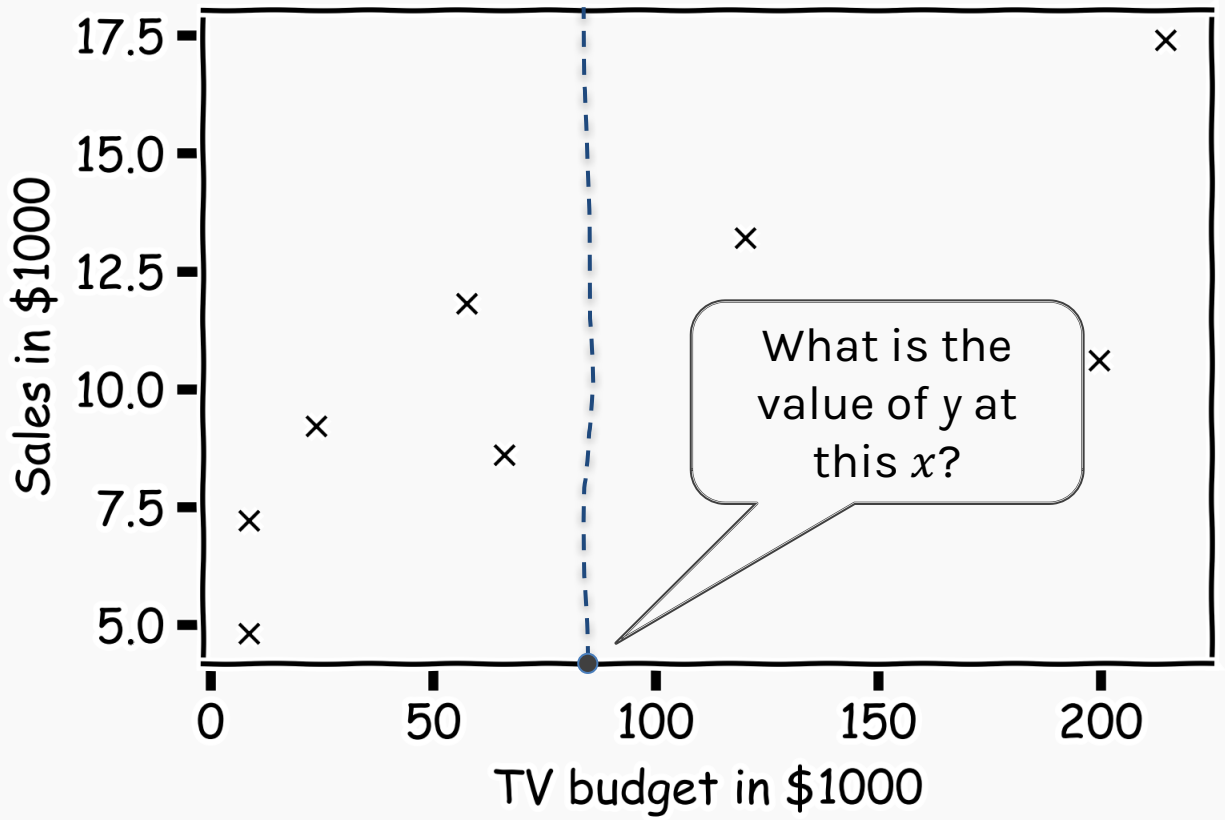
A ***statistical model*** is any algorithm that estimates f. We denote the estimated function as $\widehat{f}$.

