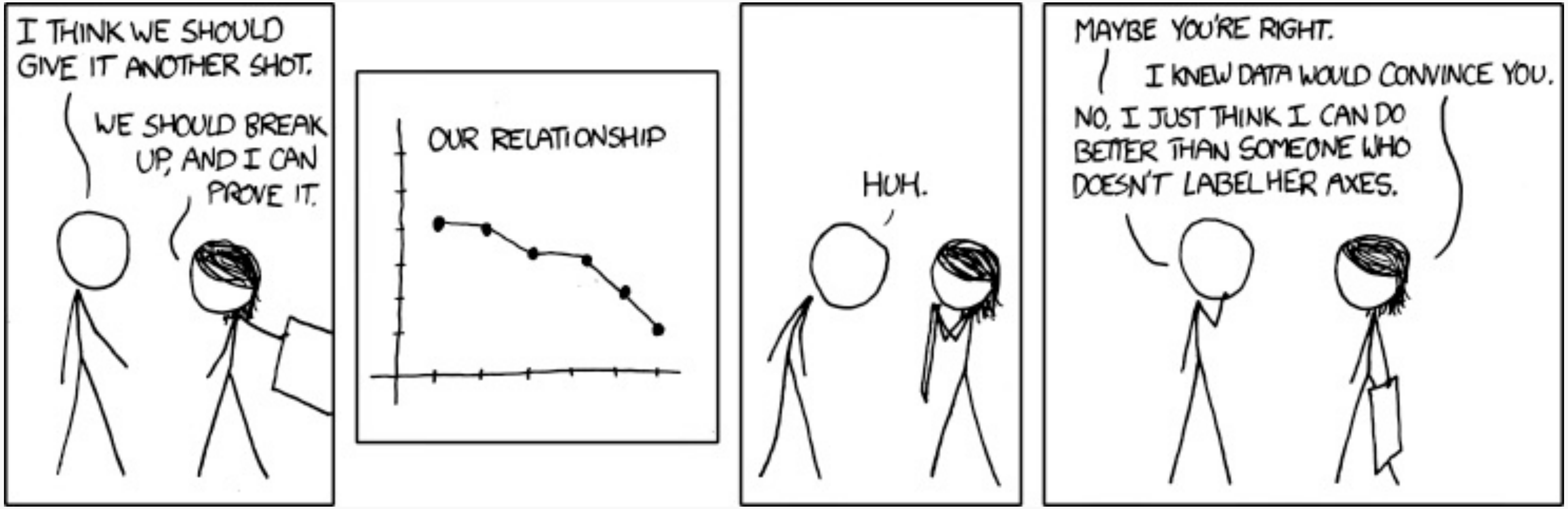
Introduction to Regression

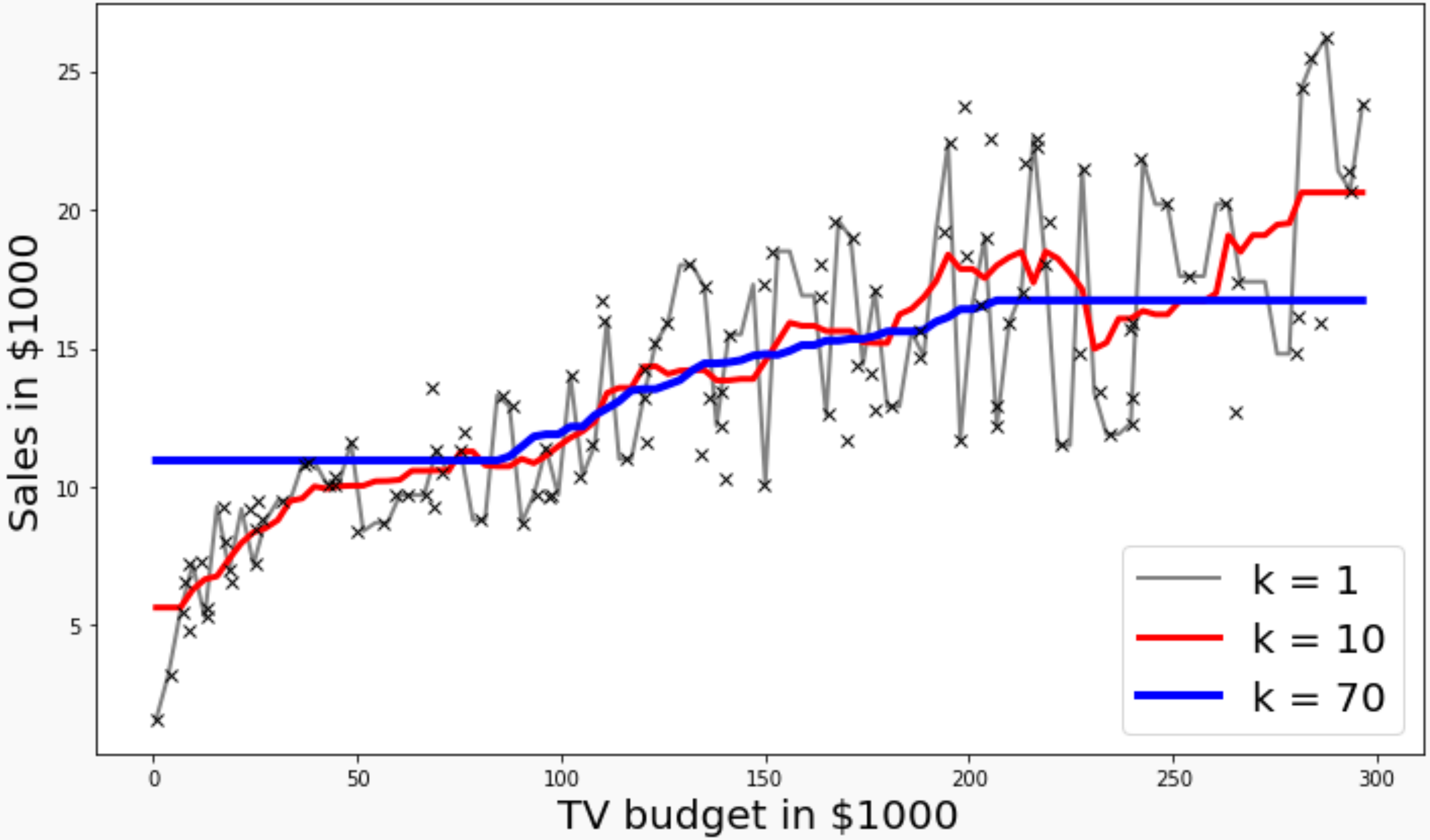
Part B: Error Evaluation and Model Comparison

