CPSC 4300/6300: Applied Data Science

2024

Introduction to Regression

Dr. Aaron J. Masino
Associate Professor, School of Computing

College of
ENGINEERING, COMPUTING
AND APPLIED SCIENCES

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 number of views a YouTube video will get next week based on video length, the date it was posted, previous number of views, etc.
- Predict a Netflix user's rating of movies they haven't viewed from their previous movie ratings, demographic data, and ratings of other "similar" users.

Let's get familiar with some terminology

Consider an **Advertising data set** that contains sales of products in 200 different markets, along with advertising budgets for three different media: TV, radio, and newspaper. Everything is given in units of $1000.

We want to predict sales based on amount spent across the different media.

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)

Definitions

- We'll assume we have n observations of $p + 1$ variables (counting the outcome)
- *outcome* or *response:* the variable to predict
  - typically denoted by Y
  - individual measurements denoted by y_i.
- *features* or *predictor:* variables used to make predictions
  - typically denoted by $X = X_1, \ldots, X_p$
  - individual measurements denoted by $x_{i,j}$.

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

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

True vs. model relation between predictors and outcome

We 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 the relation between Y and X, ε is the random amount (unrelated to X) that Y differs from the rule $f(X)$.

A ***model***, $\hat{f}$, is any algorithm that estimates f.

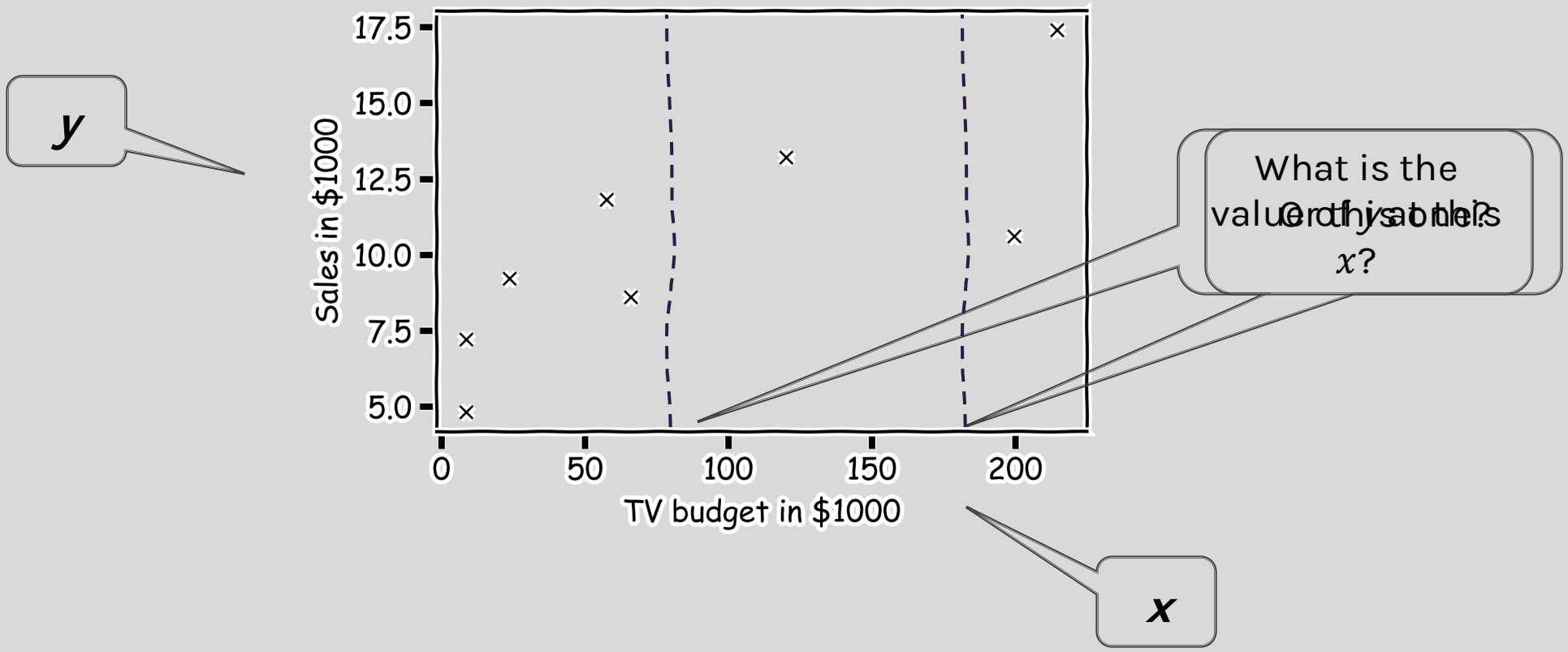
We will focus on ***statistical models*** – *those derived by algorithms that use observational data to estimate* f

All models are wrong,
but some are useful.

George E. P. Box

Statistical Model

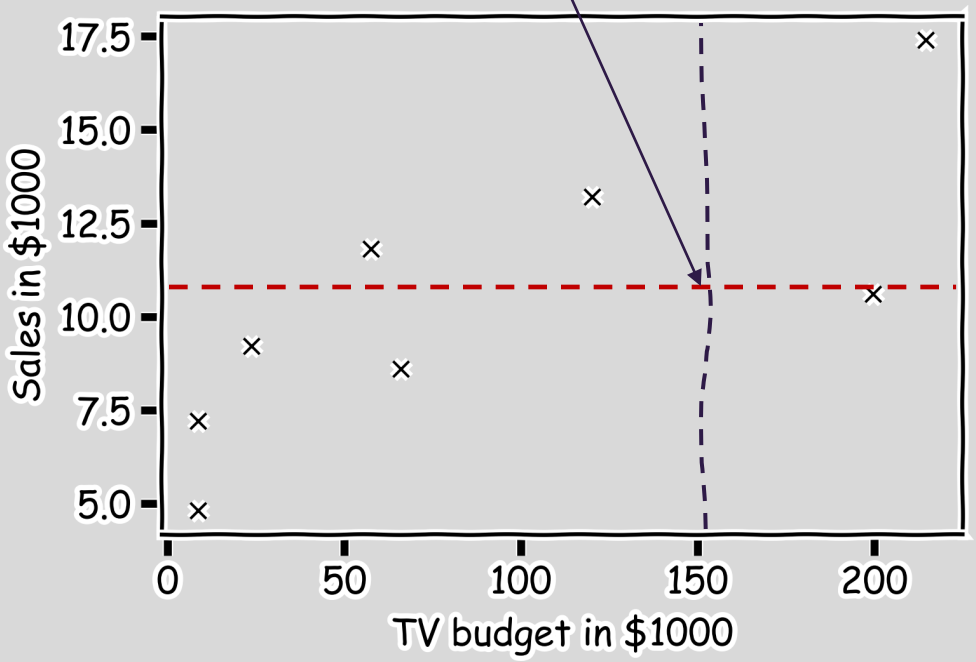
How do we find $\hat{f}(x)$ using data?



