# DS 3010

Spring 2026

*Wenting Xu*

IOWA STATE UNIVERSITY

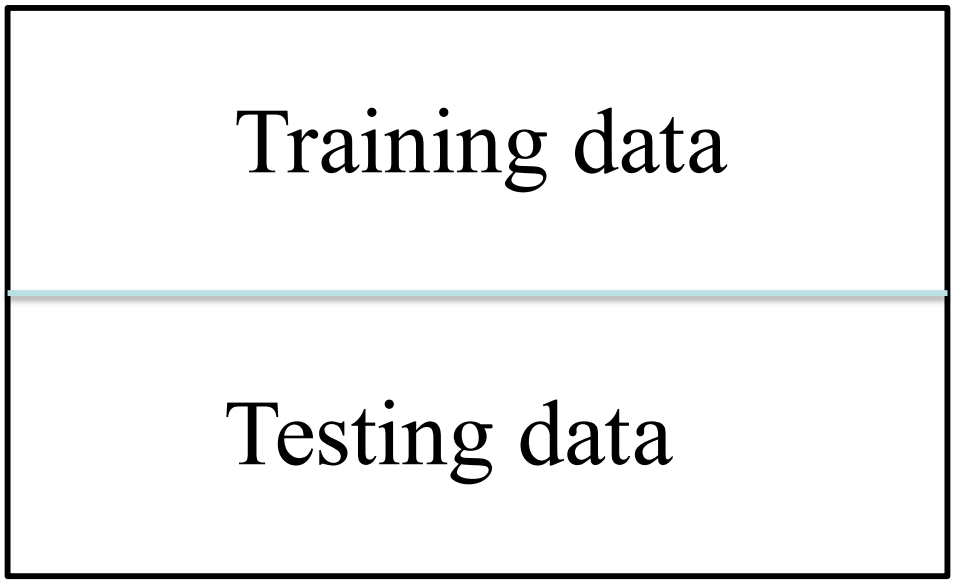
# Module 2: Statistical Decision Theory

## Part 2: Cross Validation

Wenting Xu

Iowa State University

# How have we estimated test MSE so far?



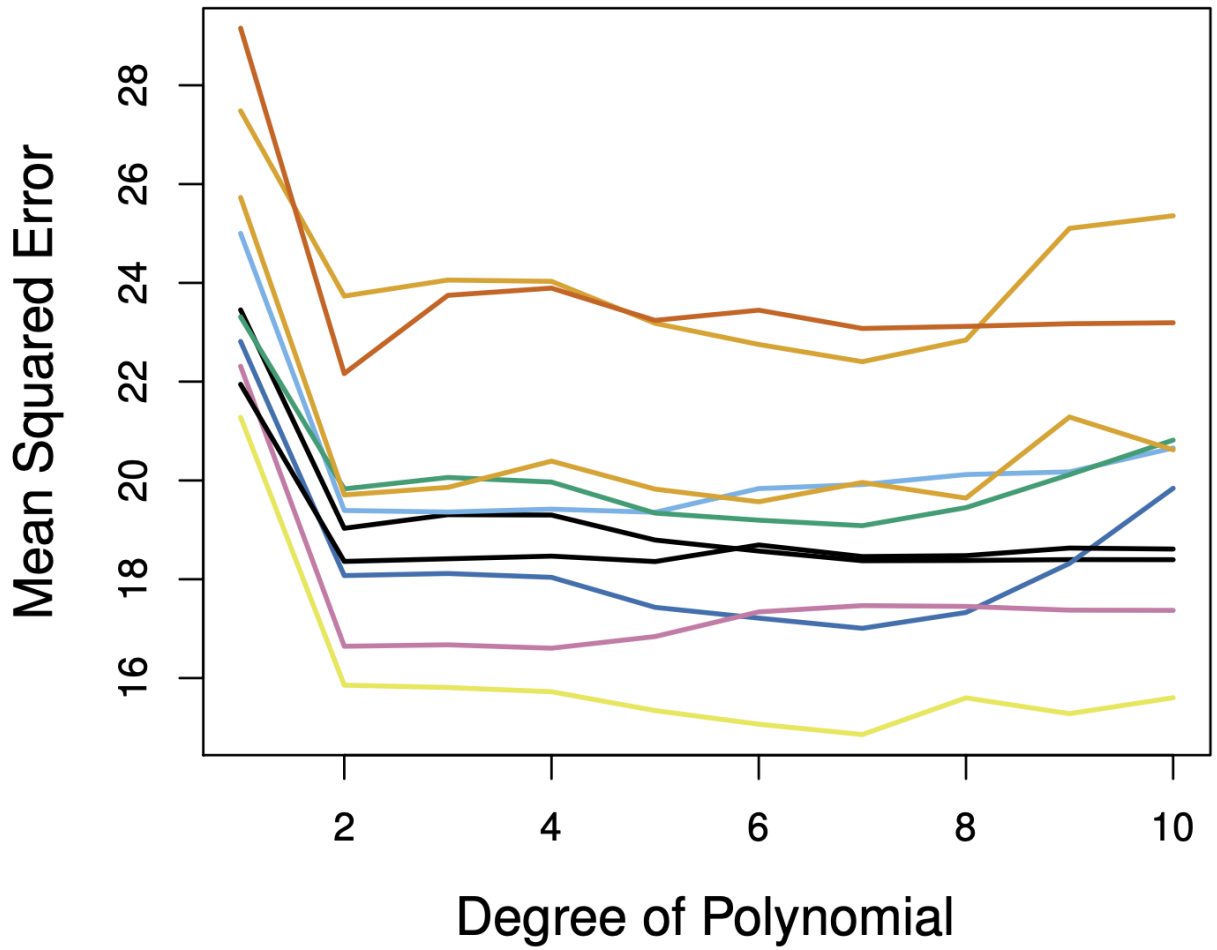

50%

50%

**Validation Set Approach:**

**Randomly** split data into training data (50%) and testing data (50%)

# How have we estimated test MSE so far ?




