Model Selection

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai

Outline

- Announcements
- Q&A from lecture 4
- Model Selection
    - Using Validation
    - Using Cross Validation

Outline

- **Announcements**
- Q&A from lecture 4
- Model Selection
  - Using Validation
  - Using Cross Validation

Announcements

- Homework 2 is due 9/29
- We will have one less homework!
- Playlist – send me your playlist and I will play it before and during exercises.

Outline

- Announcements
- **Q&A from lecture 4**
- Model Selection
  - Using Validation
  - Using Cross Validation

Are complex models always better?

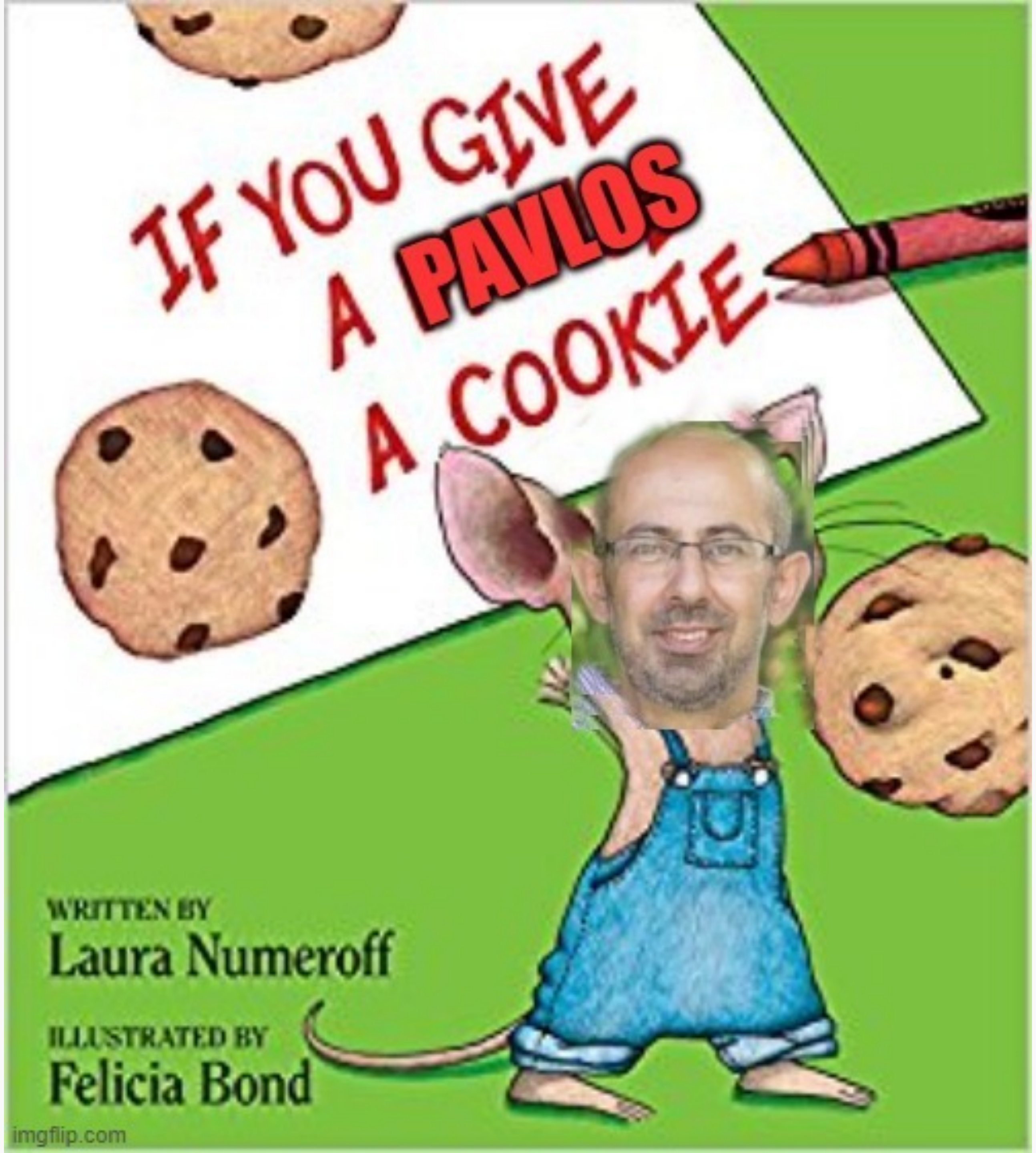
Can we have negative polynomials?

Among 2 collinear predictors, which one do we keep?

When do we need to apply the feature scaling?
What is the difference between underfitting vs overfitting with different curves?
When do we say our model is overfit?
Could you outline the difference between scaling, normalizing, and standardizing?
Be happy to do so.
Can we scale categorical predictors?