Statistical Model

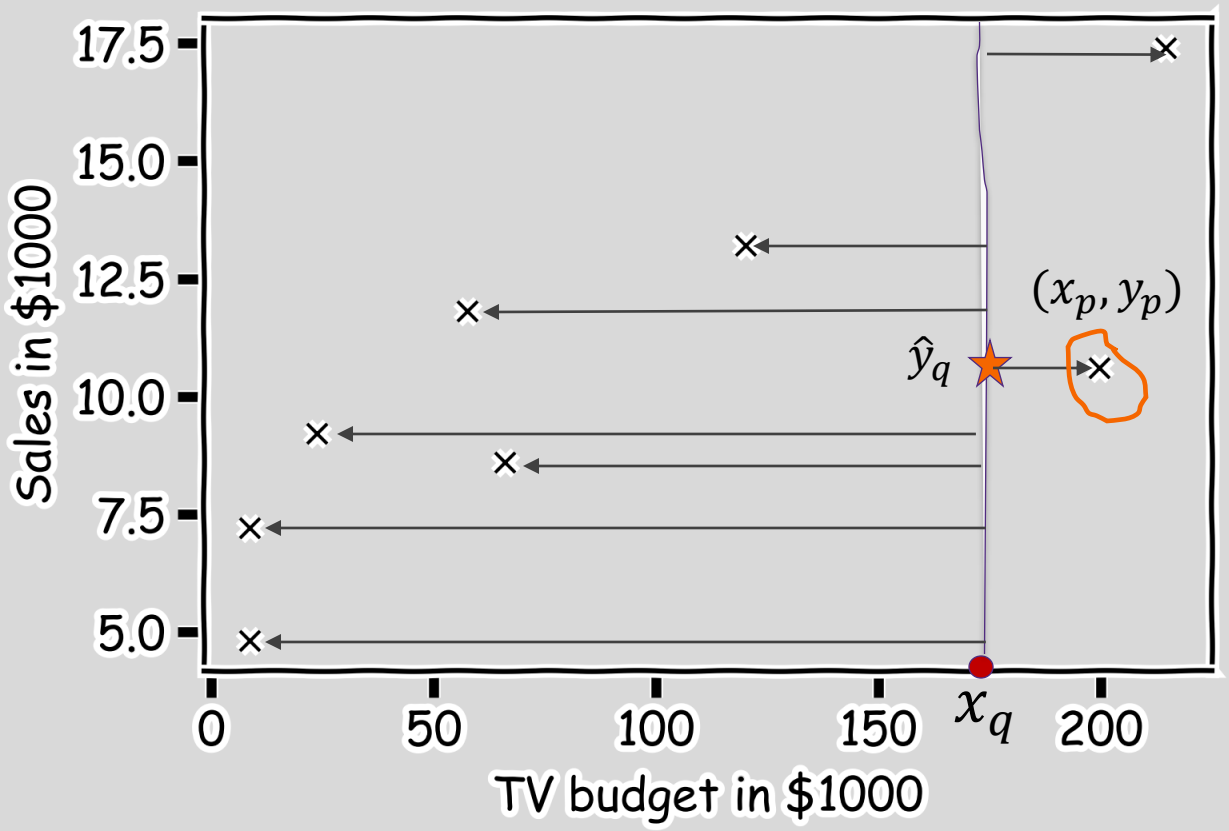
Simple idea is to take the mean of all observations

$$\hat{f}(x) = \frac{1}{n}\sum_{1}^{n} y_i$$



Is this a good model?

Simple Prediction Model



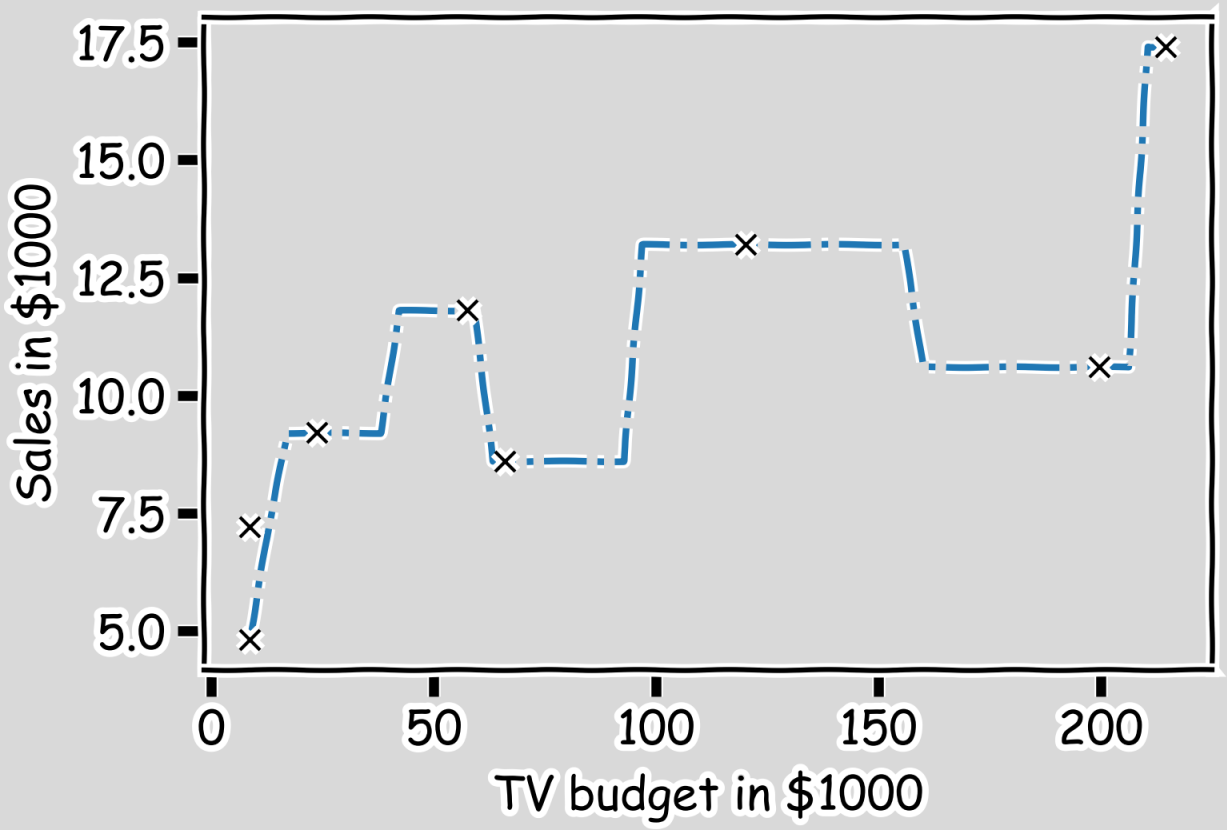
What is $\hat{y}_q$ at some x_q ?