Statistical Model



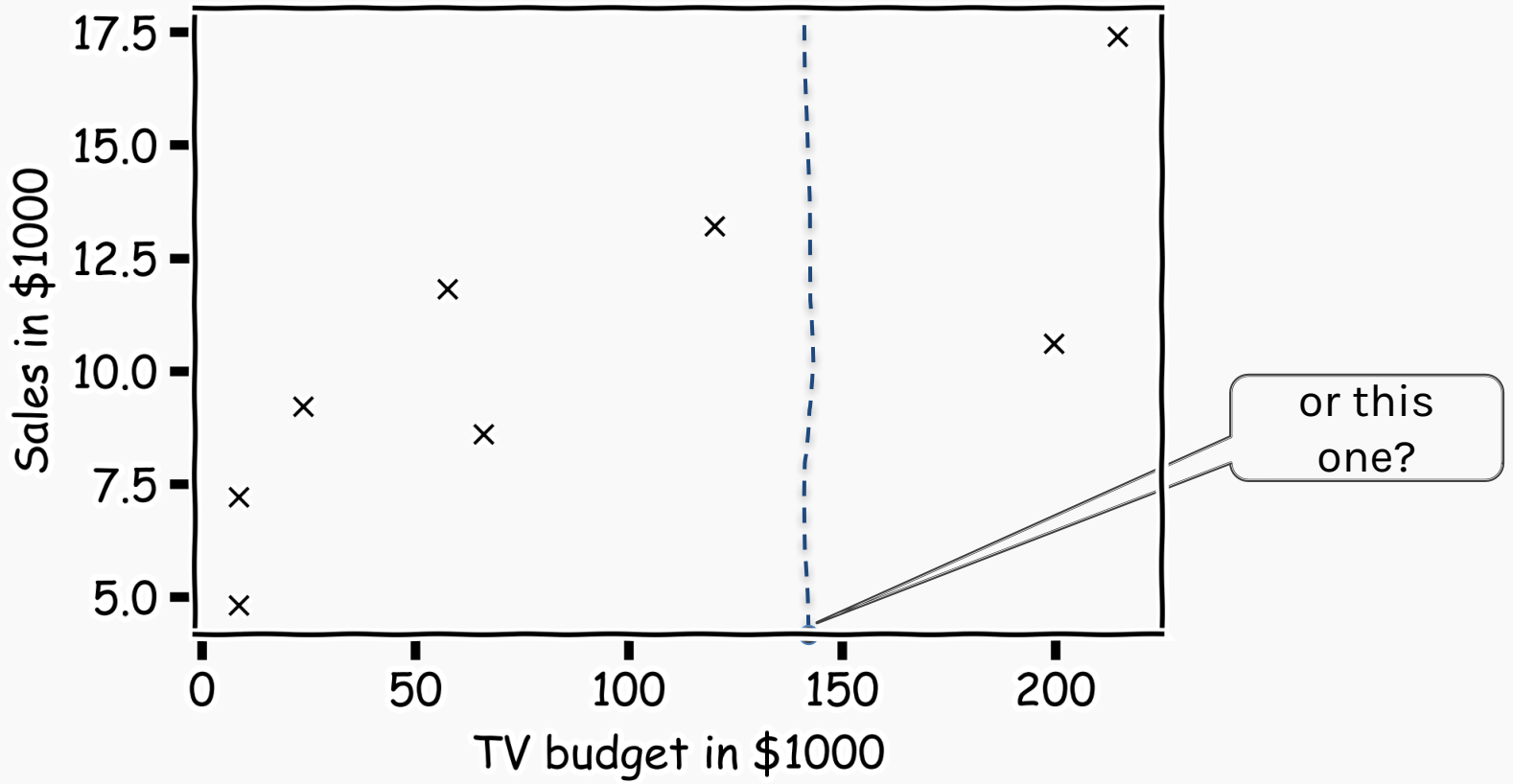
Statistical Model

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



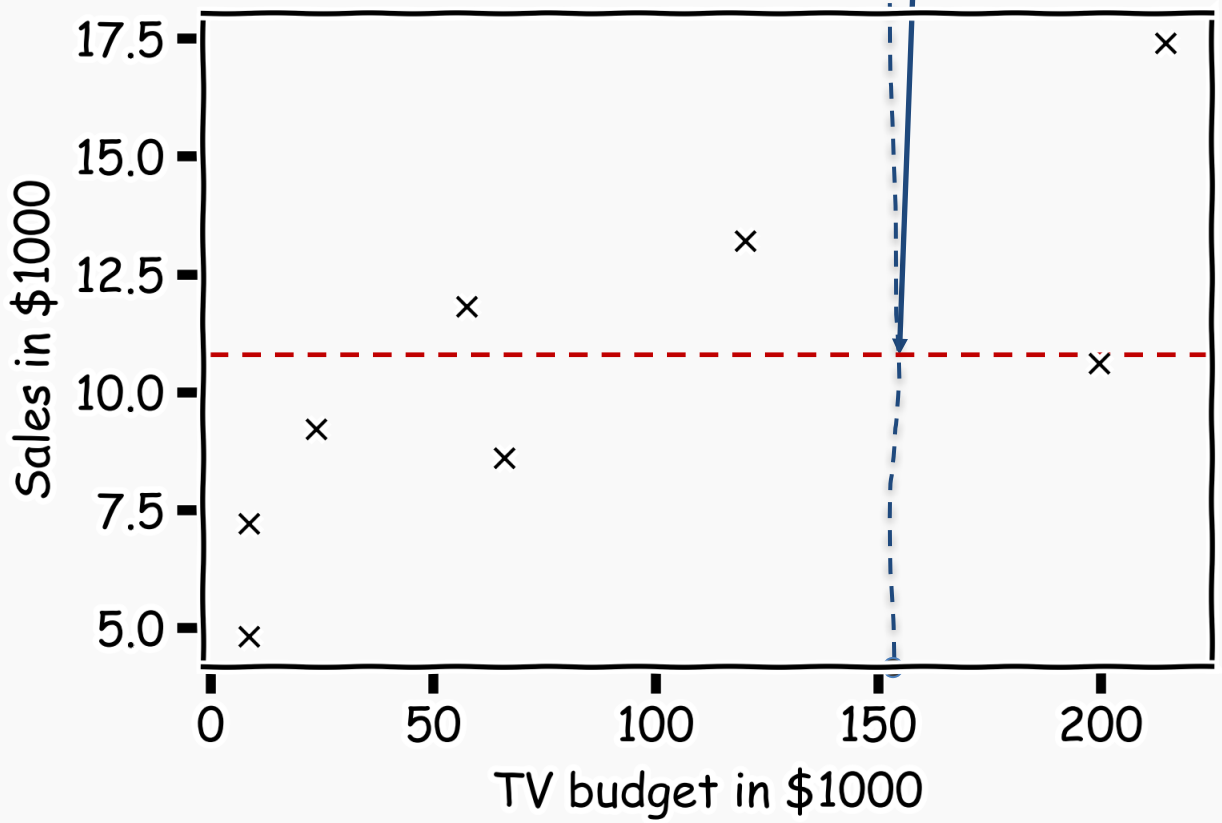
Statistical Model

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

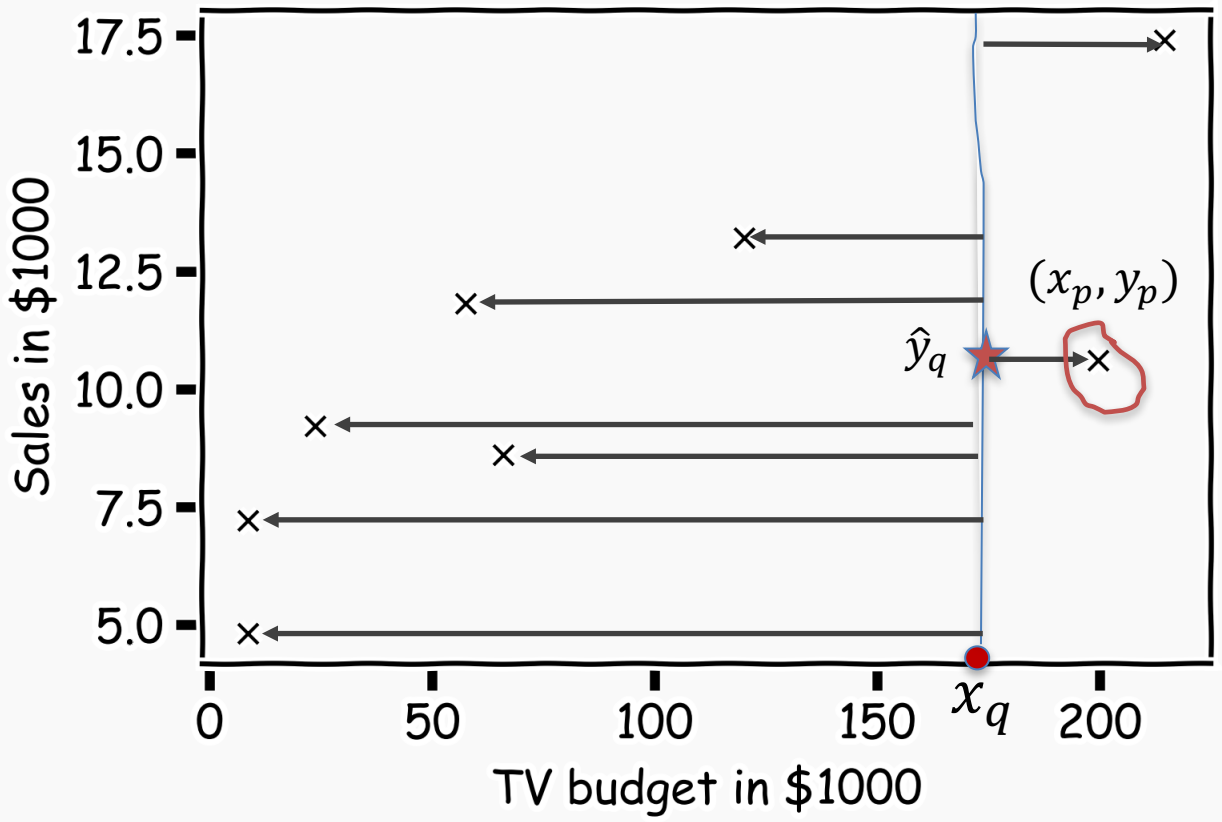
For some problems, what's important is obtaining $\hat{f}$, our estimate of f. These are called ***inference*** problems.

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

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

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

Simple Prediction Model



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

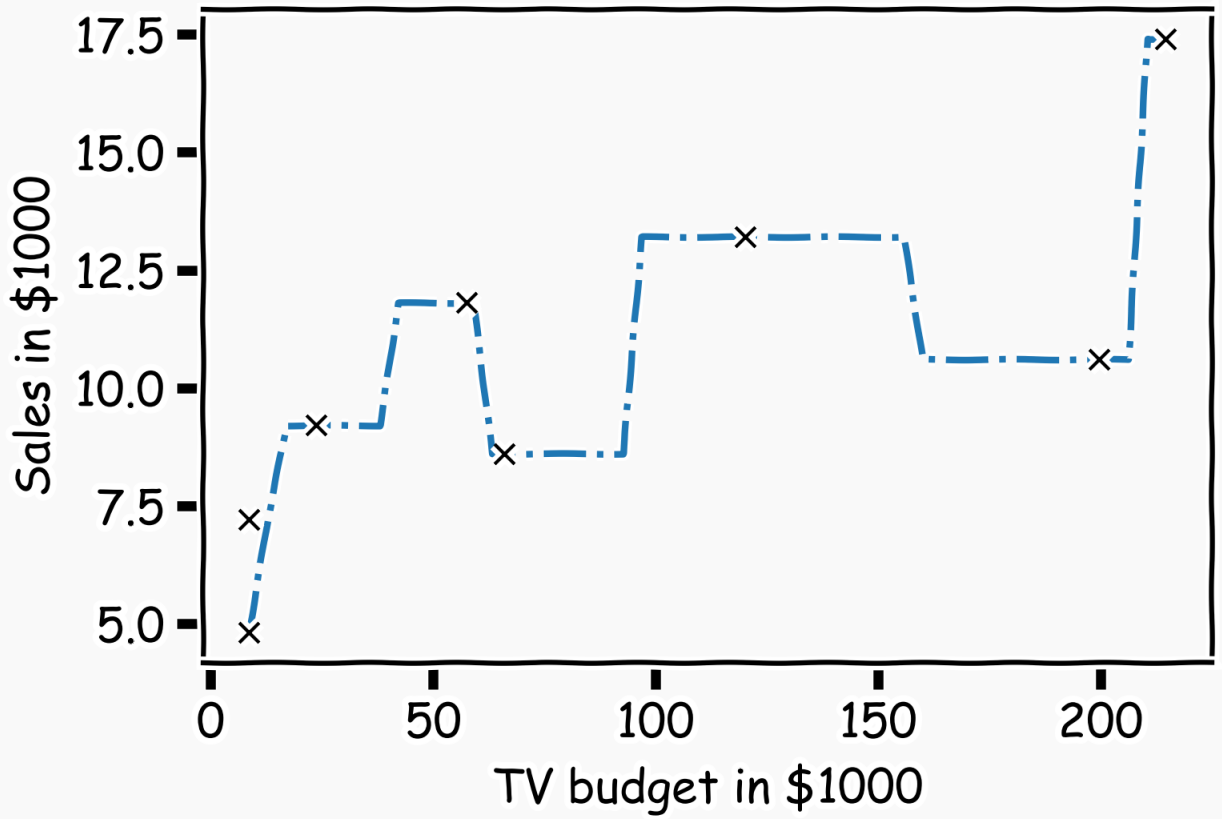
Find distances to all other points $D(x_q, x_i)$