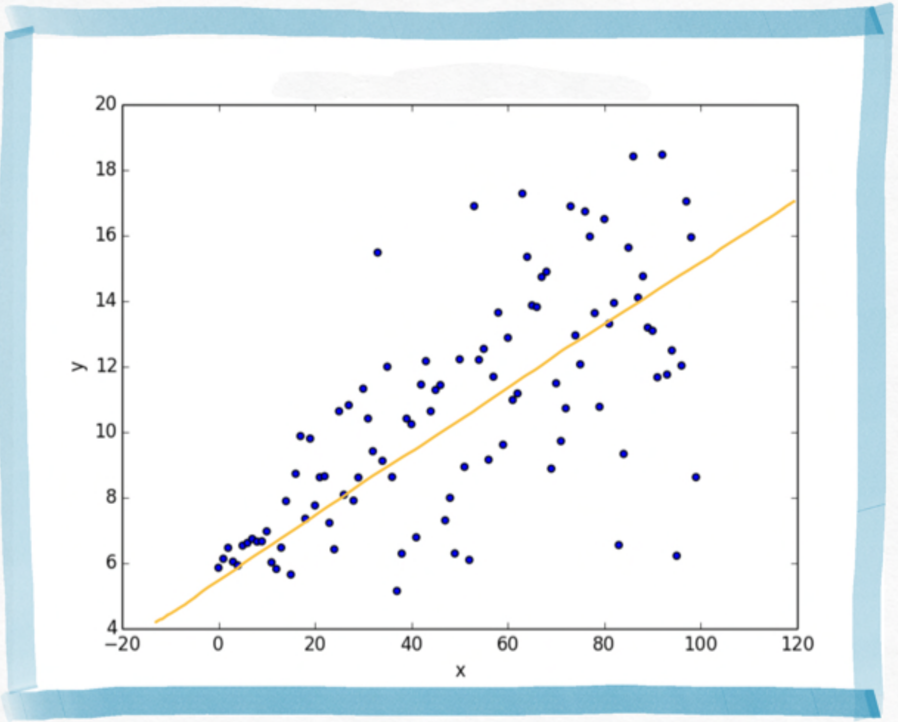
How do we know when to use scaling, normalizing, and standardizing?

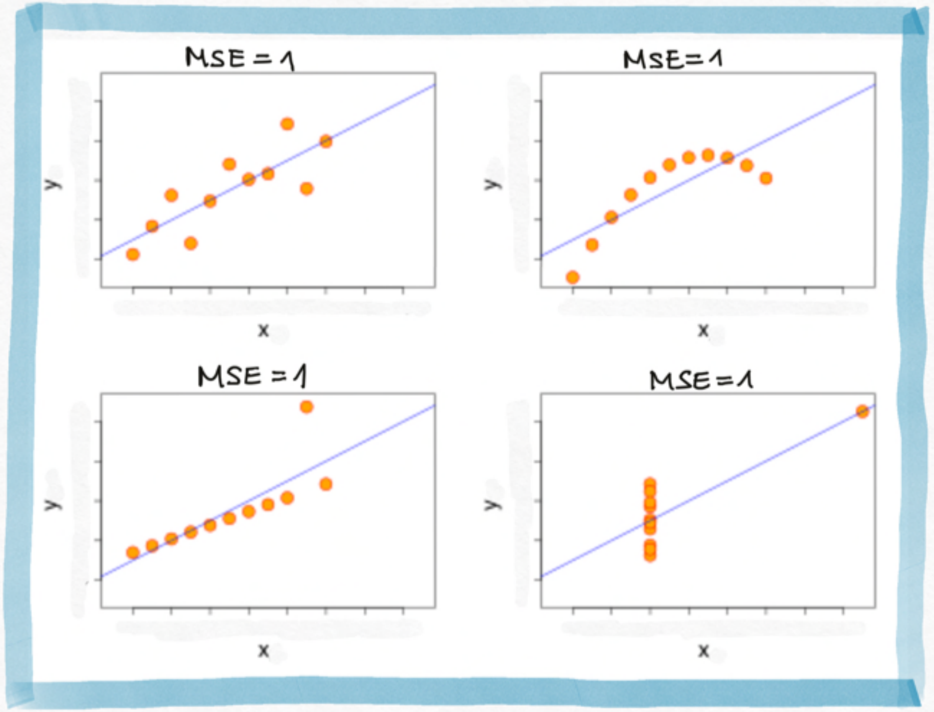
Why we are learning regression when there are built in functions that calculate the line of best fit?


IF YOU GIVE
A PAVLOS
A COOKIE
WRITTEN BY
Laura Numeroff
ILLUSTRATED BY
Felicia Bond
imgflip.com

Evaluation: Training Error

Just because we found the model that minimizes the squared error it doesn't mean that it's a good model. We investigate the R2 but also:



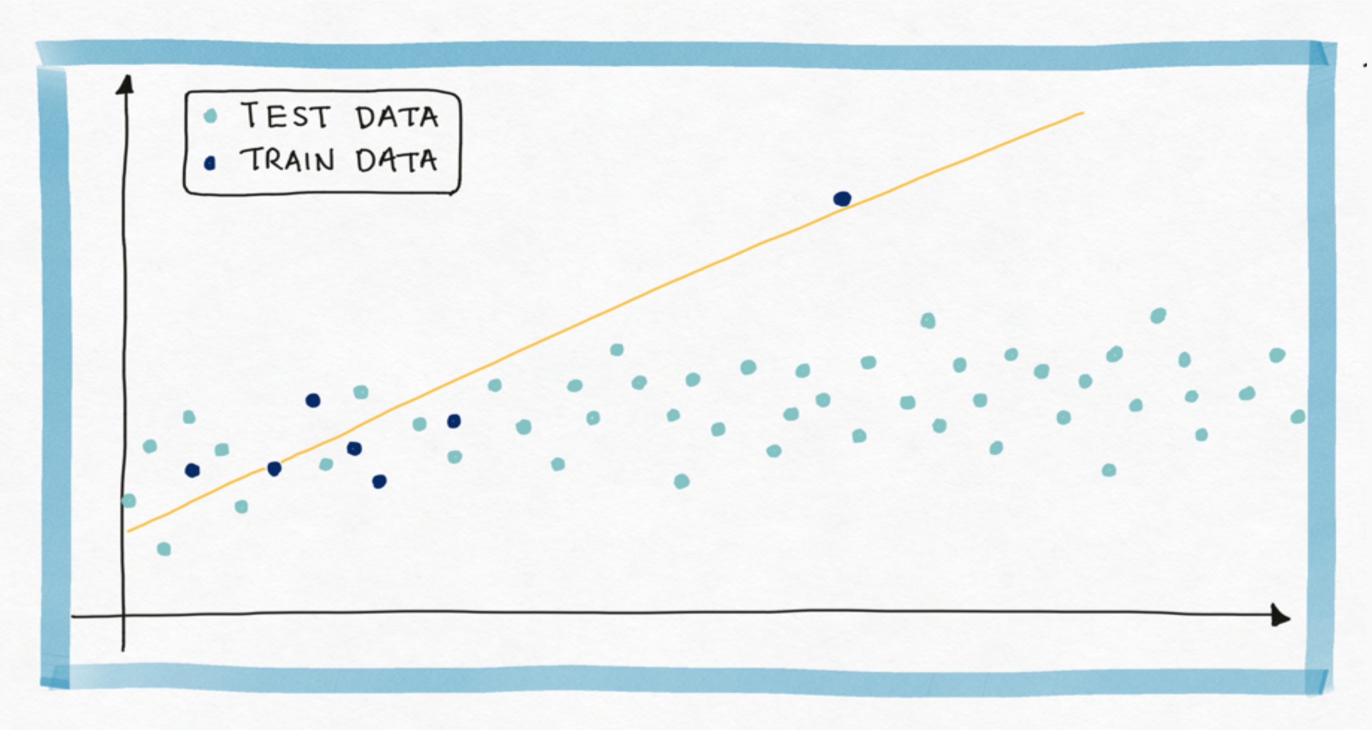


The MSE is high due to noise in the data.

The MSE is high in all four models, but the models are not equal.

Evaluation: Test Error

We need to evaluate the fitted model on new data, data that the model did not train on, the test data.



The training MSE here is 2.0 where the test MSE is 12.3.

The training data contains a strange point – an outlier – which confuses the model.

Fitting to meaningless patterns in the training is called **overfitting.**

Generalization Error

We know to evaluate the model on both train and test data, because models that do well on training data may do poorly on new data (overfitting).

The ability of models to do well on new data is called generalization.

The goal of model selection is to choose the model that generalizes the best.

Model Selection

Model selection is the application of a principled method to determine the complexity of the model, e.g., choosing a subset of predictors, choosing the degree of the polynomial model etc.