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

Find the nearest neighbor, (x_p, y_p)

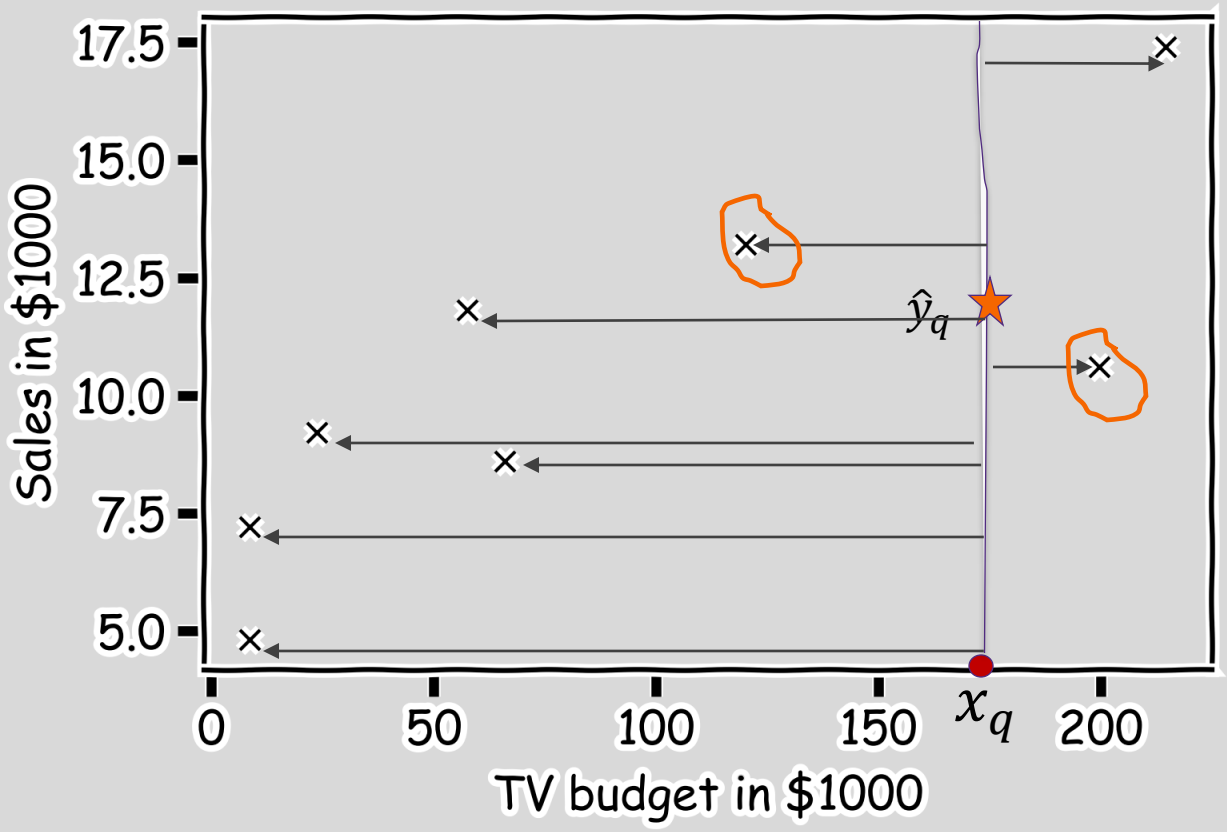
Predict $\hat{y}_q = y_p$

Simple Prediction Model



This blue line represents our model using the single nearest neighbor approach

Extend the Prediction Model



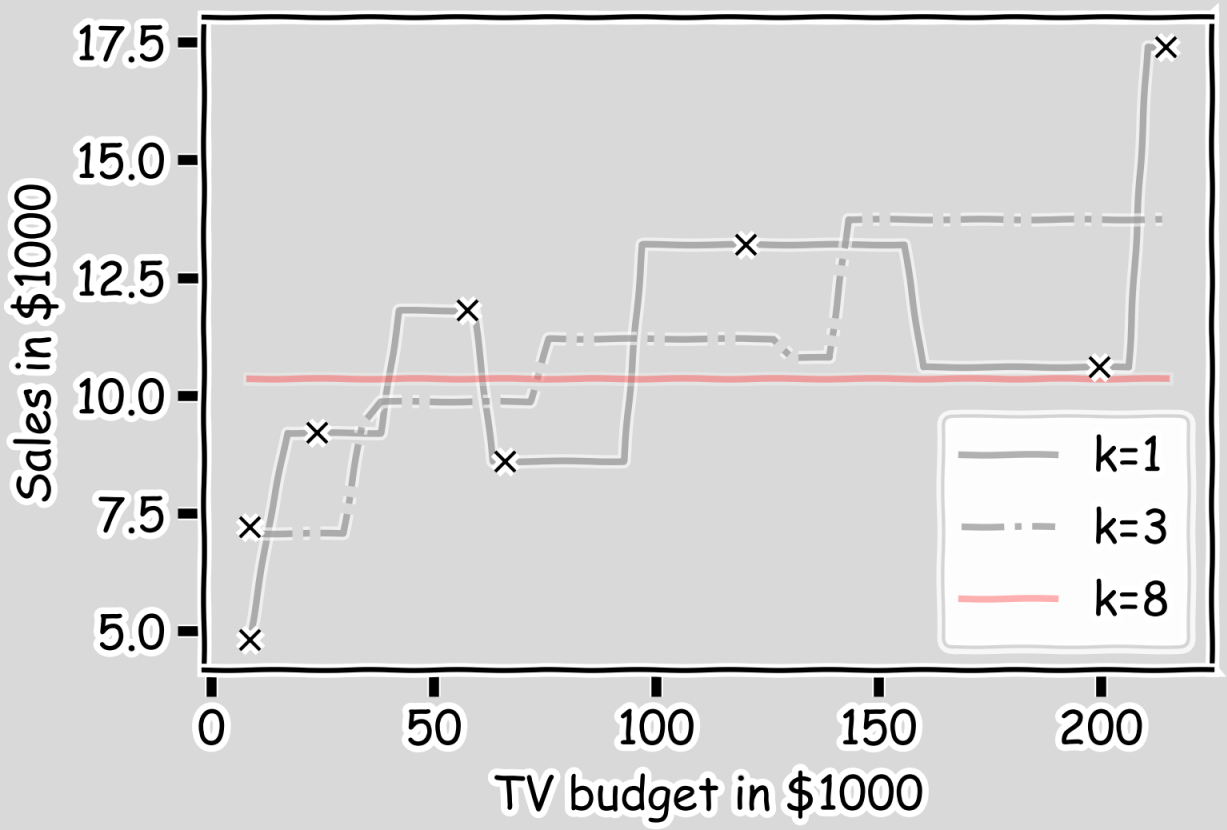
What is $\hat{y}_q$ at some x_q ?