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai



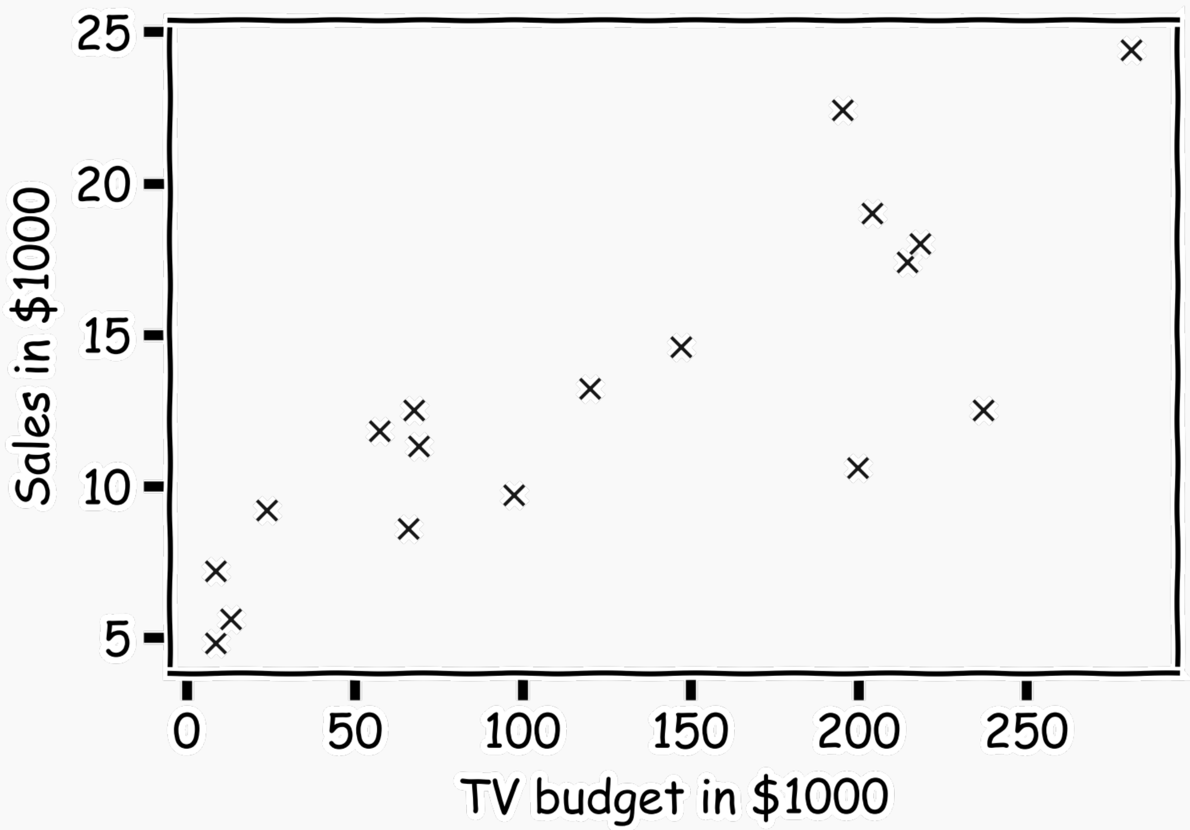
https://xkcd.com/833/


Sales in $1000
TV budget in $1000
k = 1
k = 10
k = 70
25
20
15
10
5
0
50
100
150
200
250
300