Problems:

- **Highly variable:** depending on precisely which observations are included in the training set and which observations are included in the validation set.

**Can you think of a better way to create a training and testing dataset ?**

↳ cross validation

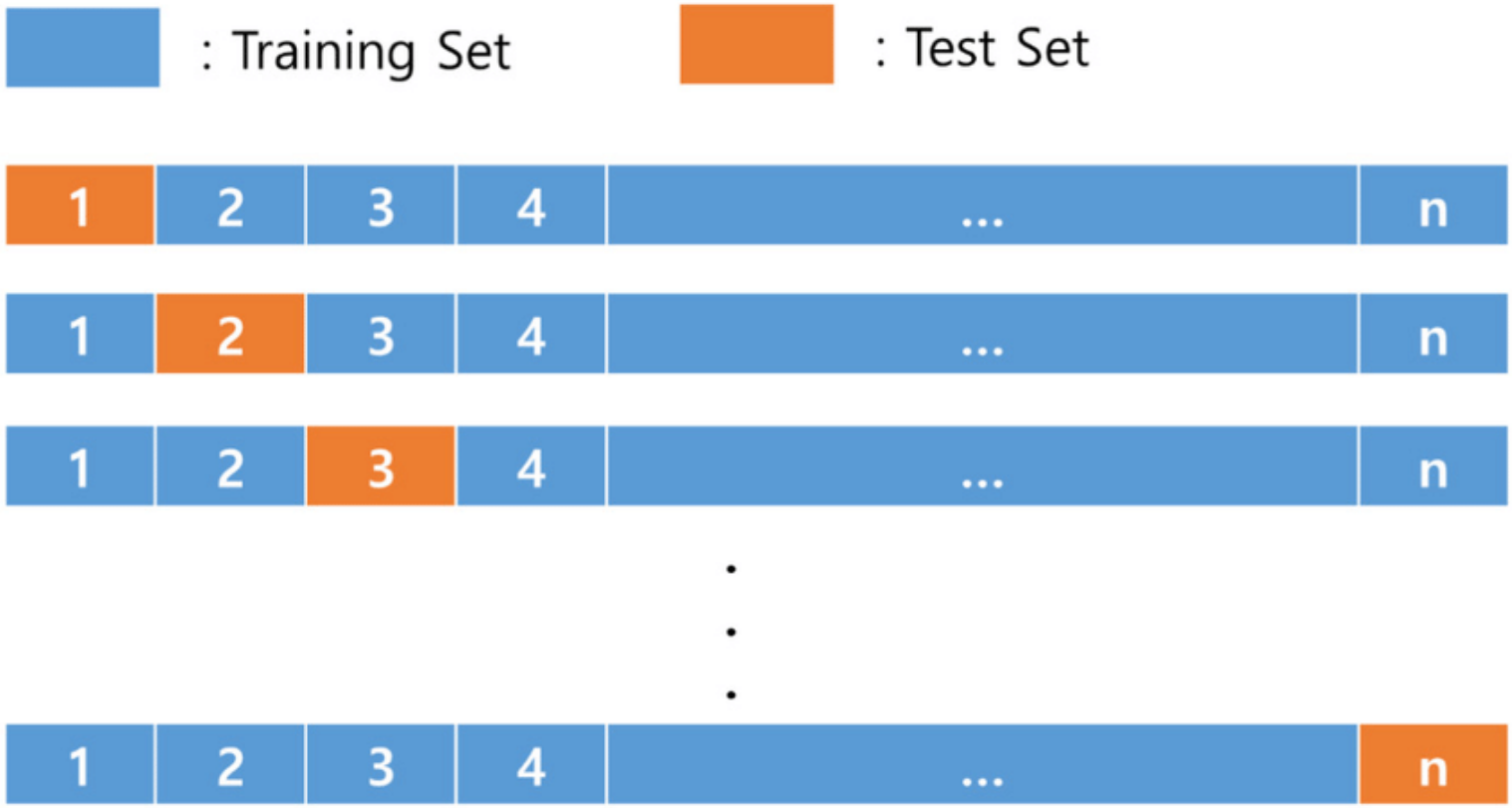
# What is Cross-Validation ?