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

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

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

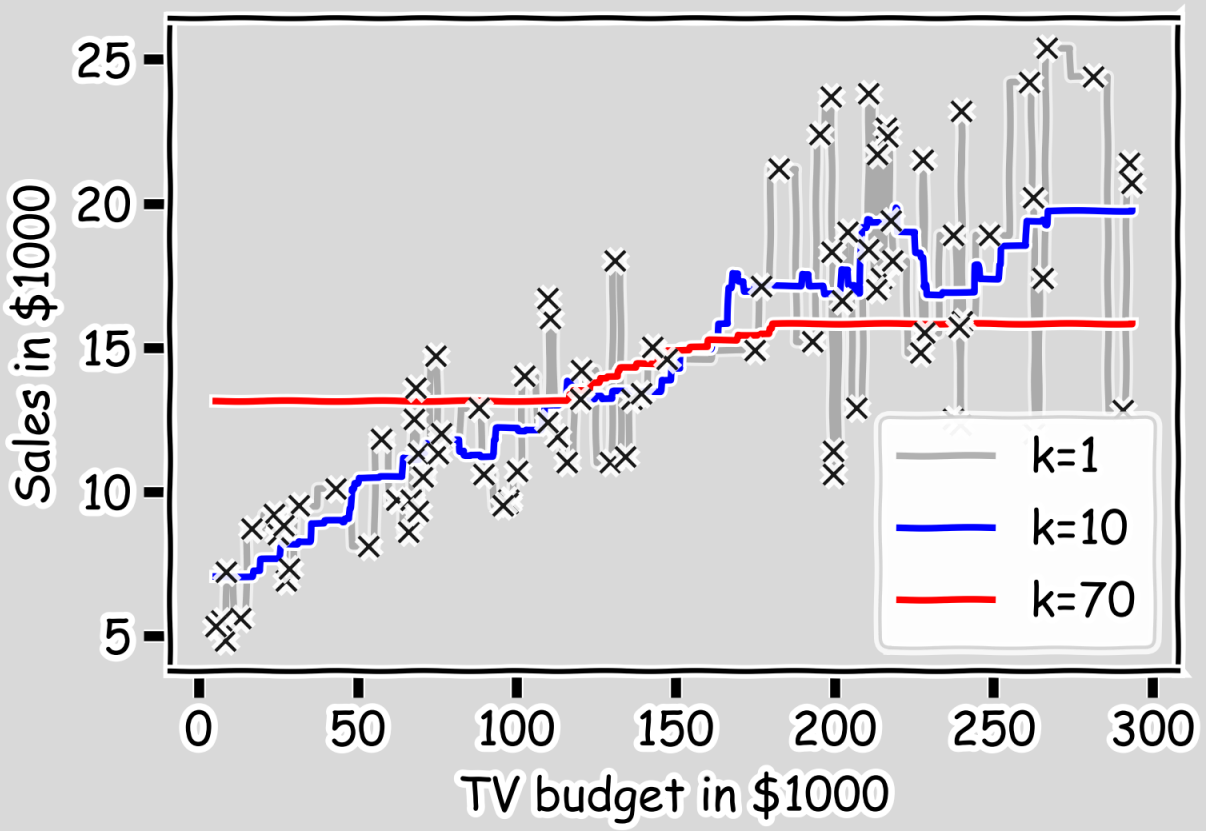
Simple Prediction Models



Models for *k* nearest neighbors approach

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'll see much more on classification models later in the course.

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

Comparison of Two Models

How do we choose from two different models?

Model Fitness

How well does the model perform?

Evaluating Significance of Predictors

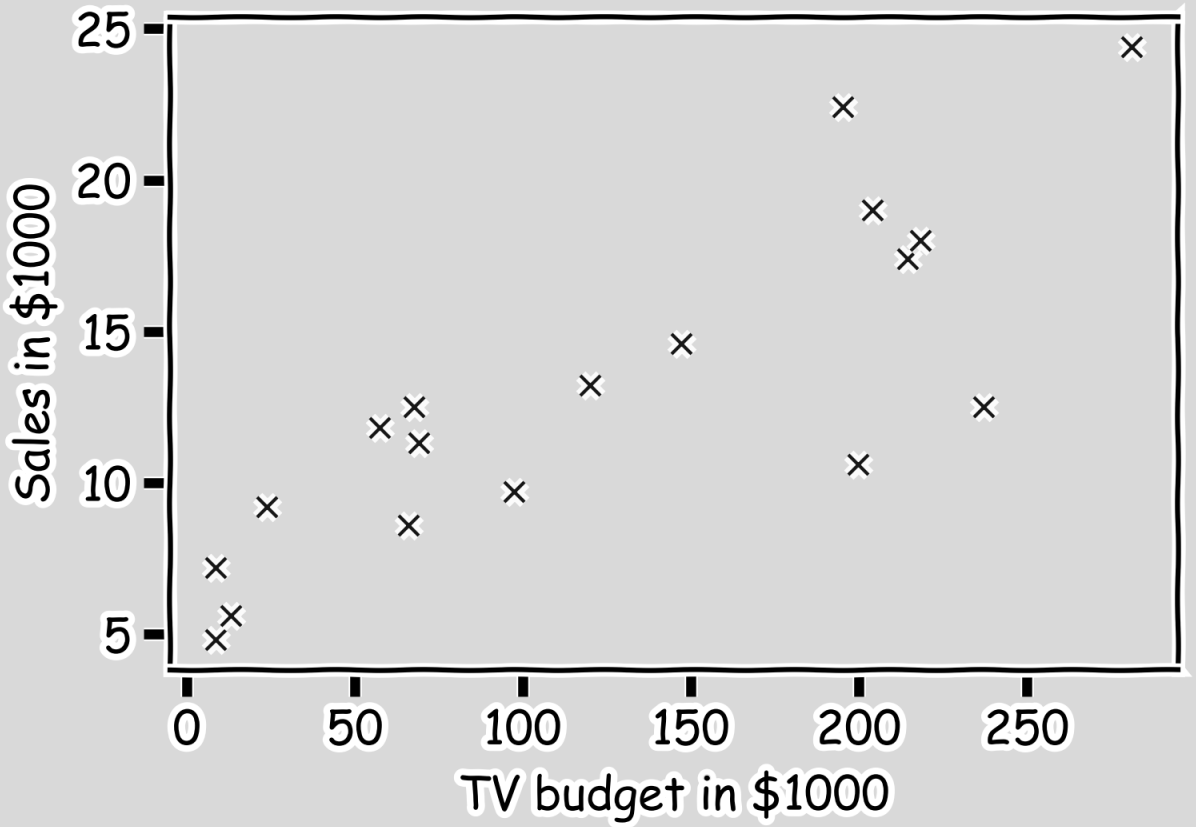
Does the outcome truly depend on the predictors?