Error Evaluation

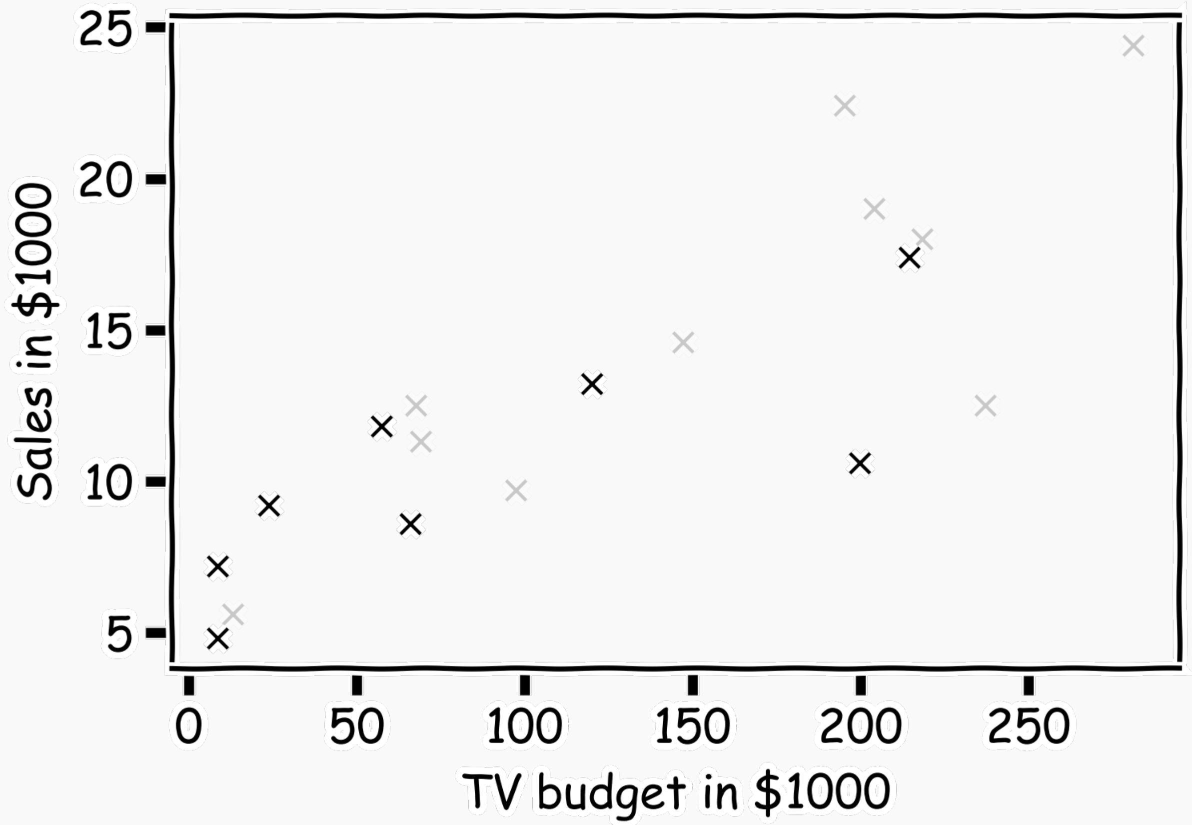
Error Evaluation

Start with some data.