Cross validation is a technique used in machine learning to evaluate the performance of a model on unseen data. It involves dividing the available data into **multiple folds or subsets**, using **one of these folds** as a validation set, and training the model on the remaining folds.

# Leave-one-out cross-validation (LOOCV)

↳changing what the test set is for each model

n - 1 1




ml

# Leave-one-out cross-validation (LOOCV)

**Testing data:** a single observation $(x_1, y_1)$

**Training data:** observations $\{(x_2,y_2), \ldots, (x_n,y_n)\}$ make up the training set.

$$MSE_1 = (y_1 - \hat{y}_1)^2$$

**Repeat** the procedure by selecting $(x_2,y_2)$, for the validation data

$$MSE_2 = (y_2 - \hat{y}_2)^2$$

……..

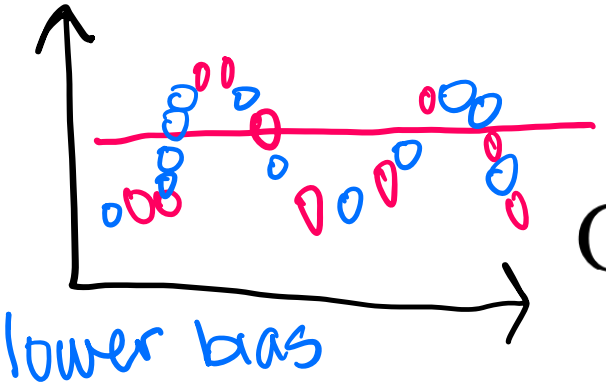
Repeating this approach **n** times produces n squared errors, $MSE_1$ , $MSE_2$ , ……$MSEn$

# Leave-one-out cross-validation (LOOCV)

The LOOCV estimate for the test MSE:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

# Leave-one-out cross-validation (LOOCV)




$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

**Lower Bias:** LOOCV trains the model on **n - 1 observations** (almost the entire dataset), it provides a more accurate estimate of the test error compared to the validation set approach, which typically uses **half size** of the data set.

**Consistency:** LOOCV is **deterministic**—it always gives the same result when repeated, unlike the validation set approach, which can vary depending on how the data is randomly split.

# Leave-one-out cross-validation (LOOCV)

This process can be **computationally expensive**

- when: **n is large**.
- **Fitting the model is slow** (e.g., complex models)