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

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

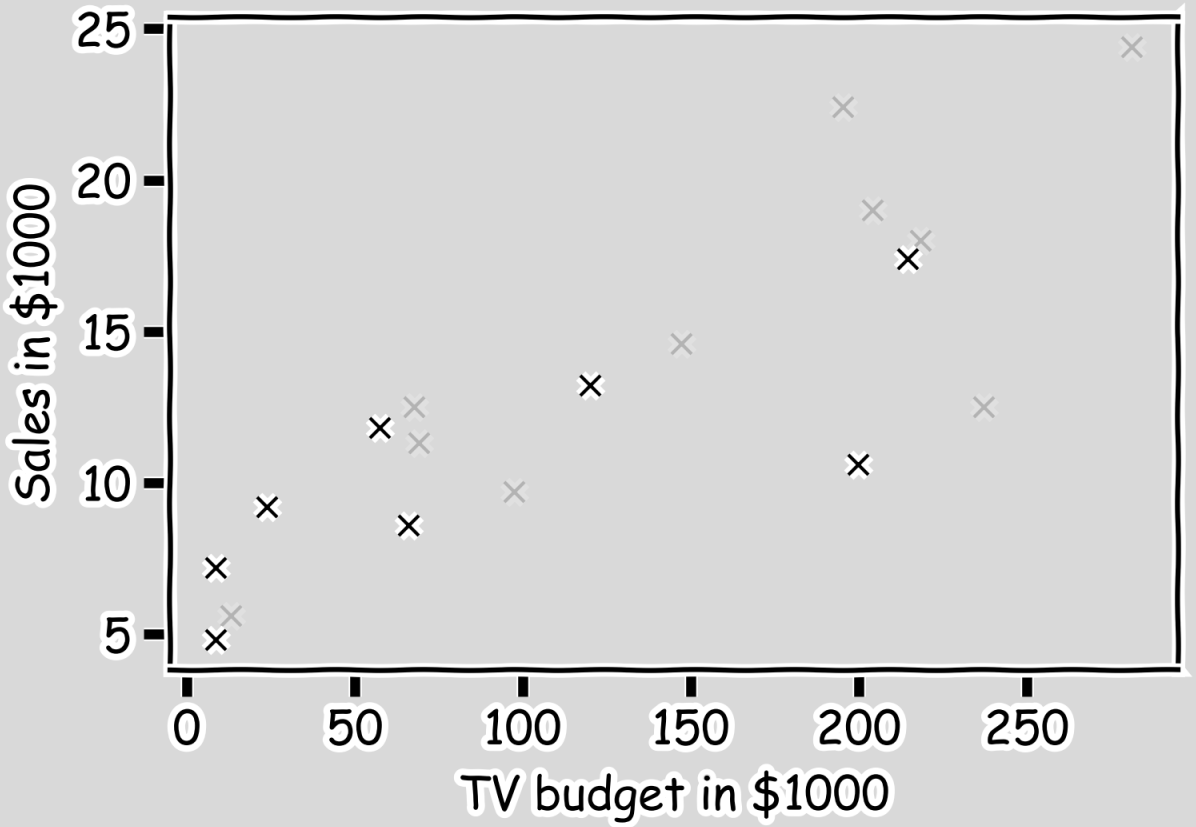
Error Evaluation

Start with some data.



Error Evaluation

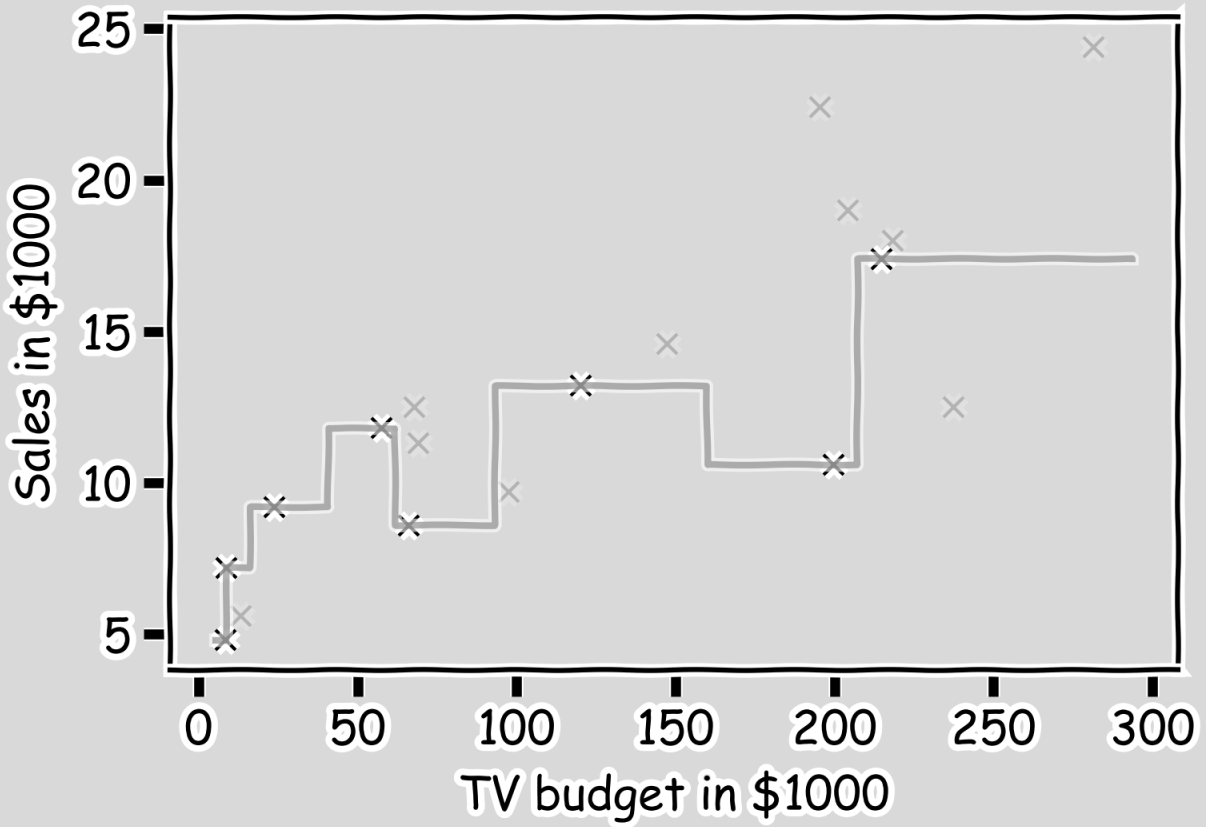
Hide some of the data from the model. This is called train-test split.