Error Evaluation

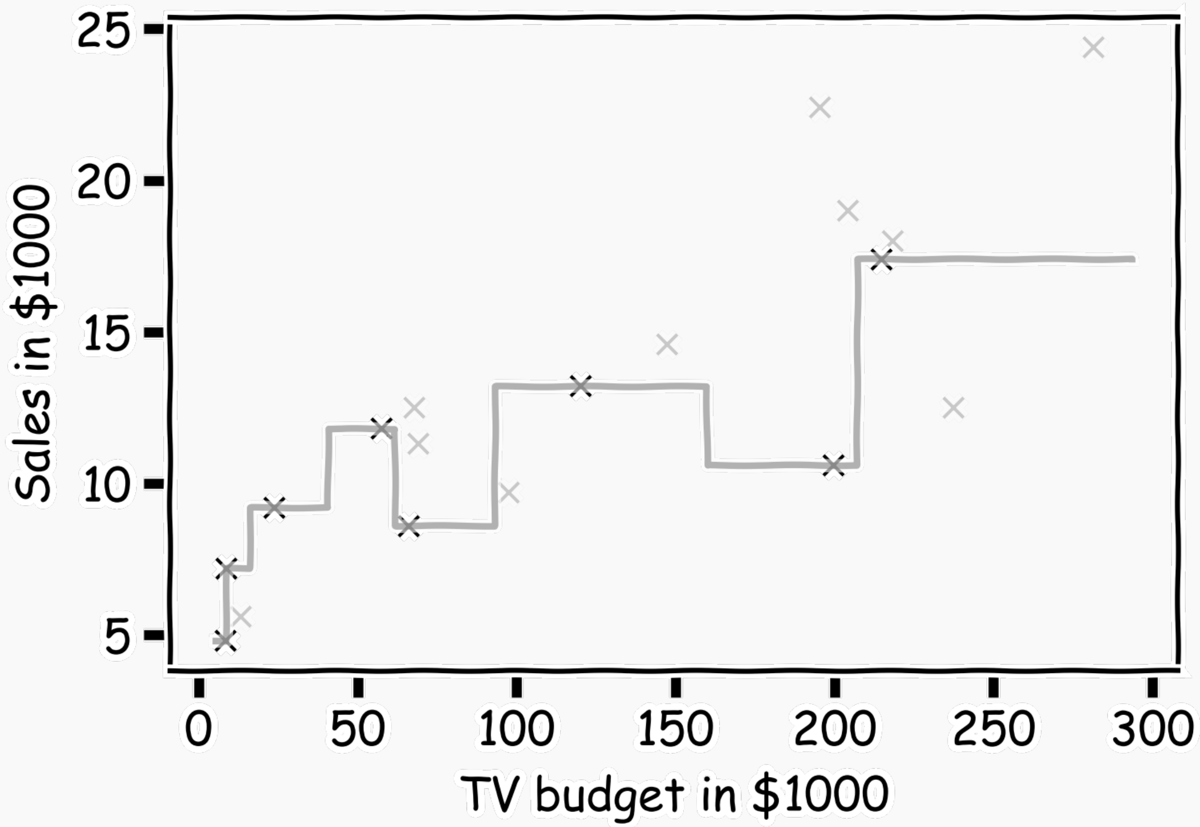
Hide some of the data from the model. This is called **train-test** split.