Find the nearest neighbor, (x_p, y_p)

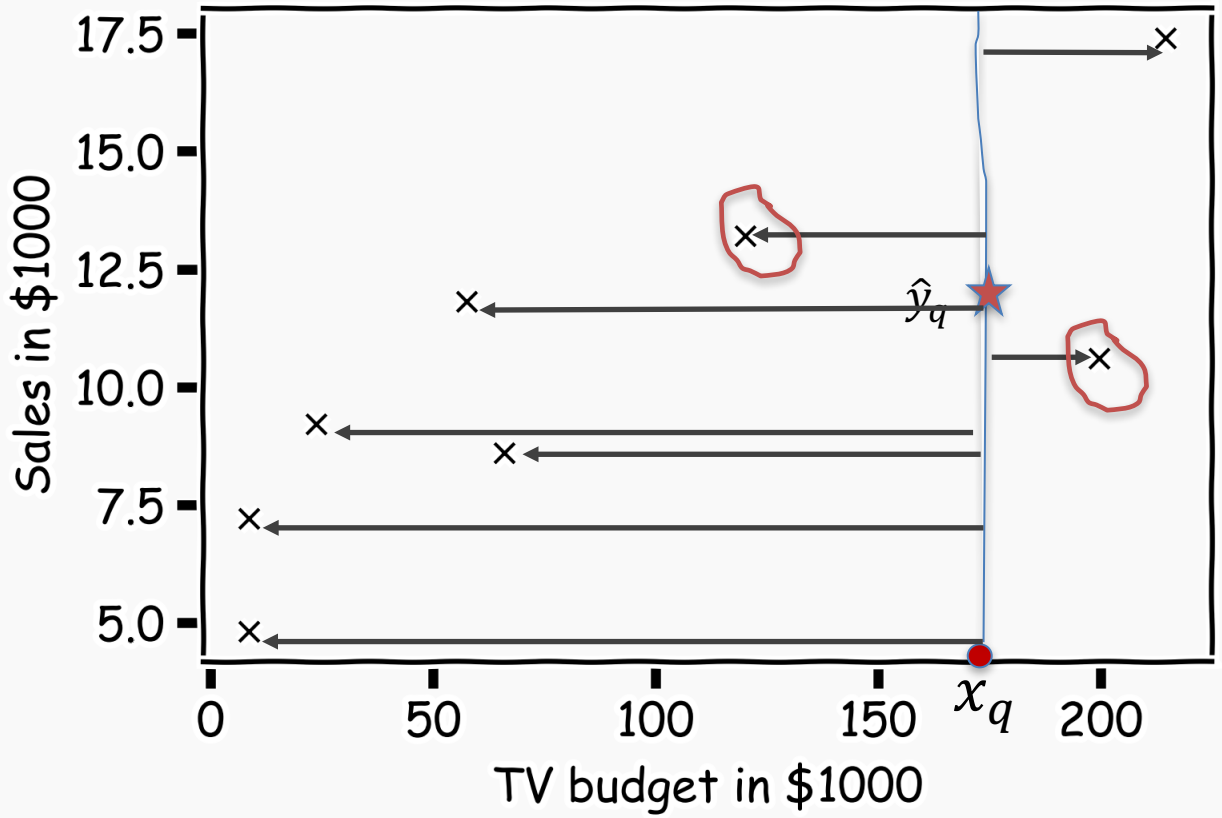
Predict $\hat{y}_q = y_p$

Simple Prediction Model

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



Extend the Prediction Model



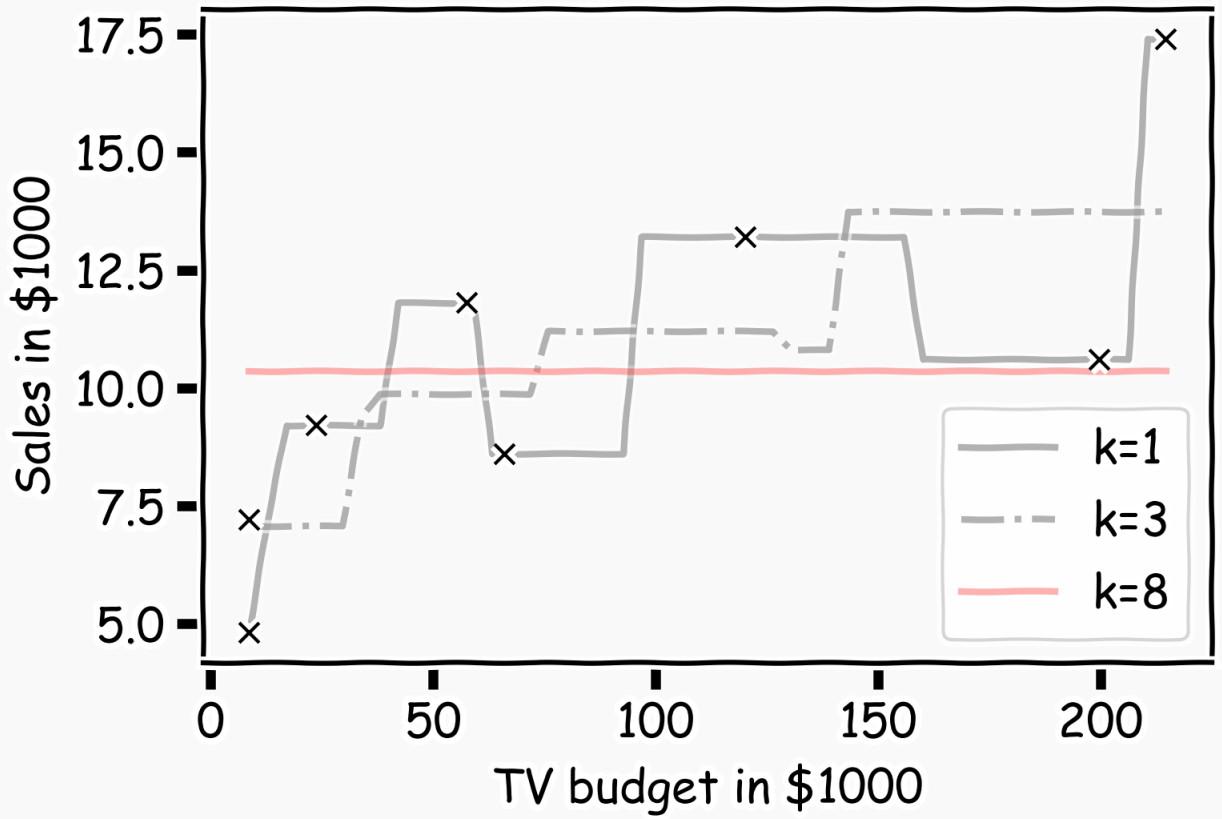
What is $\hat{y}_q$ at some x_q ?