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

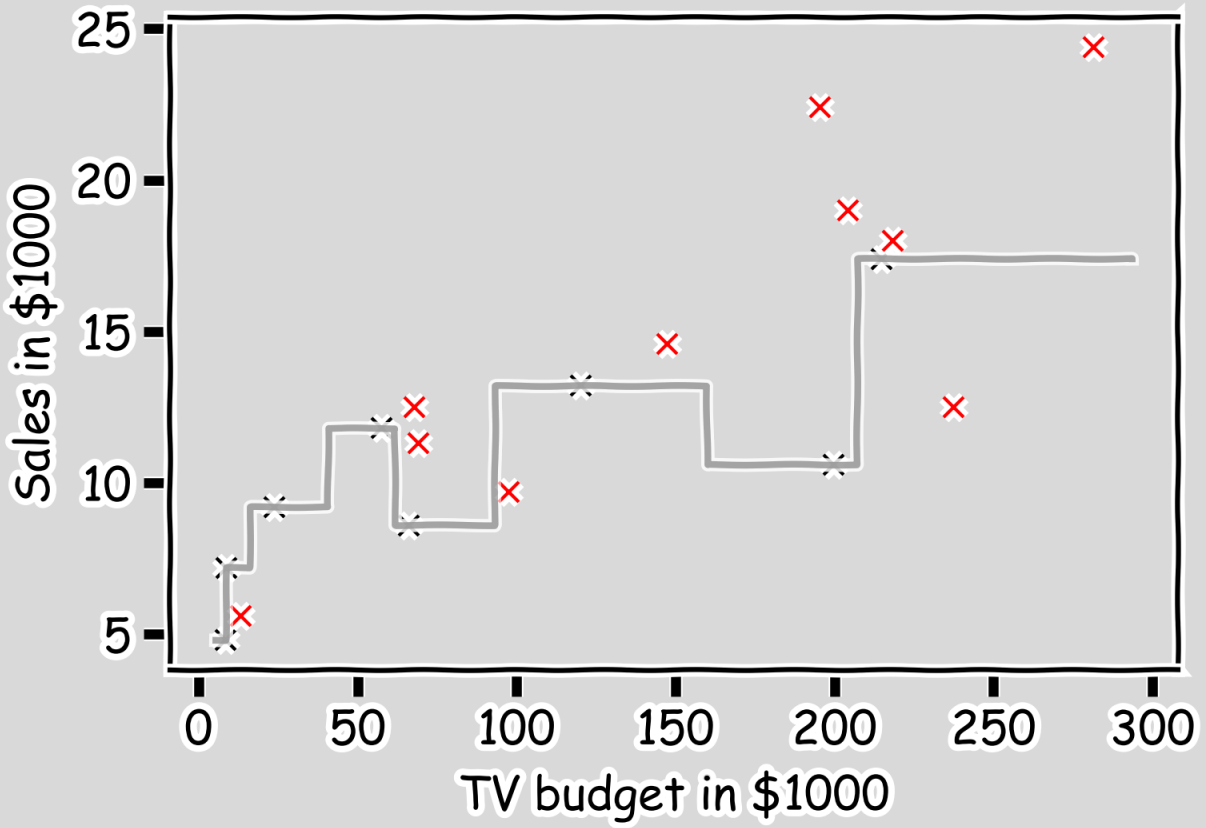
Error Evaluation

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



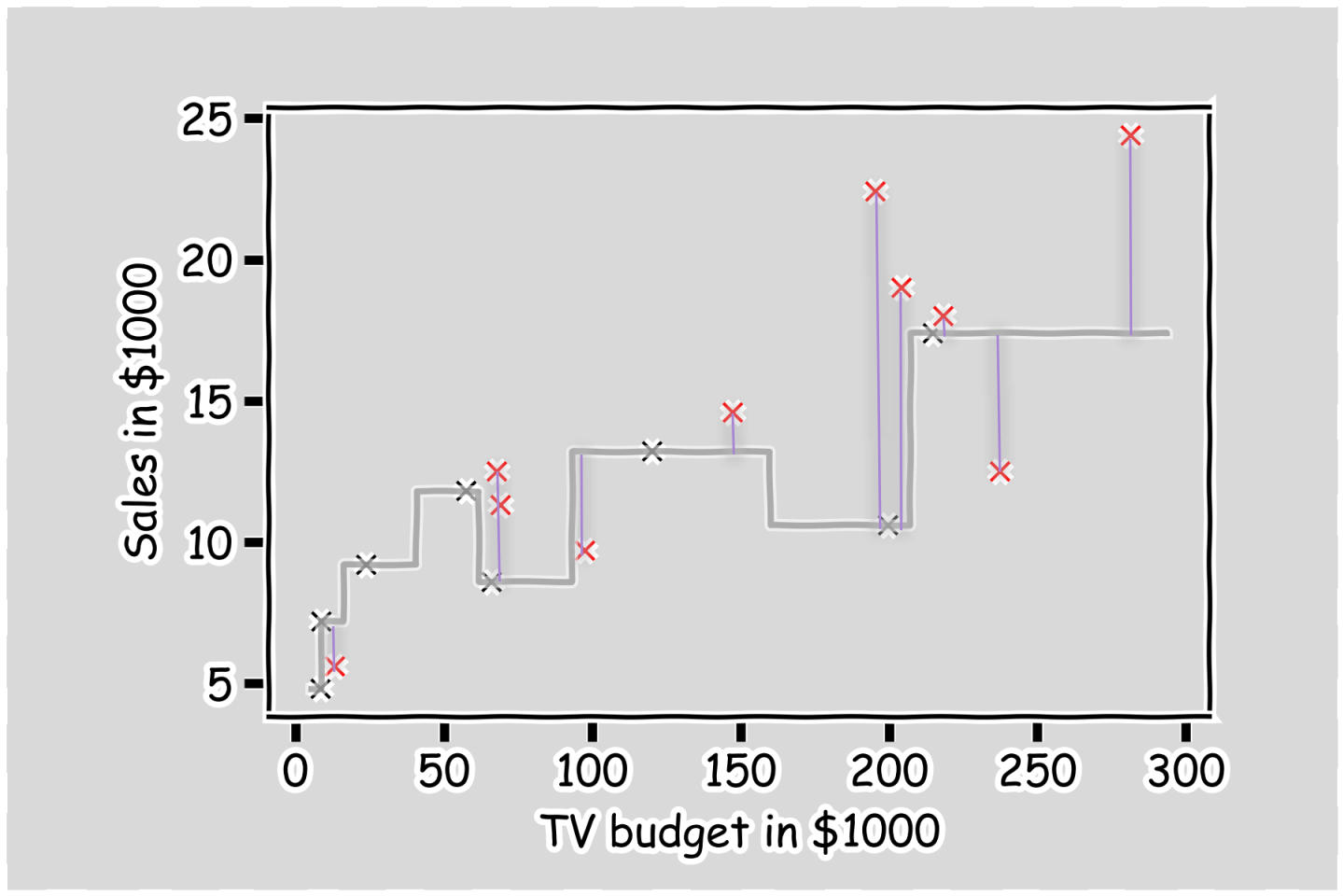
Error Evaluation

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



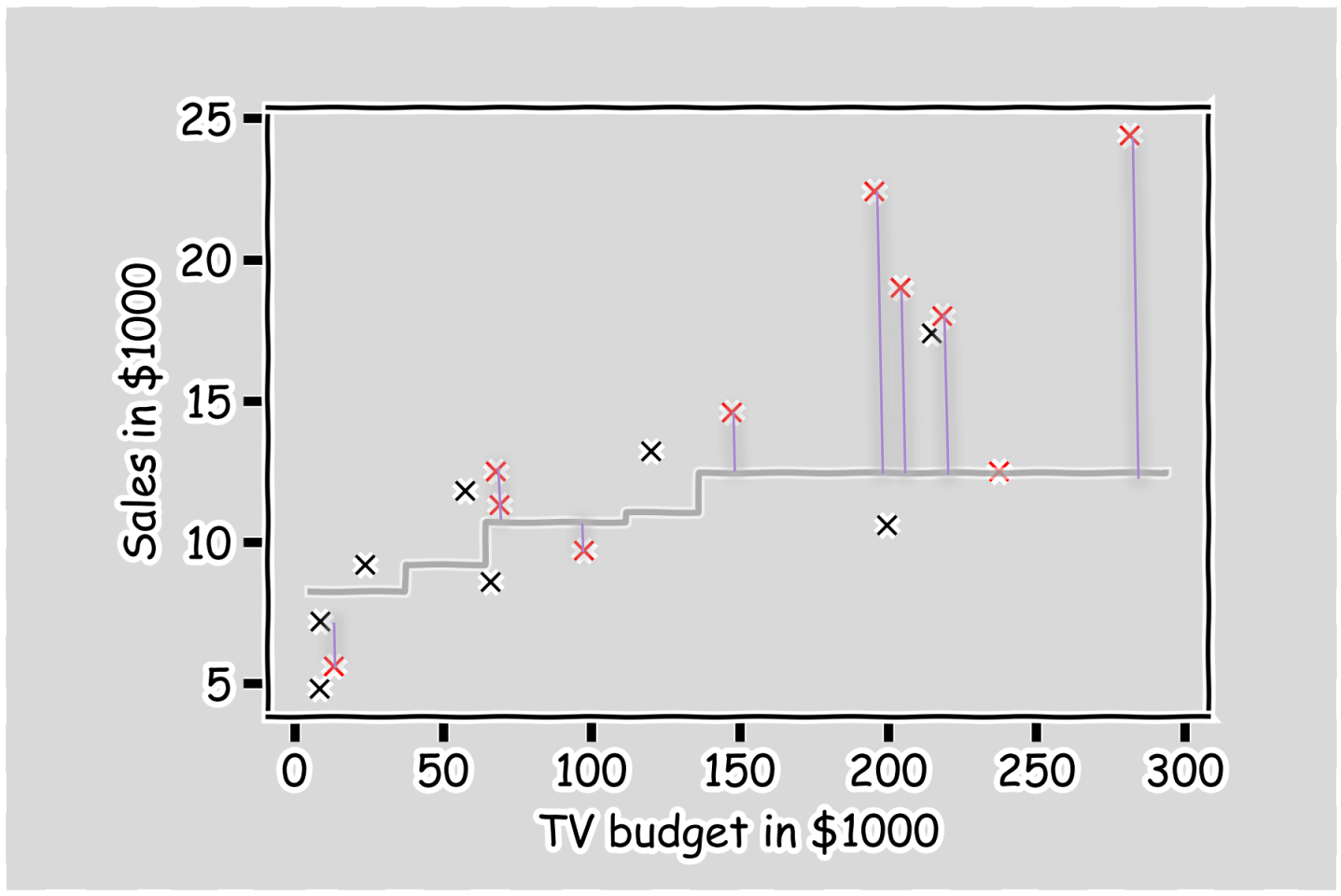
Error Evaluation

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):

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