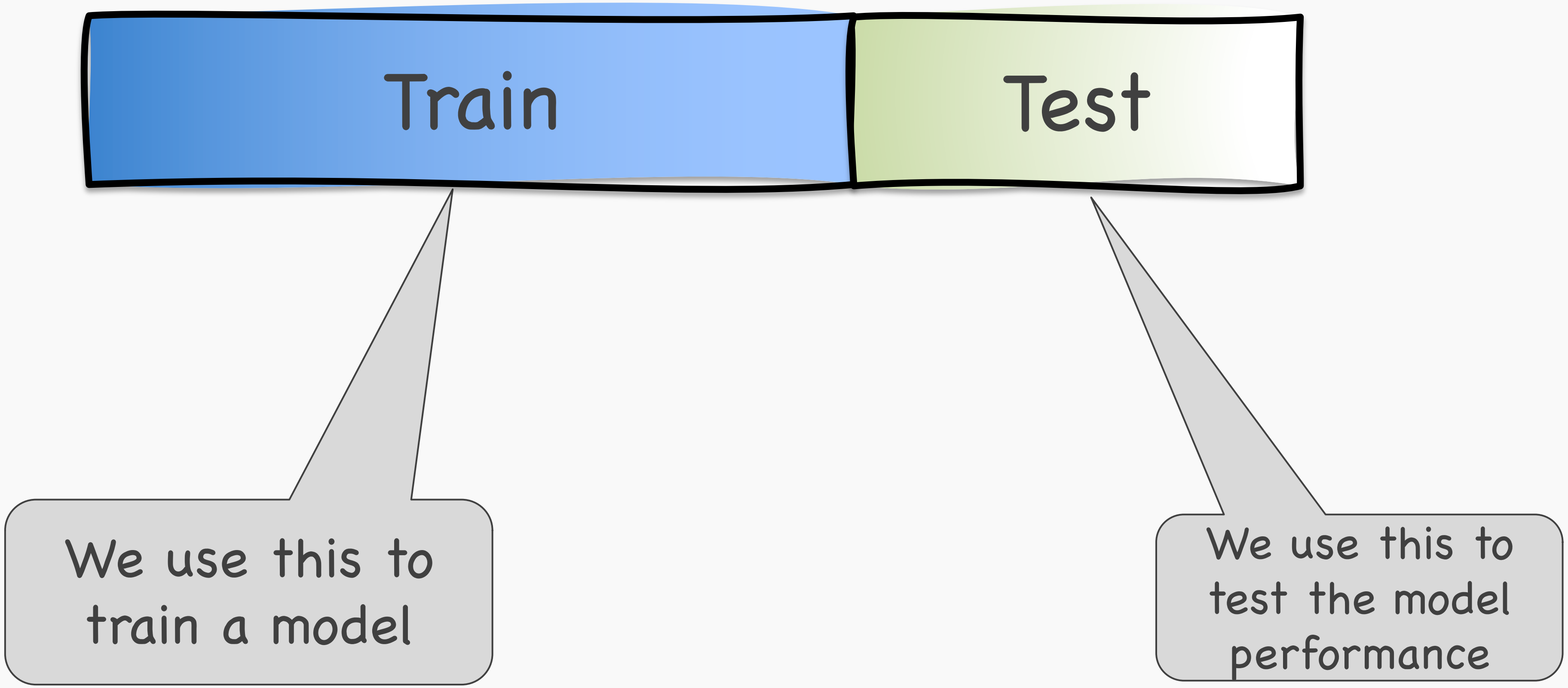
A strong motivation for performing model selection is to avoid **overfitting**, which we saw can happen when:

- there are too many predictors:
  - the feature space has high dimensionality
  - the polynomial degree is too high
  - too many cross terms are considered
- the coefficients values are too **extreme (we have not seen this yet)**

Train-Test split

How do we select a model?