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

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

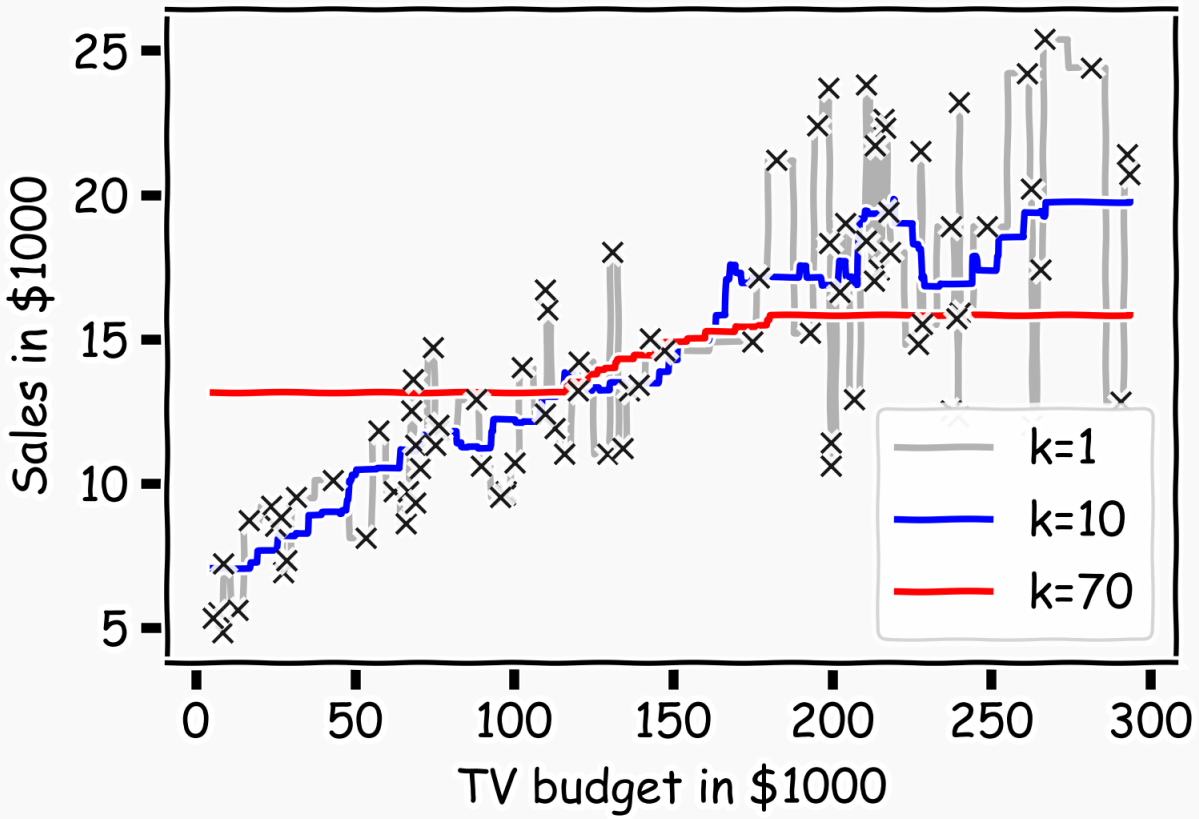
Predict $\widehat{y_q} = \frac{1}{k}\sum_i^k y_{q_i}$

Simple Prediction Models



Simple Prediction Models

We can try different k-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.

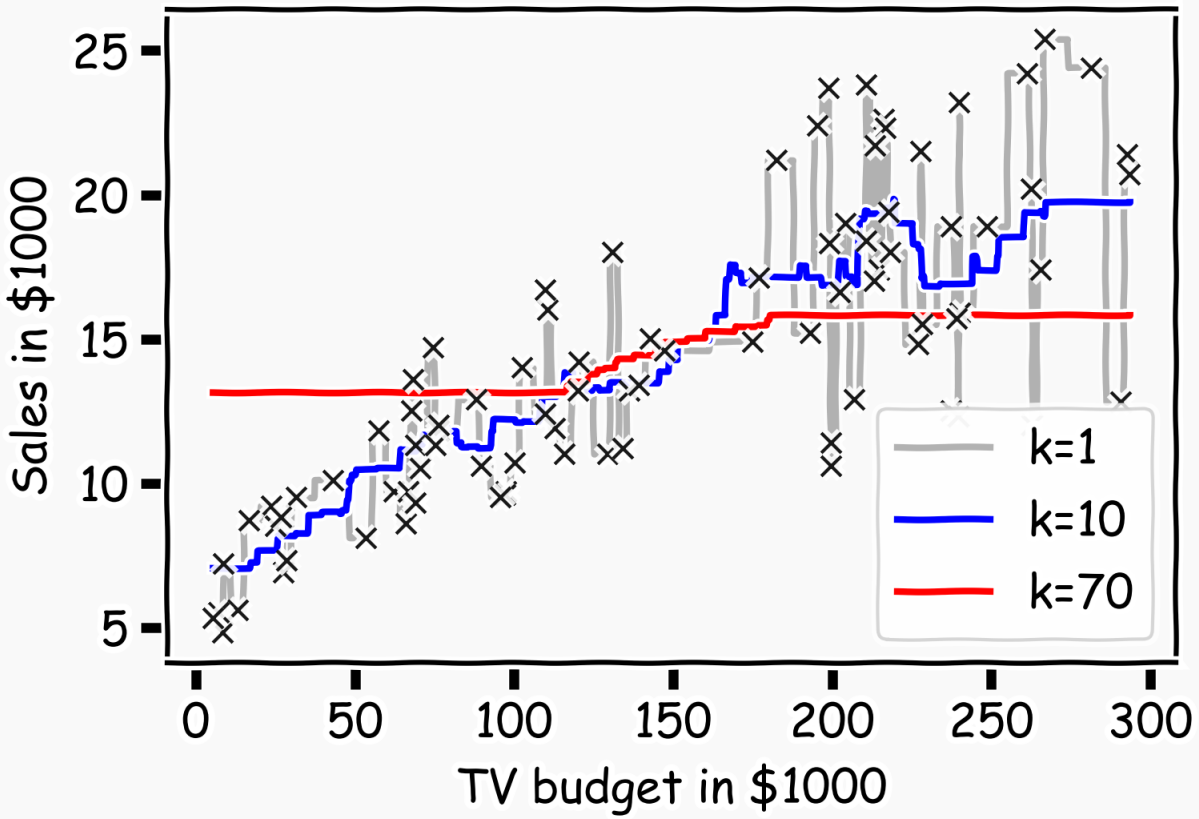
k-Nearest Neighbors - kNN

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).

ED quiz: Lecture 4 | part 1



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?