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!

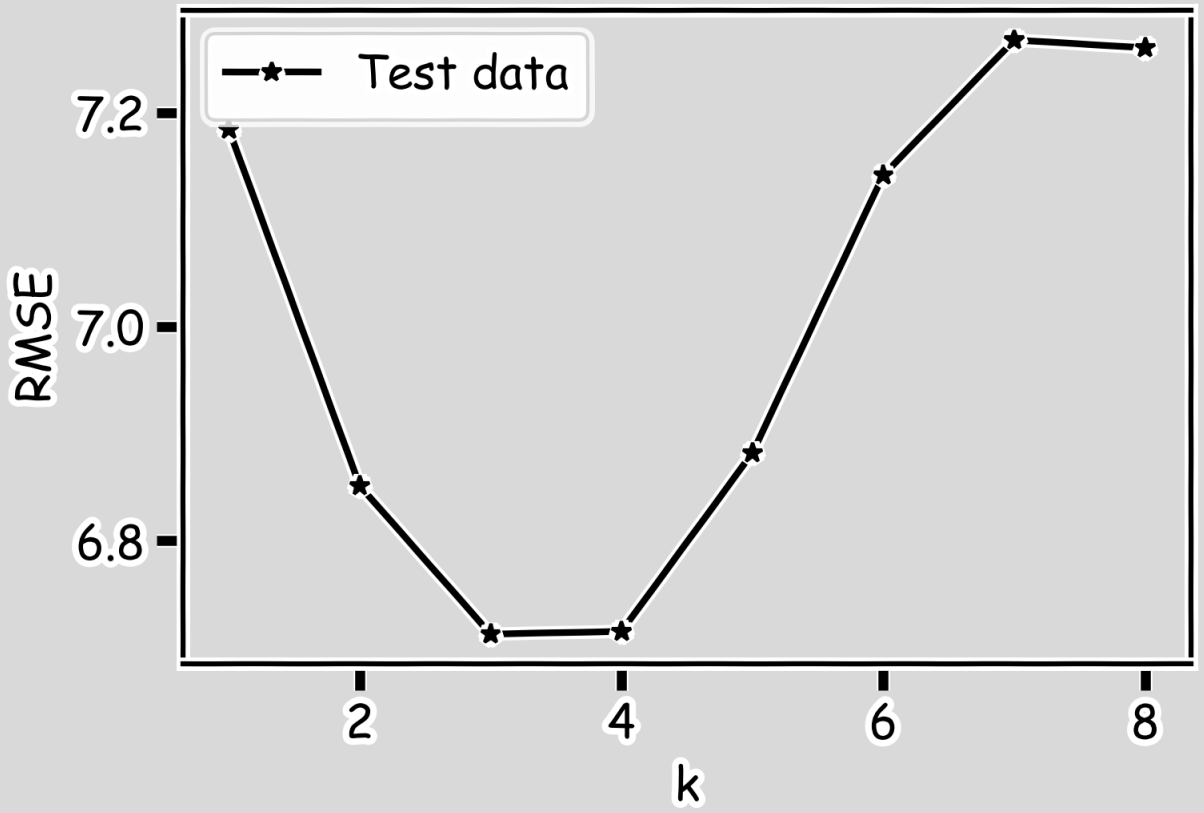
Question: What would be an intuitive loss function for predicting categorical outcomes?

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

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

Model Comparison

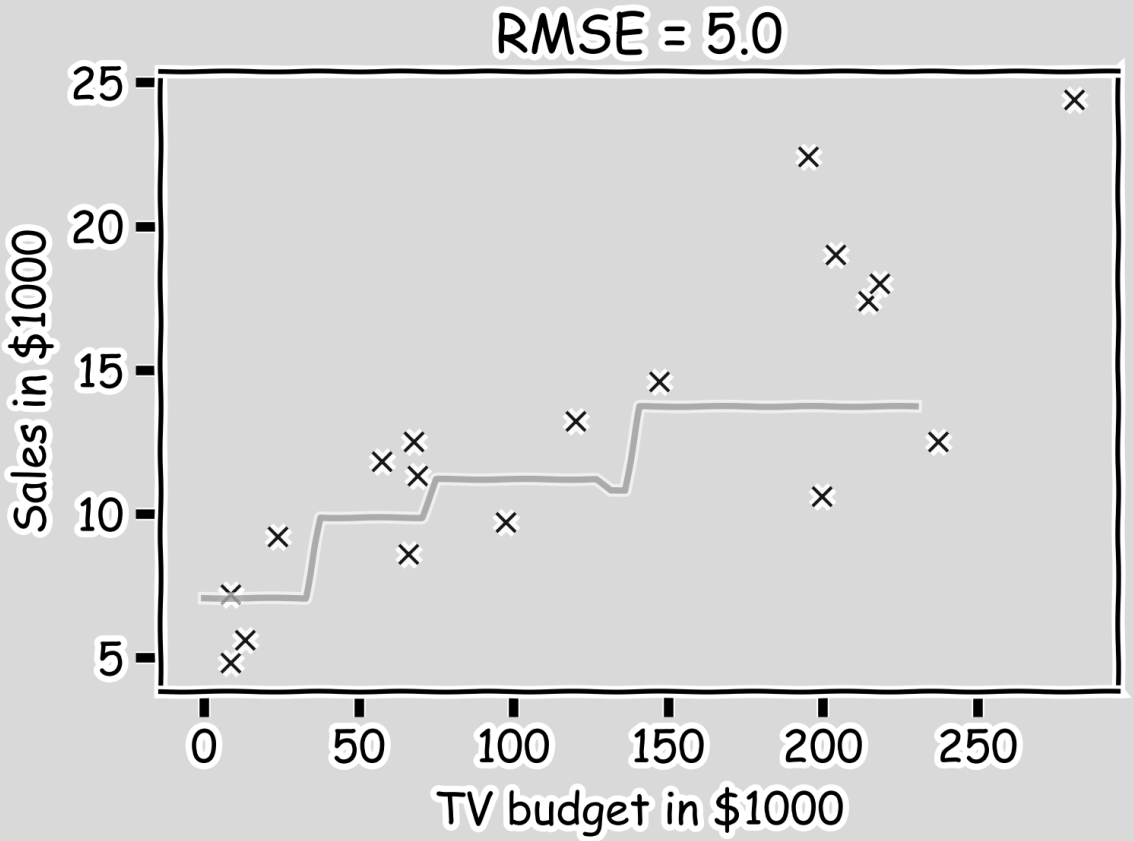
If we compare the results for different values of *k* based on RMSEs, *k*=3 seems to be the best model.