So far, we have been using train/test splits

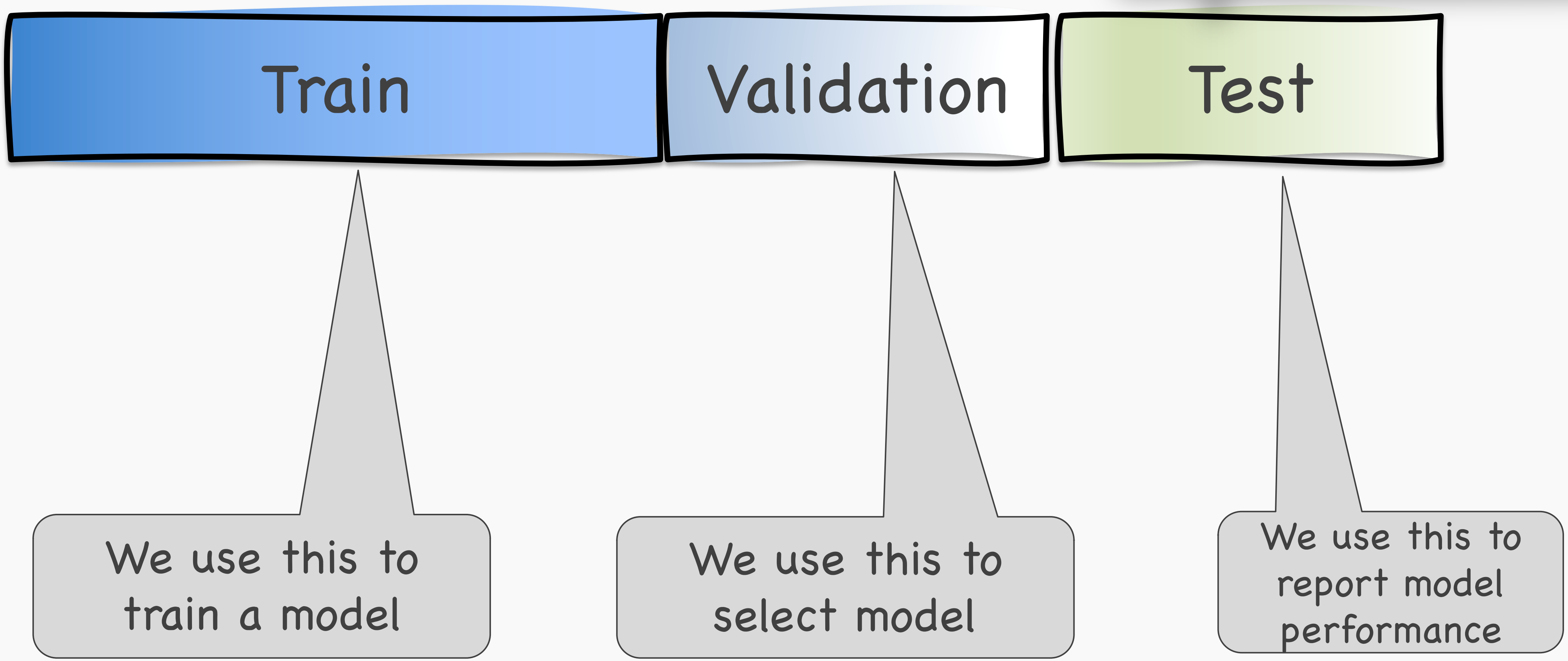


Train-Validation-Test

We introduce a different sub-set, which we called validat

select the model.

The test set should never be touched for model training or selection.



Model Selection

Ways of model selection:

- Exhaustive search
- Greedy algorithms
- Fine tuning hyper-parameters
- Regularization

Model Selection

Ways of model selection:

- **Exhaustive search**
- Greedy algorithms
- Fine tuning hyper-parameters
- Regularization

Model Selection: How many models?

Can you prove this?

?

Question:

How many different models when considering J predictors (only linear terms) do we have?

Example: 3 predictors (X_1, X_2, X_3)

- Models with 0 predictor:

  M0:

- Models with 1 predictor:

  M1: X_1

  M2: X_2

  M3: X_3

- Models with 2 predictors:

  M4: $\{X_1, X_2\}$

  M5: $\{X_2, X_3\}$

  M6: $\{X_3, X_1\}$

- Models with 3 predictors:

  M7: $\{X_1, X_2, X_3\}$

2^J models

Model Selection

Ways of model selection:

- Exhaustive search
- **Greedy algorithms**
- Fine tuning hyper-parameters
- Regularization

Stepwise Variable Selection and Validation

Selecting optimal subsets of predictors (including choosing the degree of polynomial models) through:

- stepwise variable selection - iteratively building an optimal subset of predictors by optimizing a fixed model evaluation metric each time.

- selecting an optimal model by evaluating each model on validation set.

Stepwise Variable Selection: Forward method

In **forward selection**, we find an 'optimal' set of predictors by iterative building up our set.

1. Start with the empty set P0, construct the null modelM0.

2. For $k = 1, \ldots, J$:

2.1 Let M_{k-1} be the model constructed from the best set of $k-1$ predictors, P_{k-1}.

2.2 Select the predictor X_{n_k}, not in P_{k-1}, so that the model constructed from $P_k = X_{n_k} \cup P_{k-1}$ optimizes a fixed metric (this can be p -value, F -stat; validation MSE, R^2, or AIC/BIC on training set).

2.3 Let M_k denote the model constructed from the optimal P_k.

3. Select the model M amongst $\{M_0, M_1, \ldots, M_J\}$ that optimizes a fixed metric (this can be validation MSE, R^2; or AIC/BIS on training set)

Stepwise Variable Selection Computational Complexity

How many models did we evaluate?

- 1st step, ***J* Models**
- 2nd step, *J-1* **Models** (add 1 predictor out of *J-1* possible)
- 3rd step, *J-2* **Models** (add 1 predictor out of *J-2* possible)
- …

$$O(J^2) \ll 2^J \text{ for large } J$$

Model Selection

Ways of model selection:

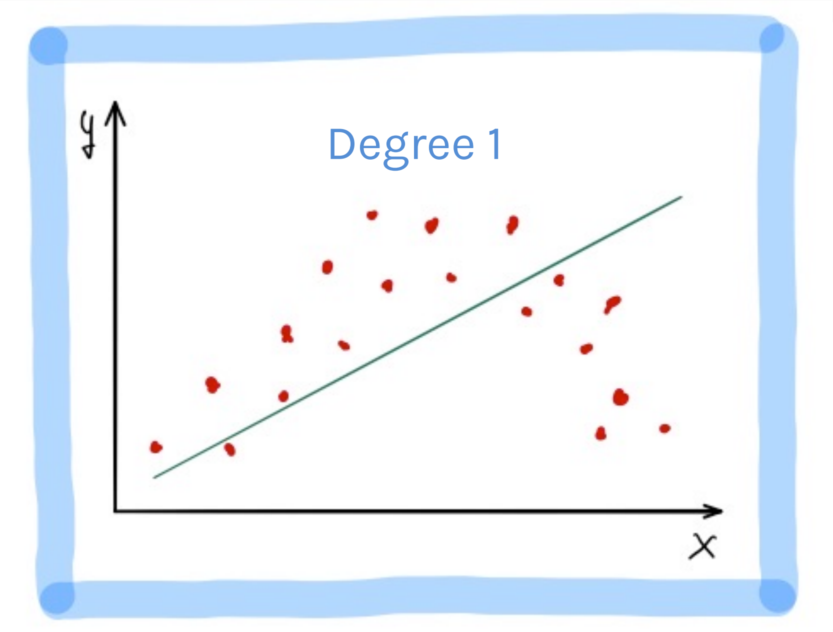
- Exhaustive search
- Greedy algorithms
- **Fine tuning hyper-parameters**
- Regularization