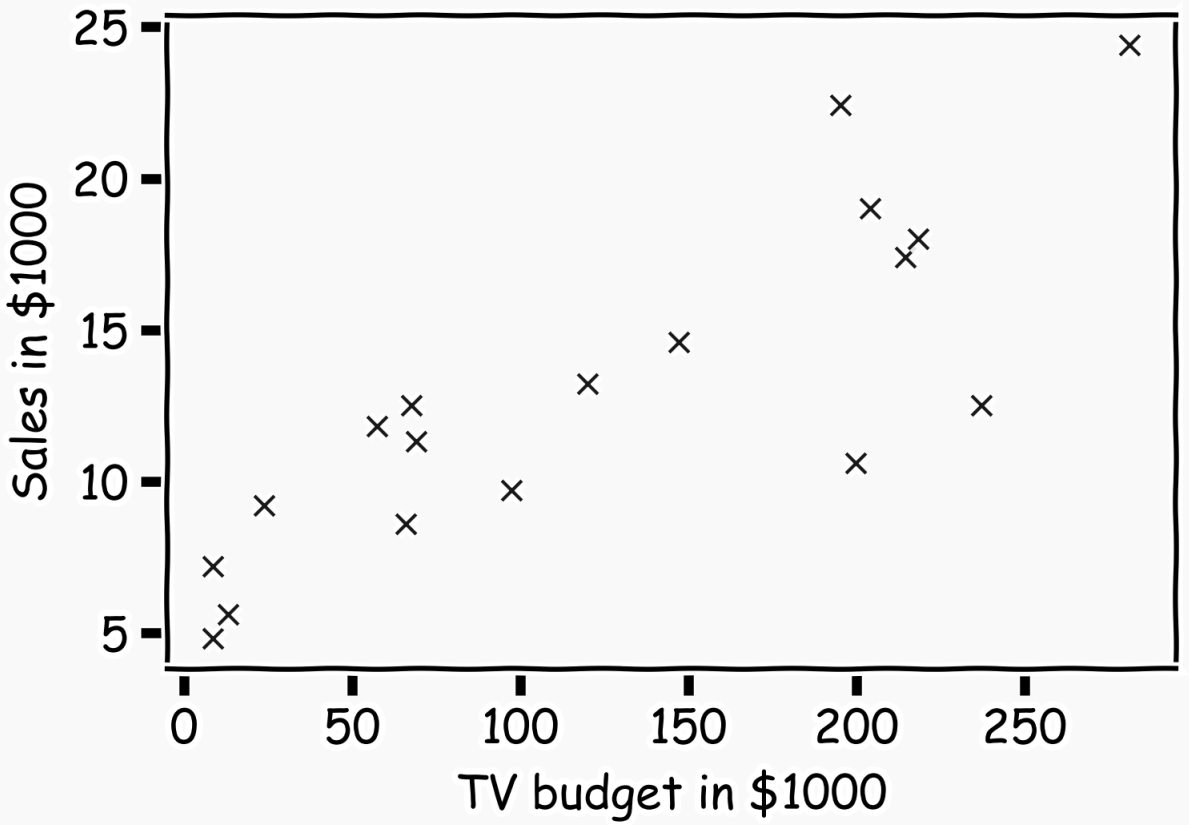
How well do we know $\hat{f}$

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

Error Evaluation

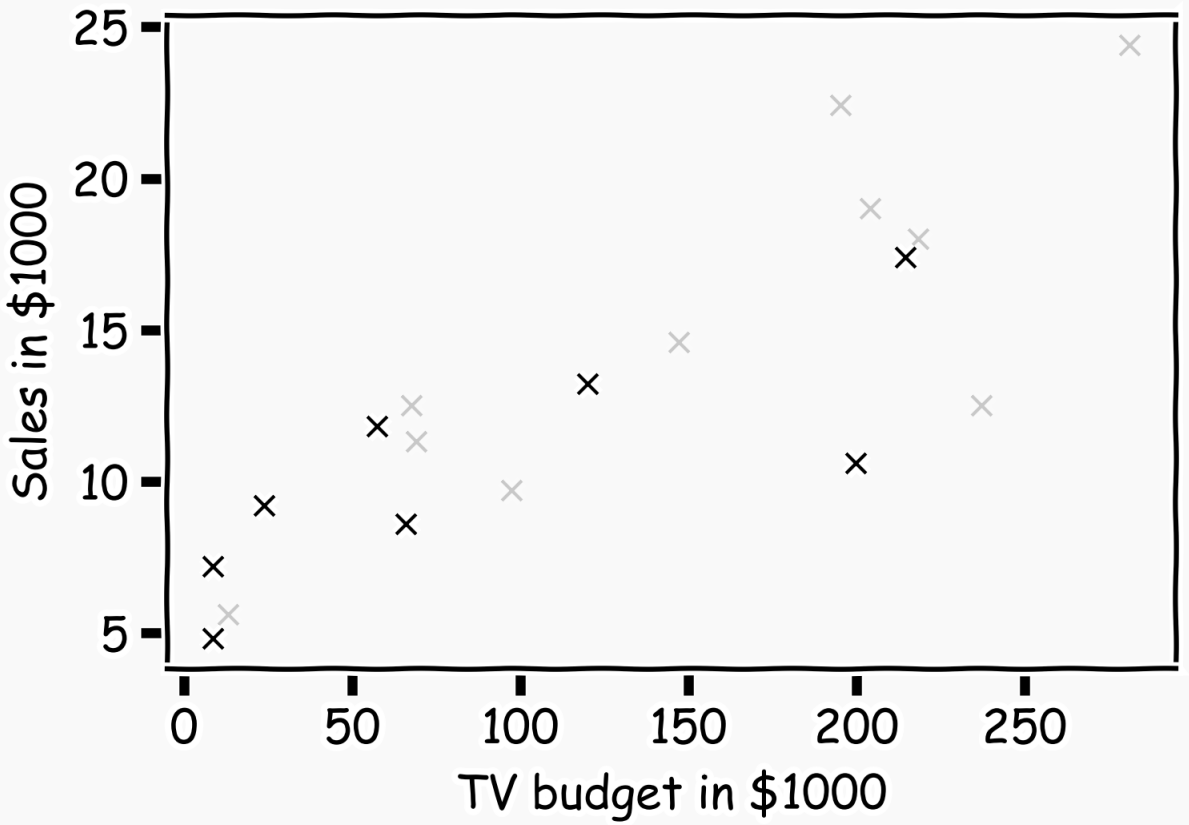
Error Evaluation

Start with some data.



Error Evaluation

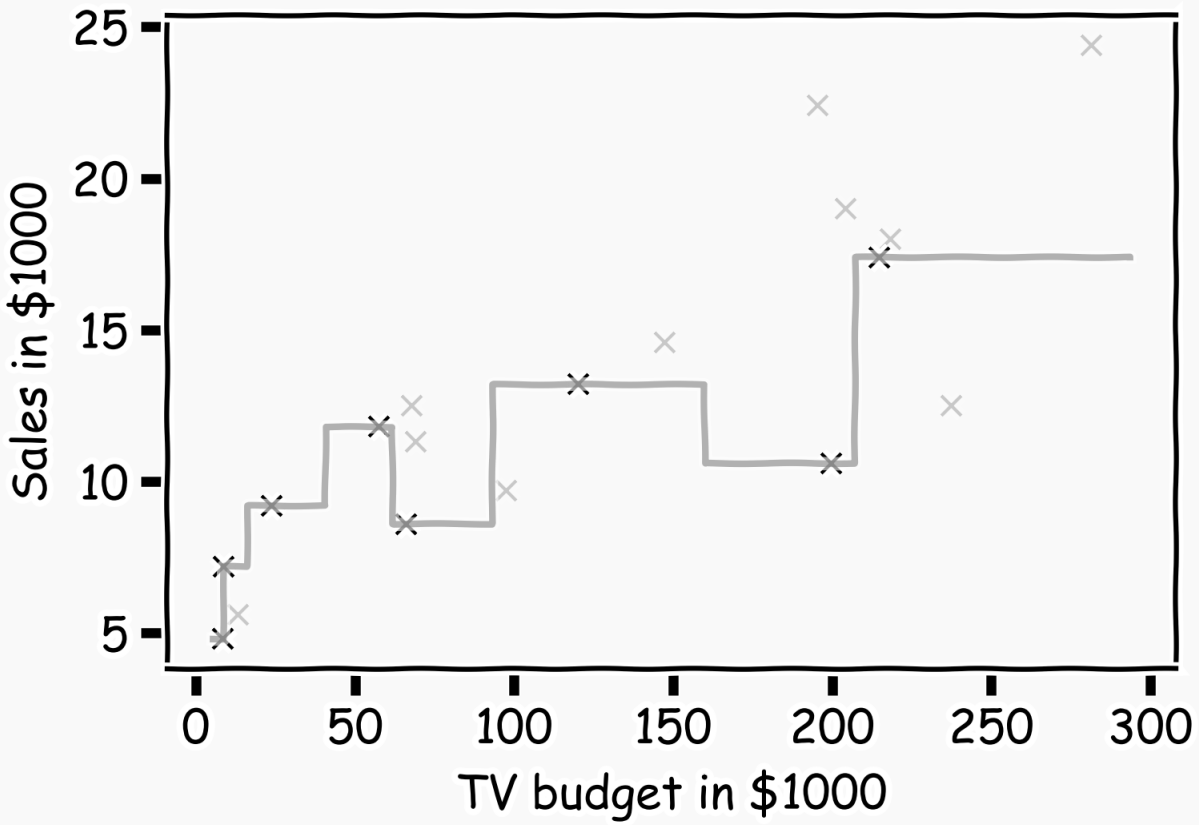
Hide some of the data from the model. This is called **train-test** split.