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

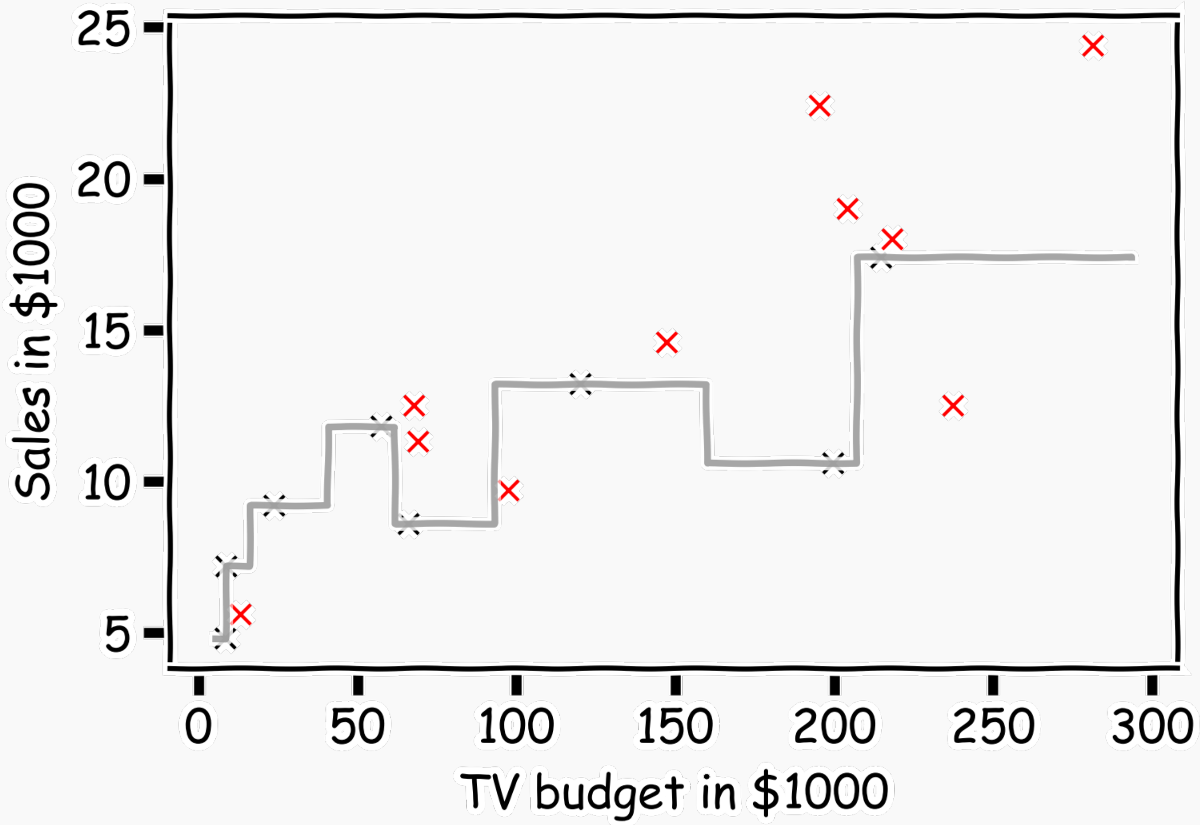
Error Evaluation

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



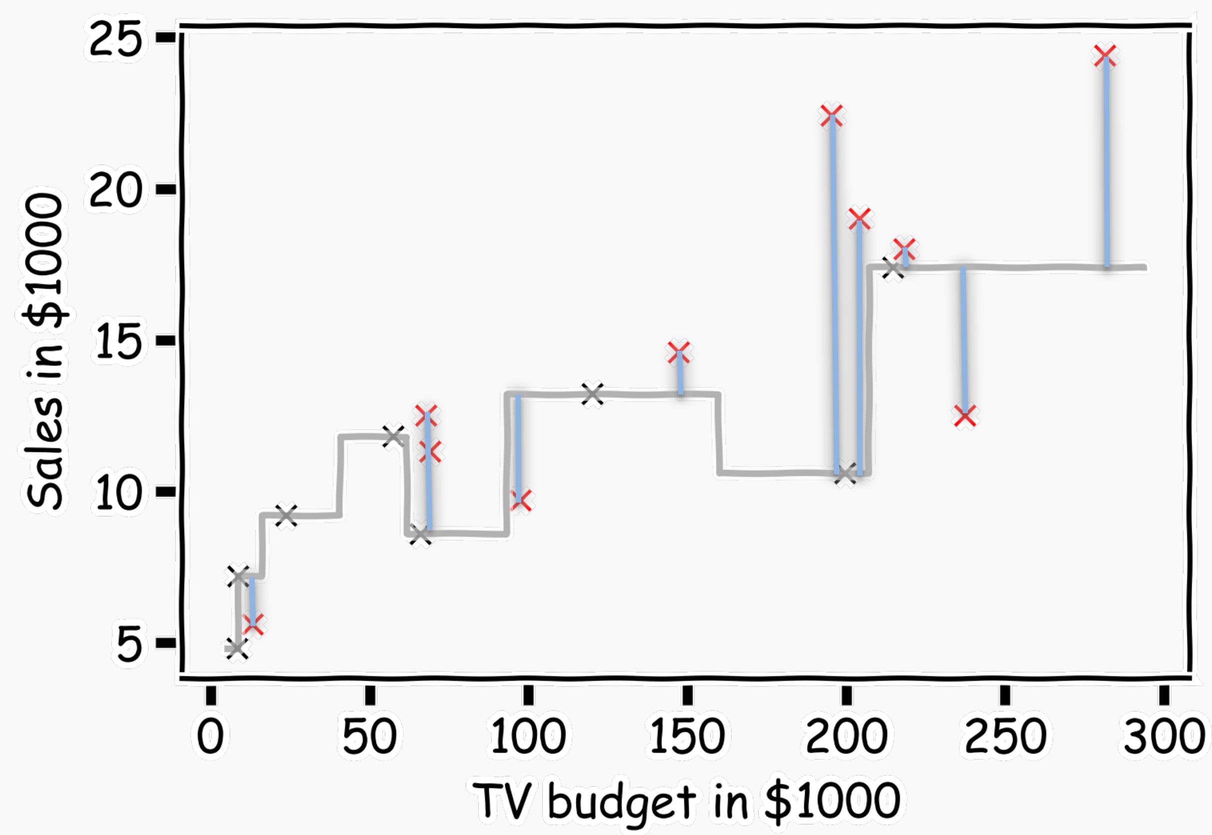
Error Evaluation

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



Error Evaluation

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



For each observation (x_n, y_n), the absolute residuals, $r_i = |y_i - \hat{y}_i|$ quantify the error at each observation.

Error Evaluation

In order to quantify how well a model performs, we aggregate the errors, and we call that the ***loss*** or ***error*** *or* ***cost function.***

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

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

Note: Loss and cost function refer to the same thing. Cost usually refers to the total loss where loss refers to a single training point.

Error Evaluation

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

Other choices for loss function:

1. Max Absolute Error
2. Mean Absolute Error
3. Mean Squared Error

We will motivate MSE when we introduce probabilistic modeling.