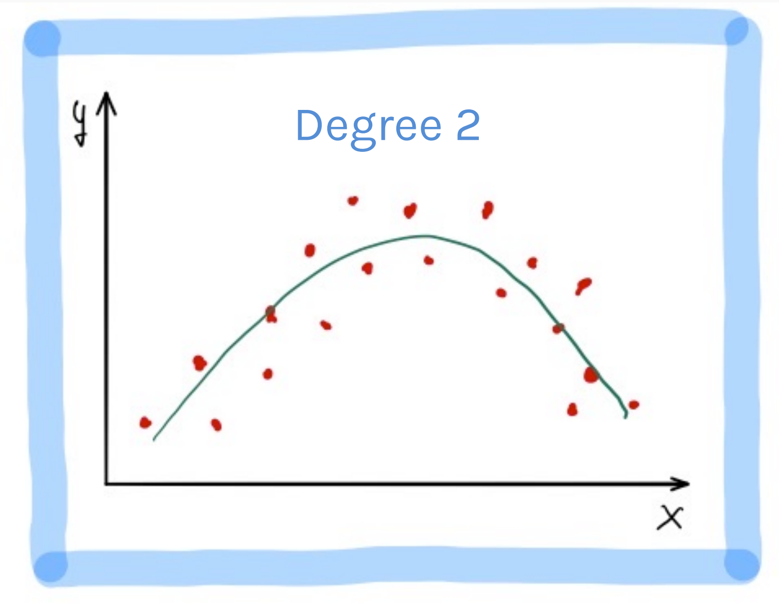
Choosing the degree of the polynomial model

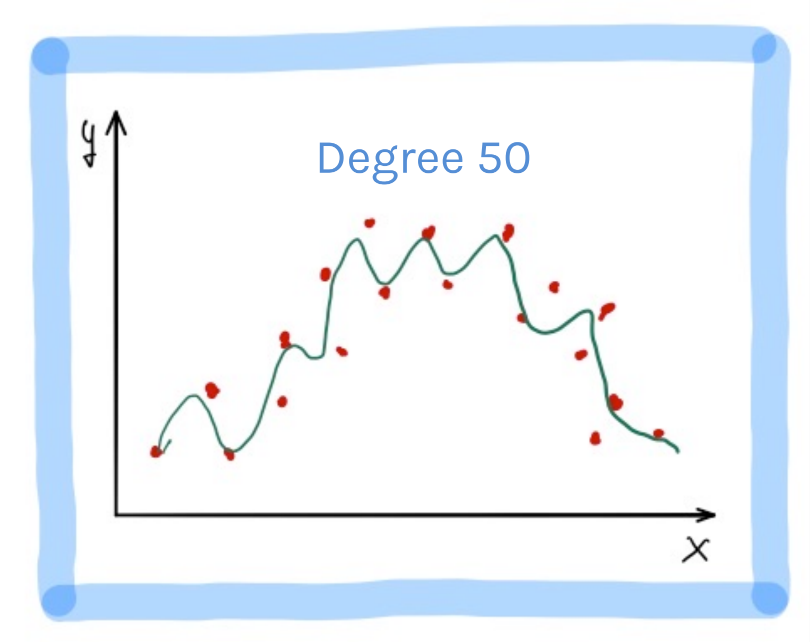
kNN: k was a hyper-parameter

?

We turn model selection into choosing a hyper-parameter. For example, polynomial regression requires choosing a degree – this can be thought as model selection – and we select the model by tuning the hyper-parameter.





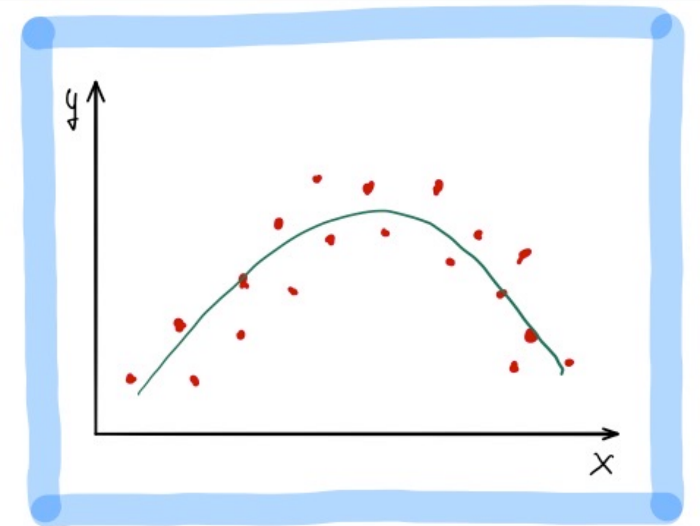


Underfitting: when the degree is too low, the model cannot fit the trend.