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

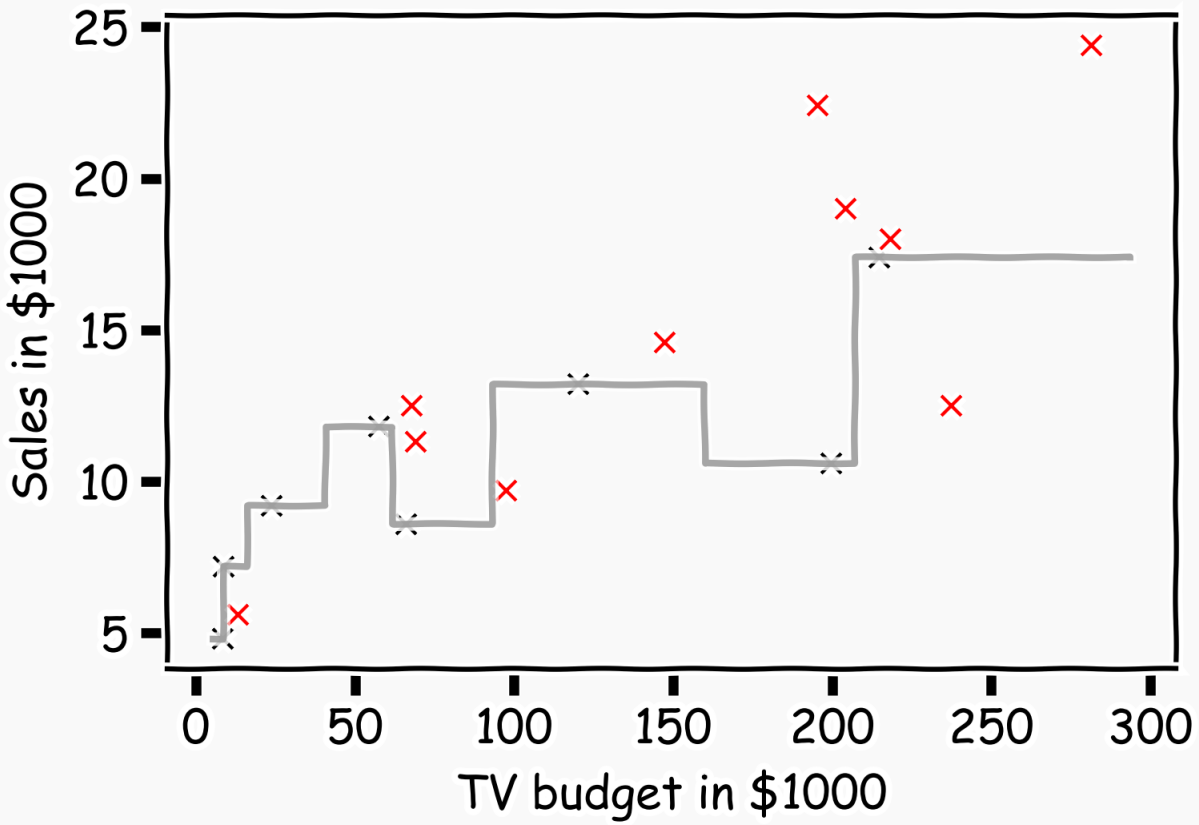
Error Evaluation

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



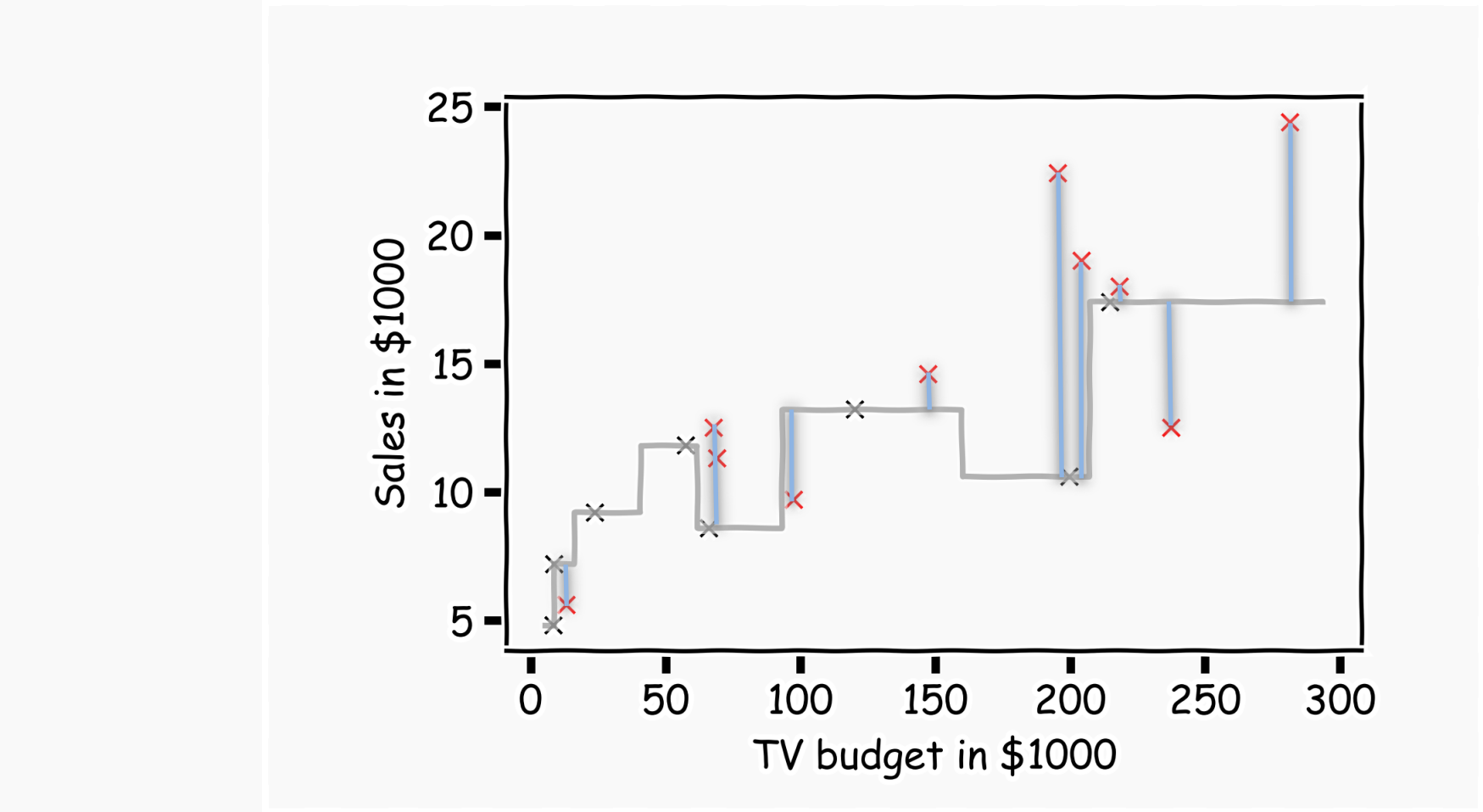
Error Evaluation

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