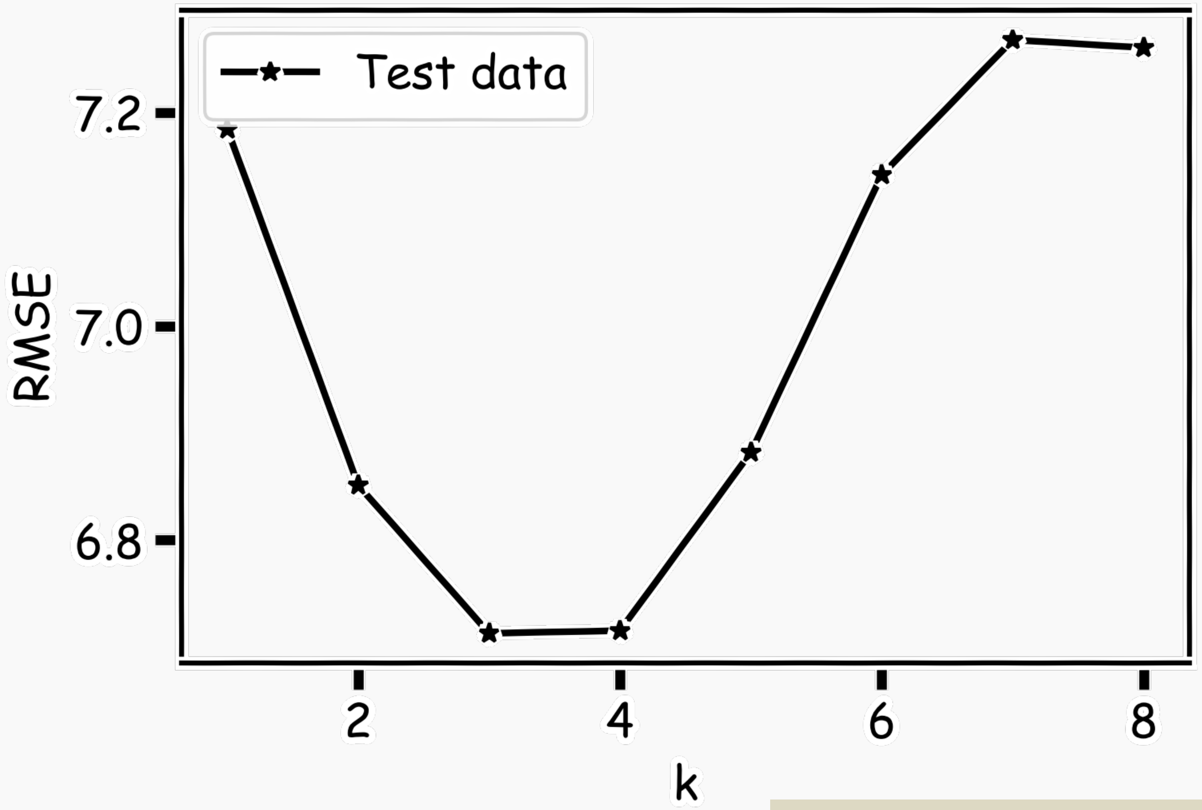
Note: The square **R**oot of the **M**ean of the **S**quared **E**rrors (RMSE) is also commonly used.

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

Model Comparison

Model Comparison

Do the same for all *k*'s and compare the RMSEs.



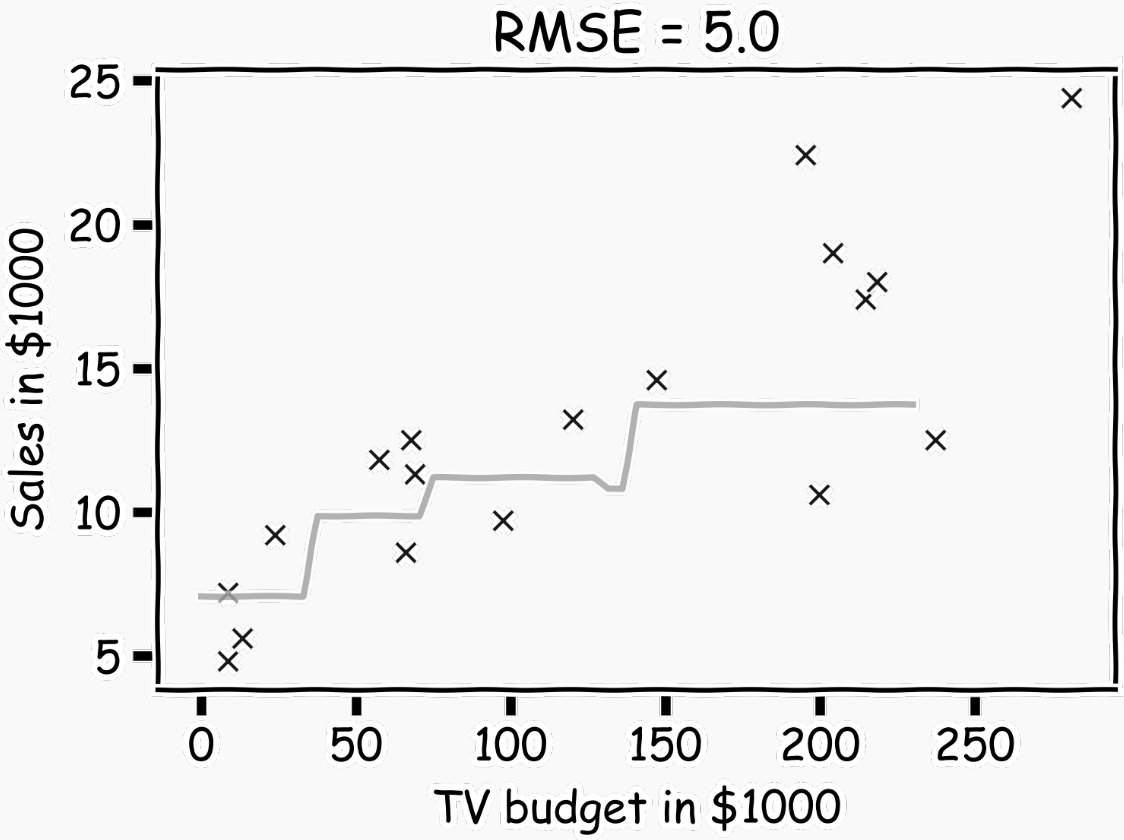
Which model is the best?