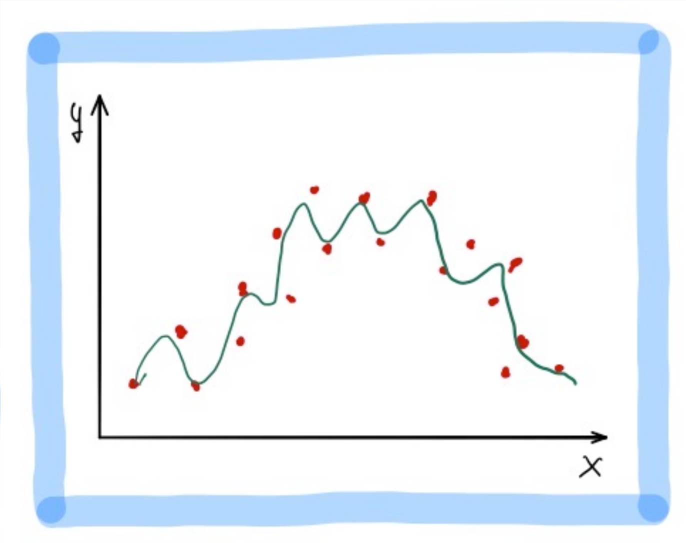
We want a model that fits the trend and ignores the noise.

Overfitting: when the degree is too high, the model fits all the noisy data points.

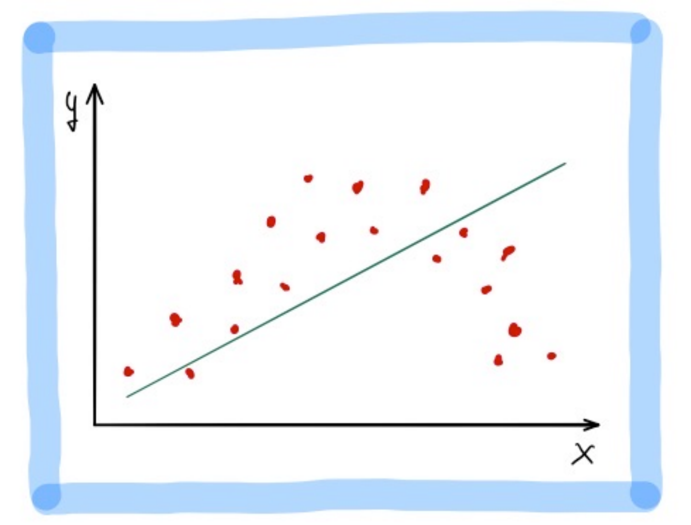
Best model: validation error is minimum.

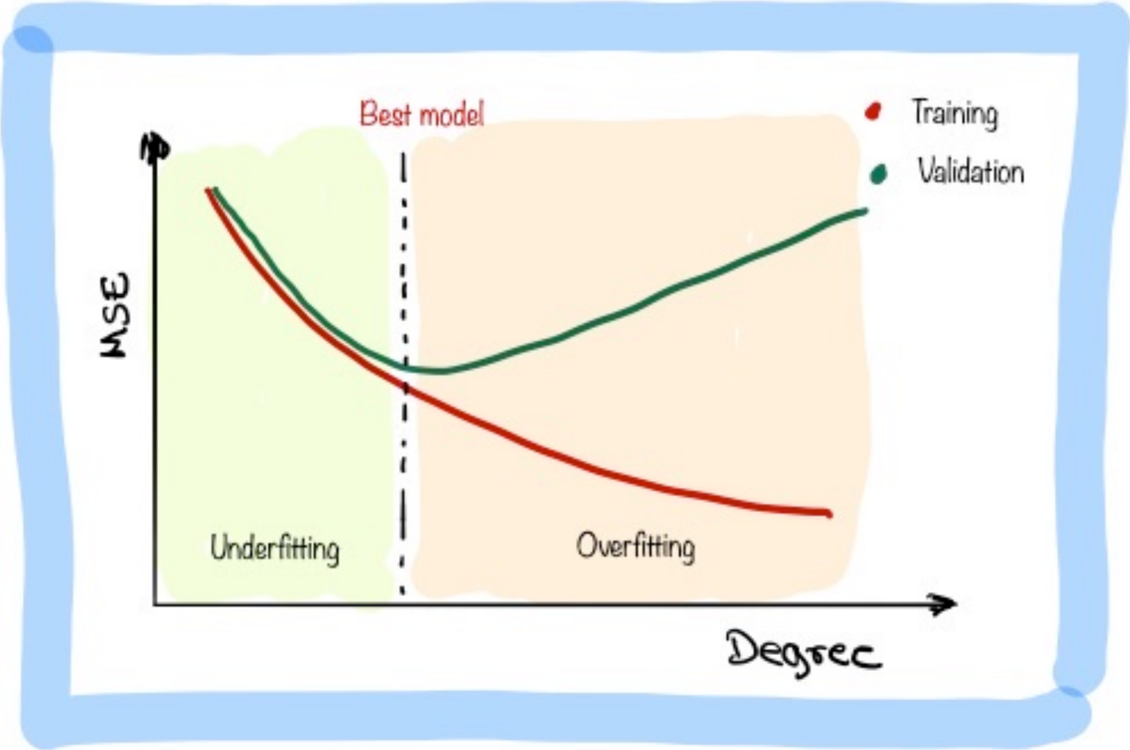


Overfitting: train error is low, validation error is high.