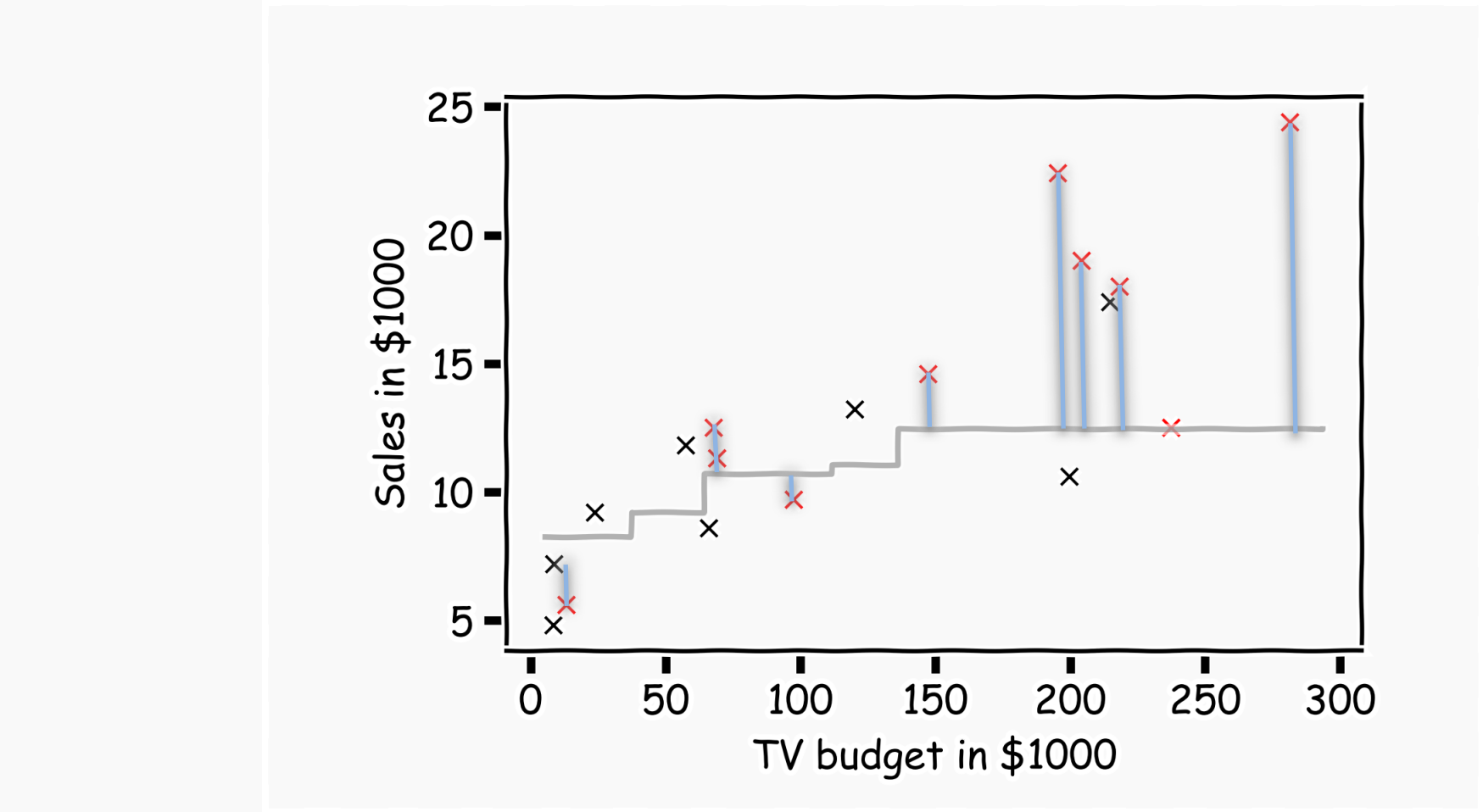
Error Evaluation

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?

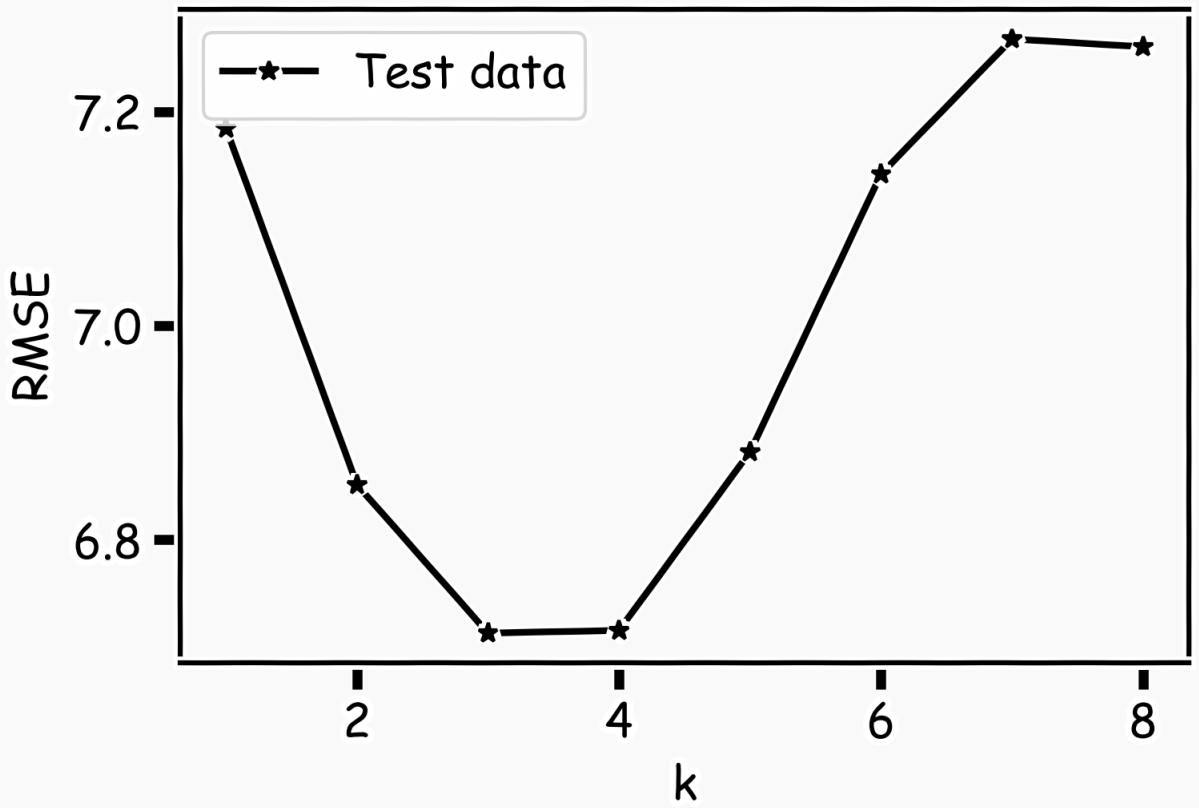
How well do we know $\hat{f}$

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?