k=3 seems to be the **best model**.

Model Fitness

Model fitness

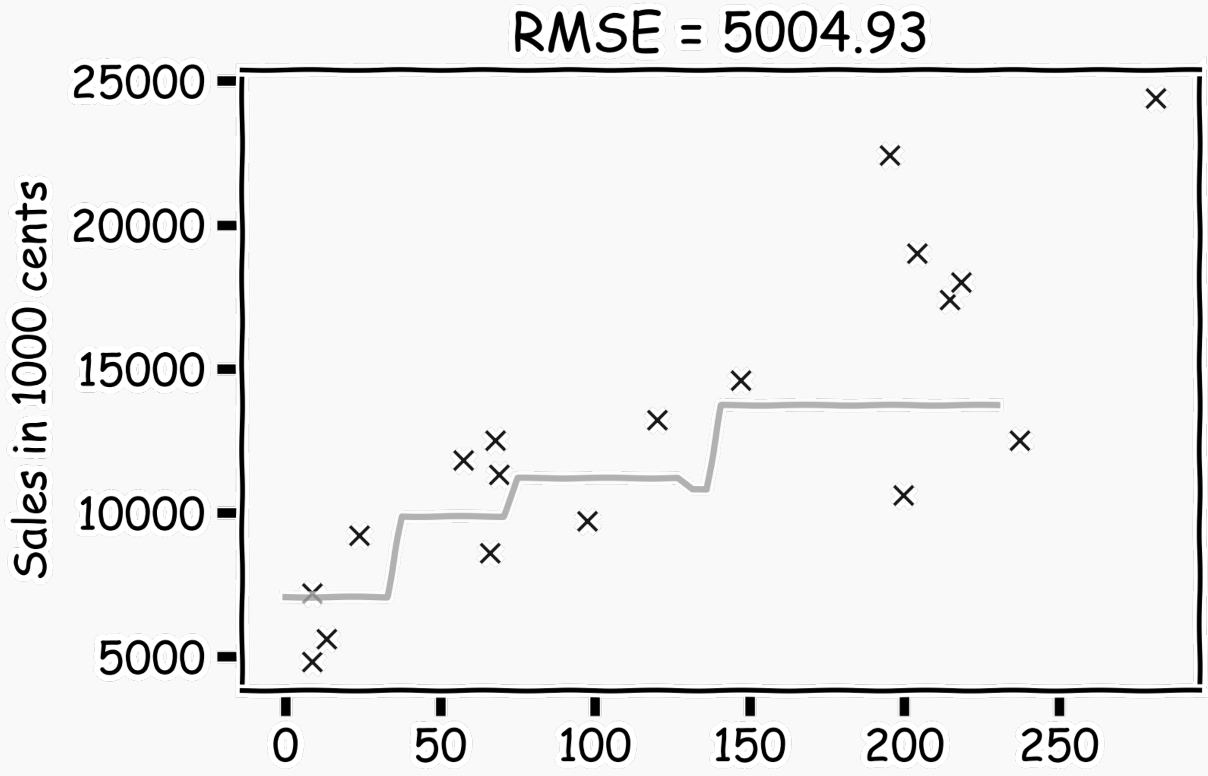
For a subset of the data, calculate the RMSE for *k=3*.



Is RMSE=5.0 good enough?