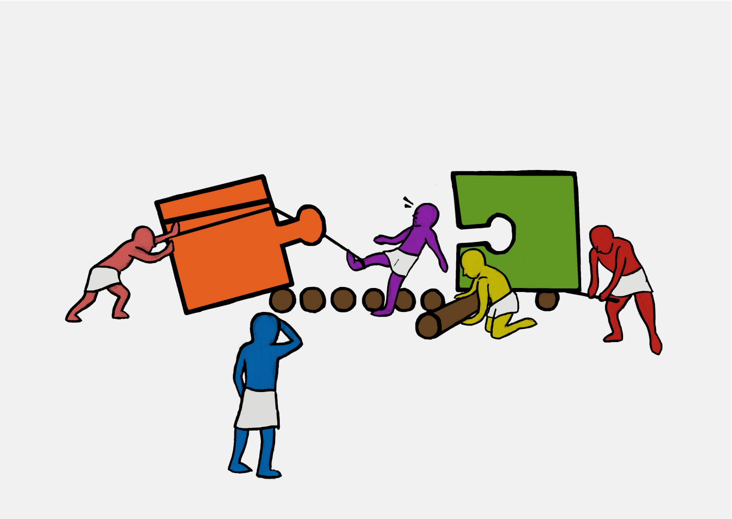
Underfitting: train and validation error is high.





?

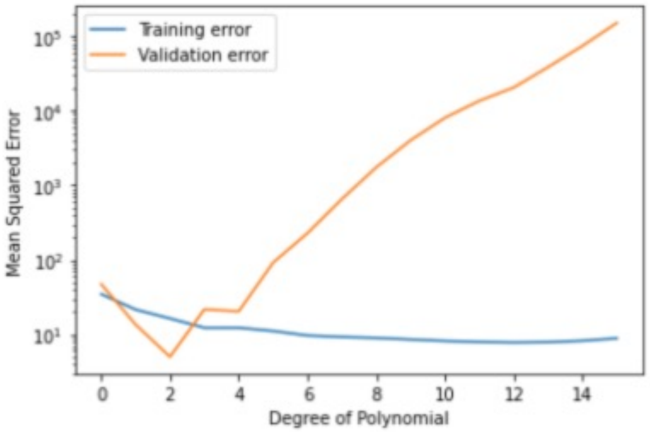
What are the parameters of the models and what are the hyperparameters?



Description

Exercise: Best Degree of Polynomial with Train and Validation sets

The aim of this exercise is to find the **best degree** of polynomial based on the MSE values. Further, plot the train and validation error graphs as a function of degree of the polynomial as shown below.

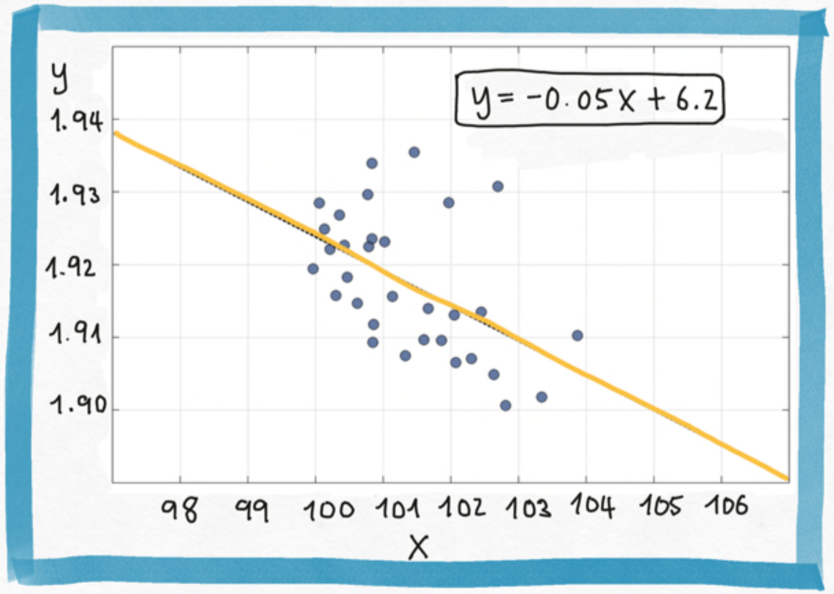


Instructions:

- Read the dataset and split into train and validation sets.
- Select a max degree value for the polynomial model.
- Fit a polynomial regression model on the training data for each degree and predict on the validation data.
- Compute the train and validation error as MSE values and store in separate lists.
- Find out the best degree of the model.
- Plot the train and validation errors for each degree.

Evaluation: Model Interpretation

For linear models it's important to interpret the parameters



The MSE of this model is very small. But the slope is -0.05. That means the larger the budget the less the sales.

The MSE is very small, but the intercept is -0.5 which means that for very small budget we will have negative sales.