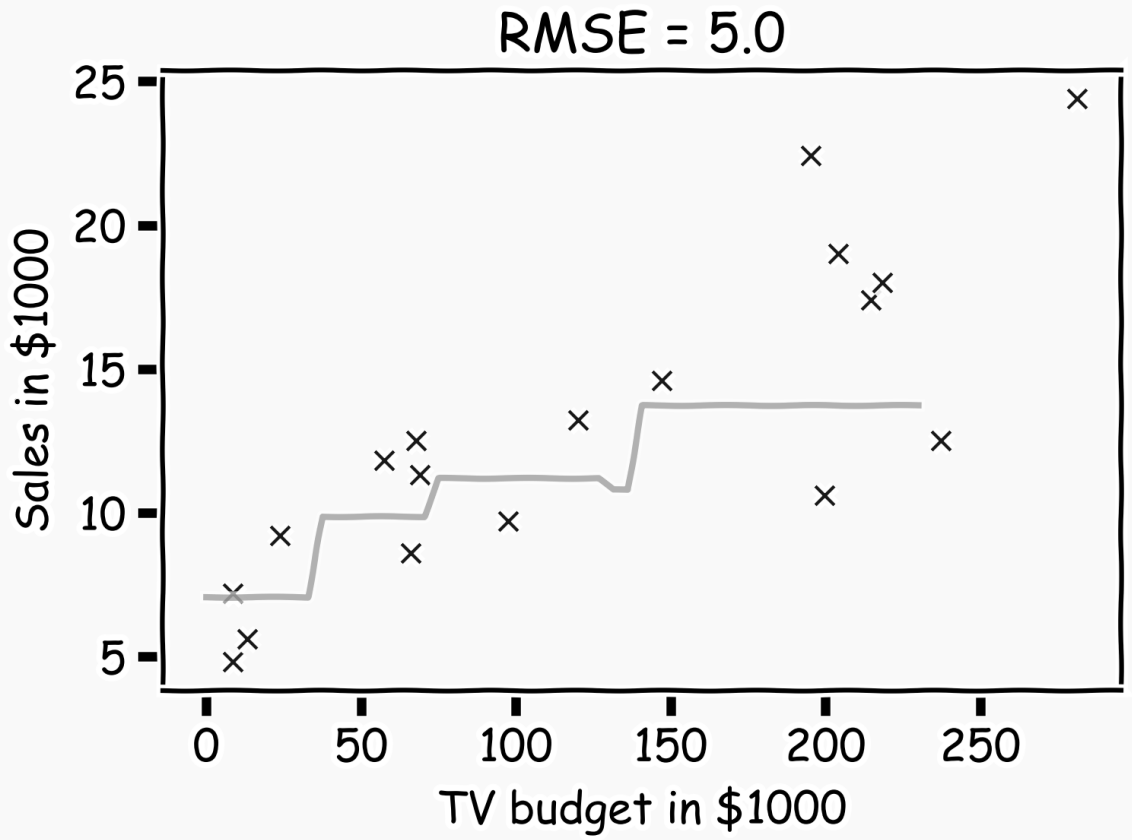
How well do we know $\hat{f}$

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

Model Fitness

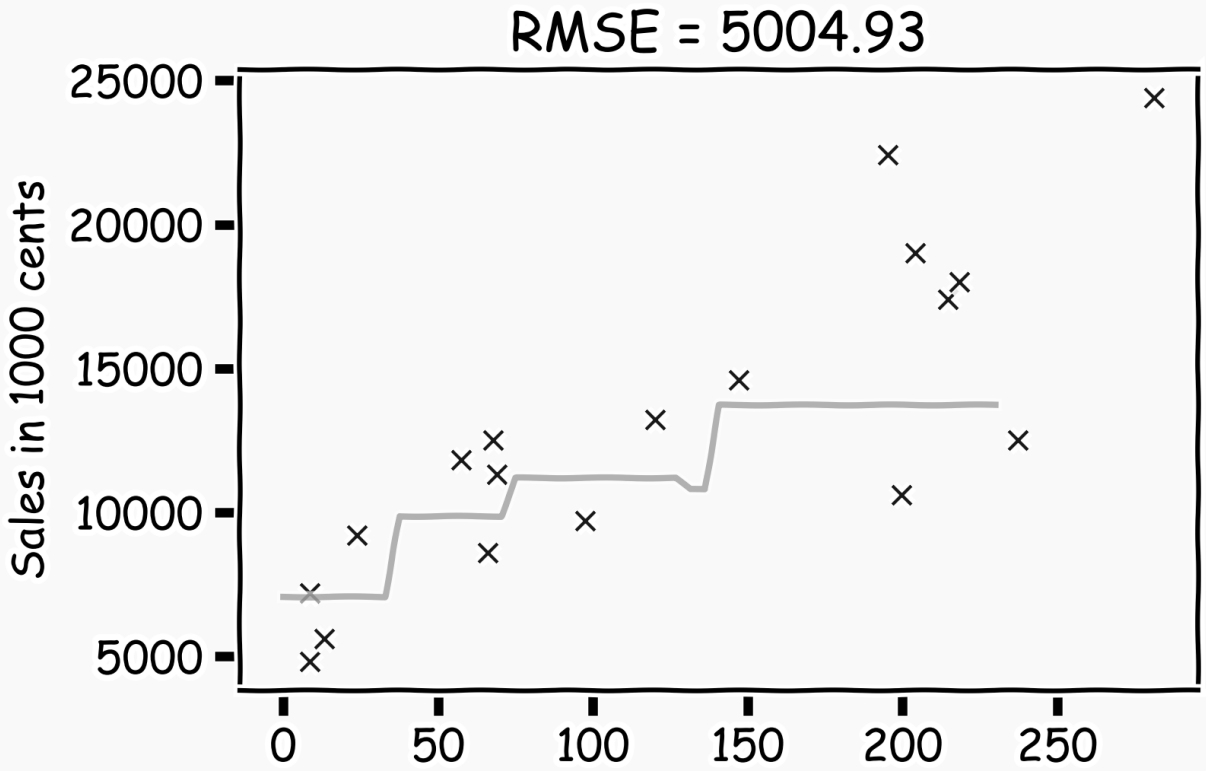
Model fitness

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



Model fitness

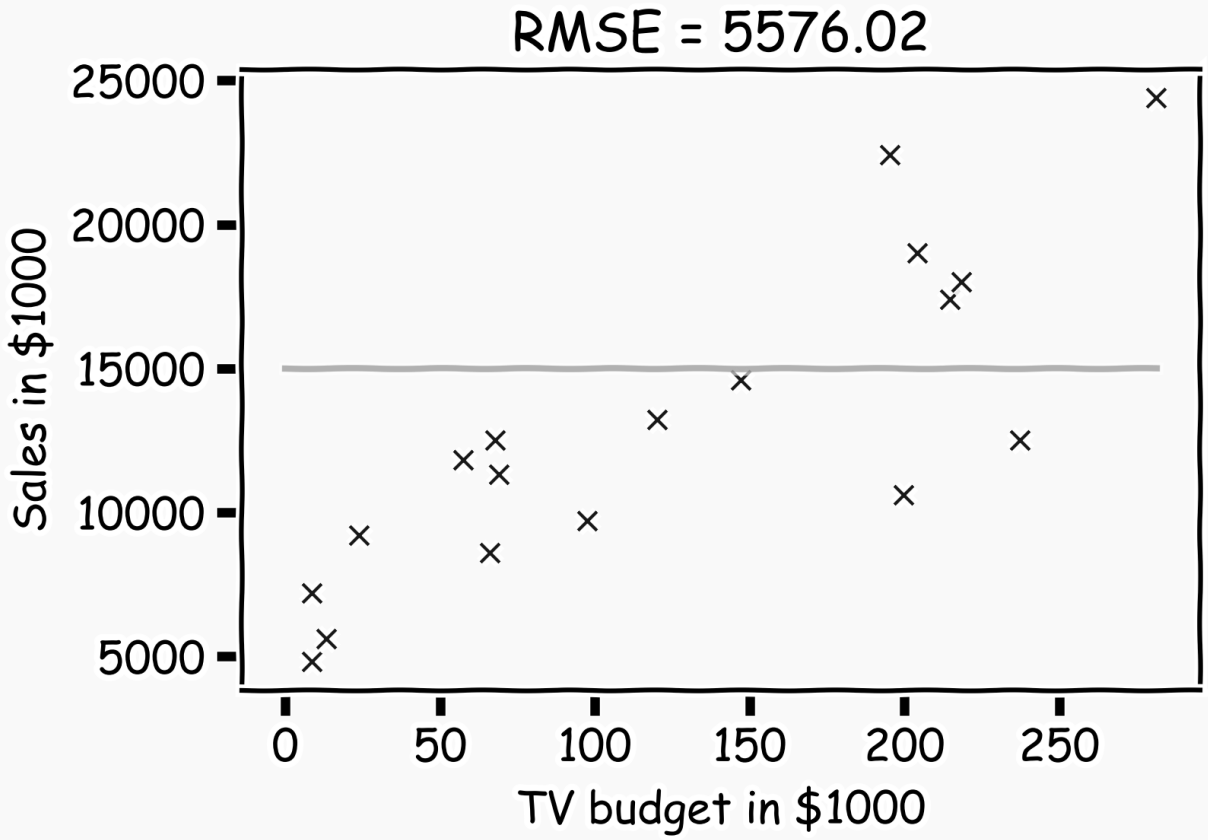
What if we measure the Sales in cents instead of dollars?



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.

Summary

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 $\hat{f}$

Summary

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 do we know $\hat{f}$

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

Next lecture