# PRESS (Predicted Residual Sum of Squares)

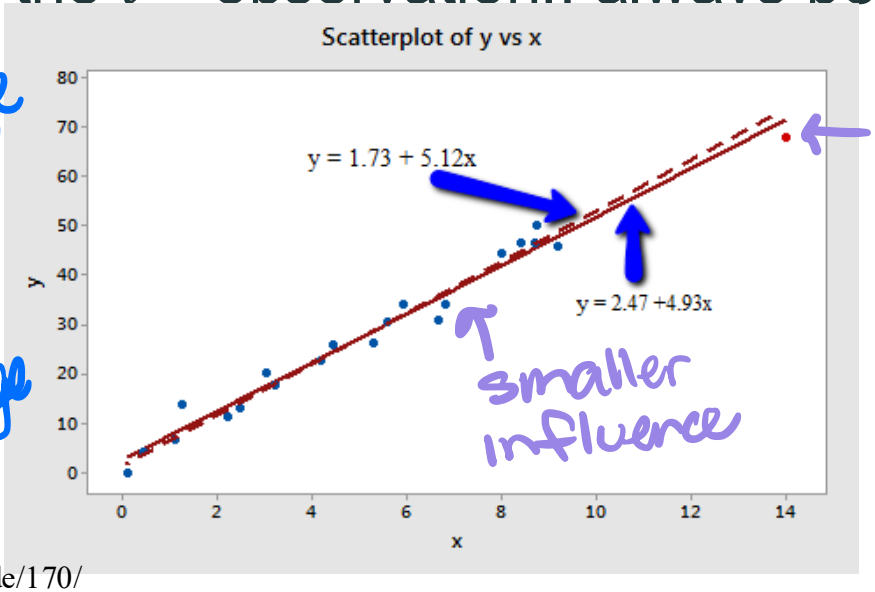
$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

$\hat{y}_i$ is the $i^{th}$fitted value

$y_1$ is the true value

$h_i$: the leverage fmeasures how much an observation influences its own fitted valueor the $i^{th}$ observation.. alwavs between **1/n** and **1**.

ex: a high-leverage point is like a seesaw (data close to the center won't change the balance, but far away from center will)