Model fitness

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

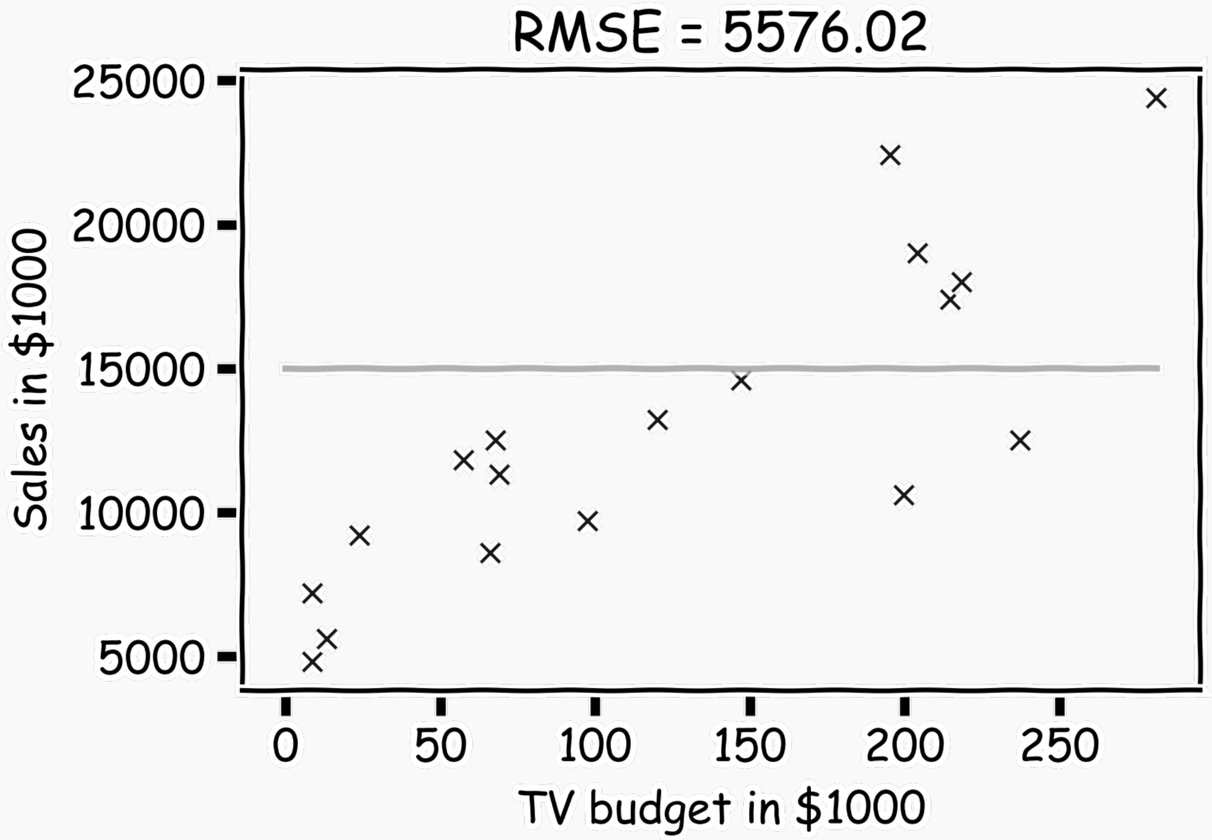


RMSE is now 5004.93.

Is that good?

Model fitness

It is better if we compare it to something.



We will use the simplest model:

$$\hat{y} = \bar{y} = \frac{1}{n}\sum_i y_i$$

as the worst possible model and

$$\hat{y_i} = y_i$$

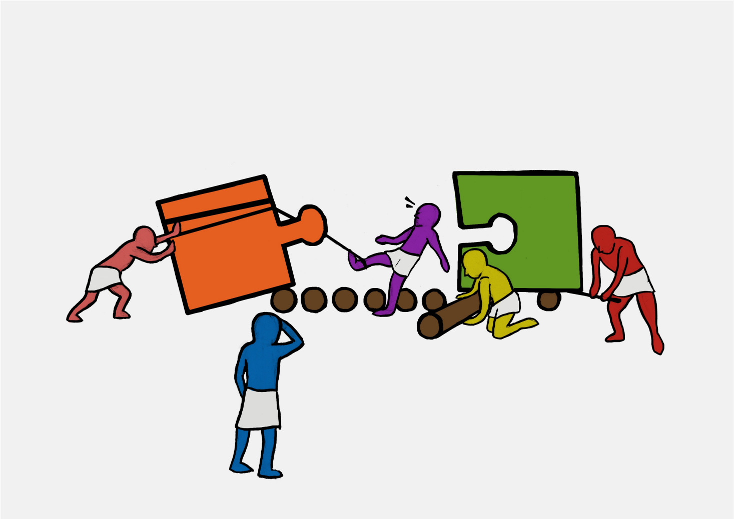
as the best possible model.

R-squared

Though is called R-squared, it is not the square of R

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

- 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.



🧑‍🏫 Use the loss to do model selection (10 min)

🧑‍🏫 Exercise: Finding the Best k in kNN Regression

The goal here is to **find the value of k of the best performing model** based on the test MSE.