Model Fitness

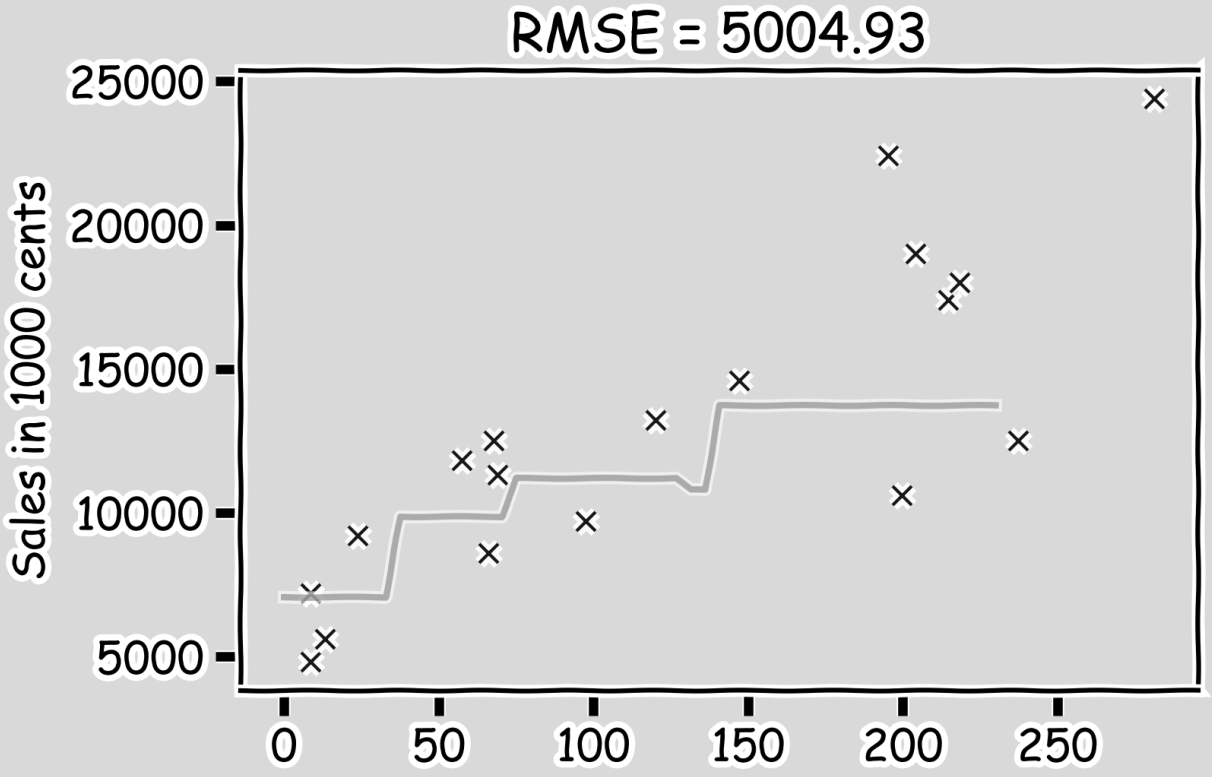
How well did the “best” model do?

For our test data the RMSE=5.0 for *k*=3. Is that good enough?



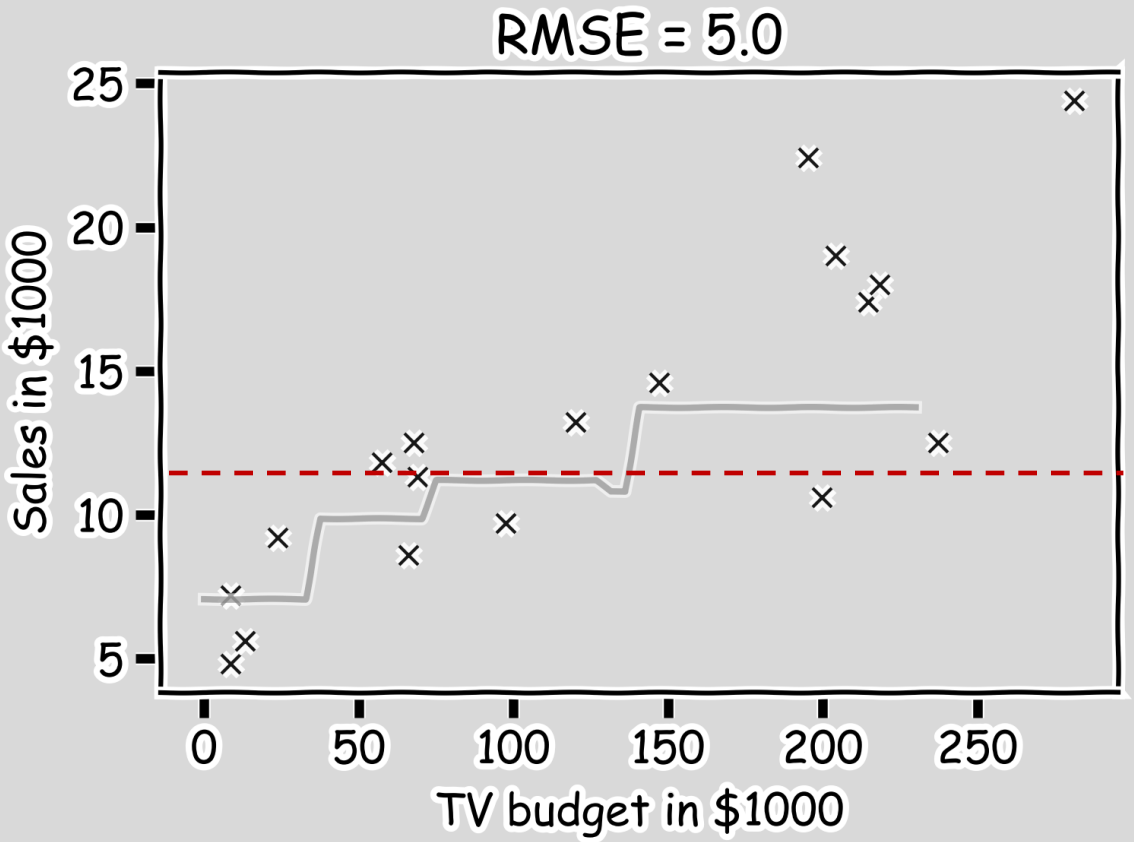
Model Fitness

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



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

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

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

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 $\hat{f}$

Future lecture

Questions?