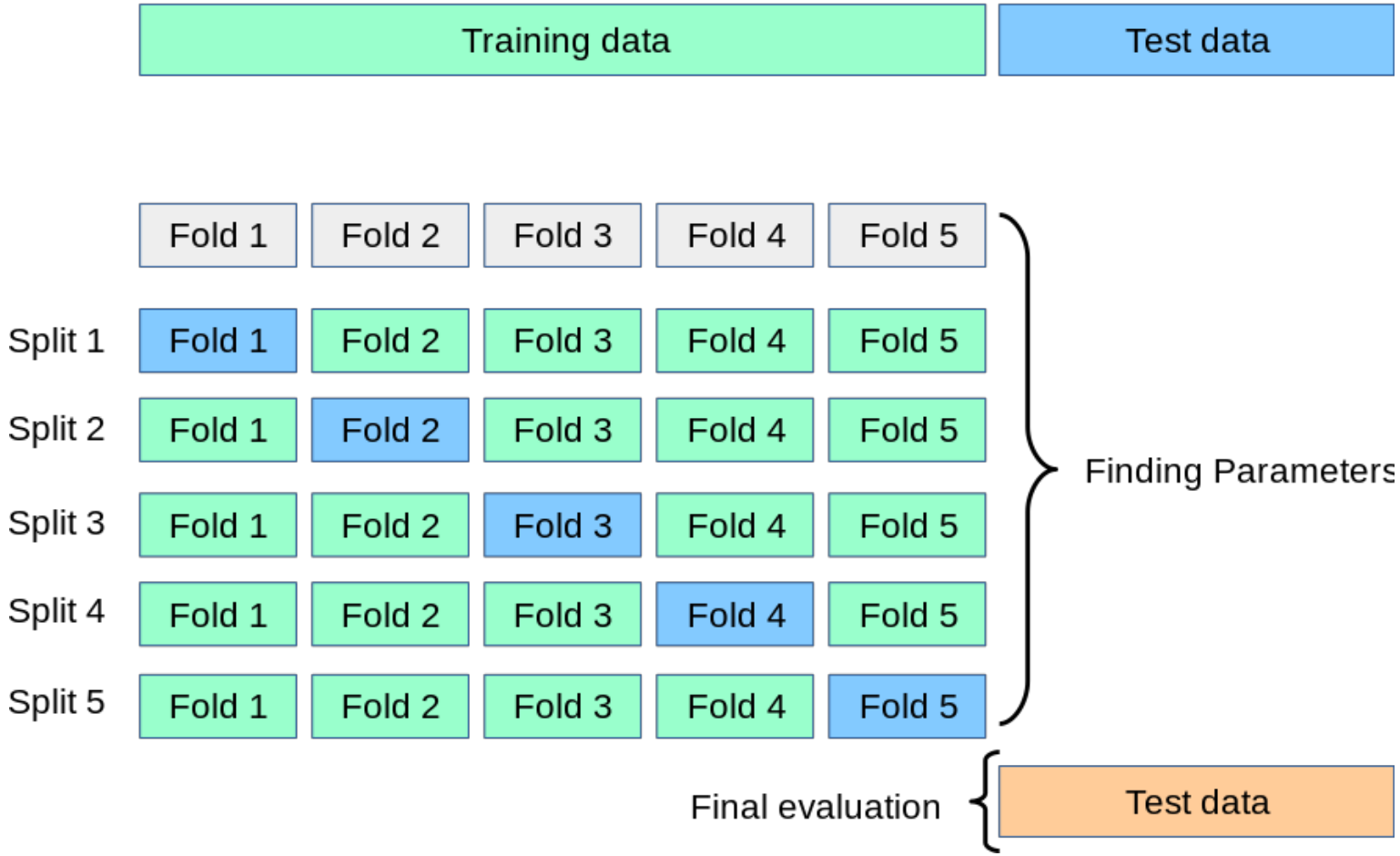

[1] https://online.stat.psu.edu/stat462/node/170/

# PRESS

**Faster Computation: PRESS** uses a shortcut formula involving leverage values, making it as fast as fitting the model **once**, while **LOOCV** requires fitting the model **n times**.

**Same Result:** For linear regression, **PRESS** gives the **same result as LOOCV**, but is computationally efficient.

# k-fold cross-validation




- Randomly dividing the set of observations into **k groups**, or *folds*, of approximately equal size.
- The **k-th fold** is treated as a validation set, and the method is fit on the remaining **k − 1 folds**.

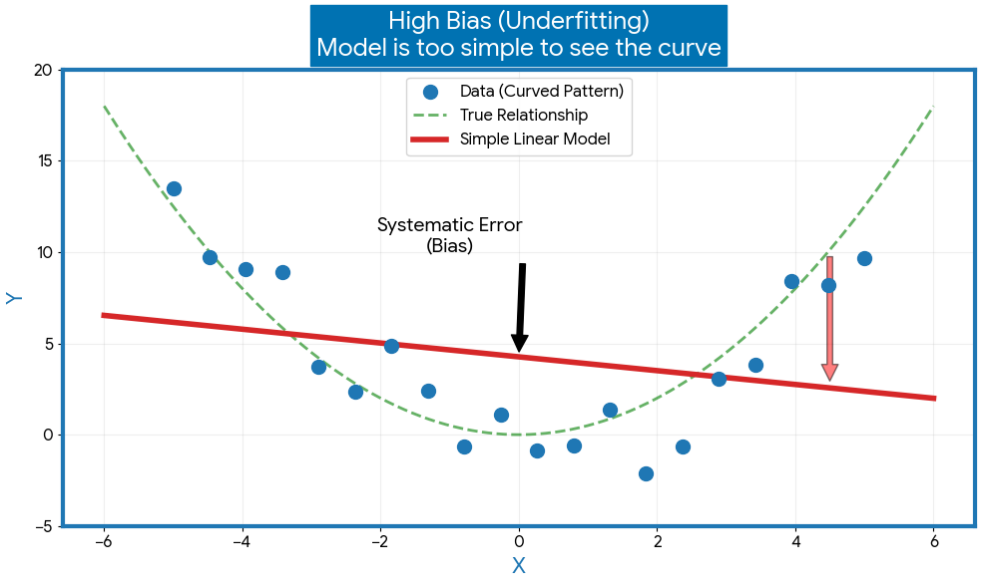
# k-fold cross-validation

$$\mathrm{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{MSE}_i .$$

# Bias and Variance of the Model (True Prediction Performance)

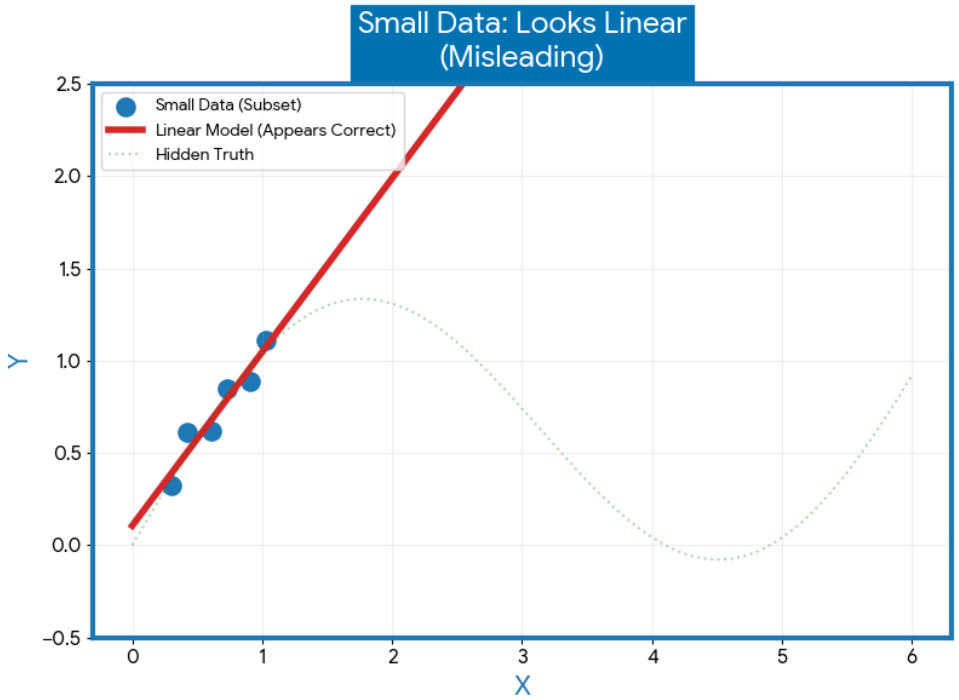
**Bias** : Measures how far off the model's predictions are on average from the true relationship.

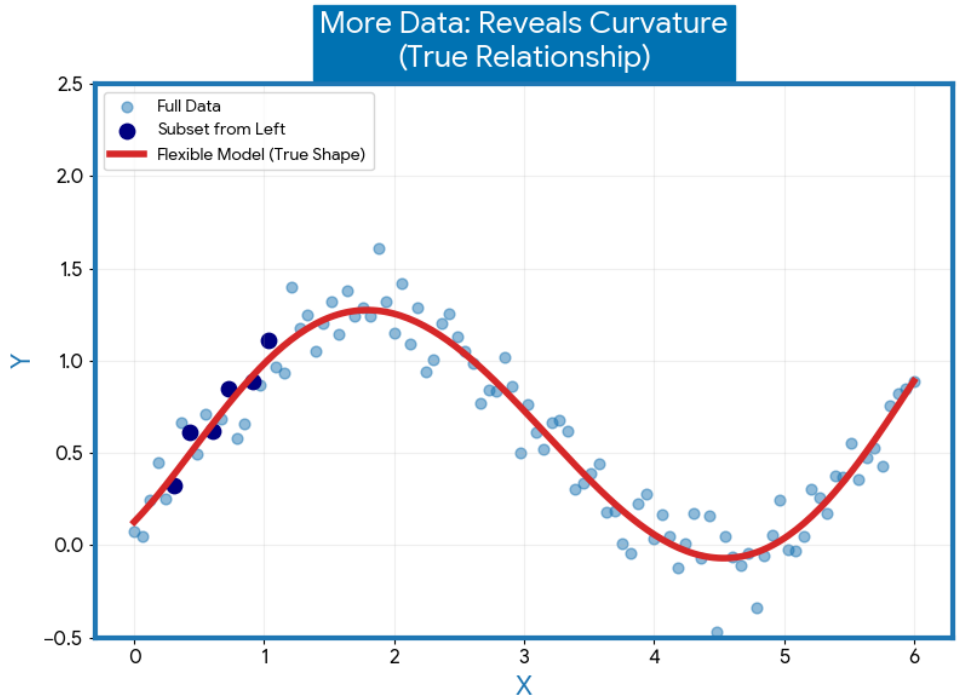
**High bias** → The model is too simple, underfits the data, and makes systematic errors.




**Variance** : Measures how much the model's predictions would vary if we trained it on different datasets.

**High variance** → The model is too sensitive to the training data, and small changes in the data lead to large changes in predictions.




High bias

Low bias

# Bias-variance trade-off k-fold CV

**Validation Set Approach (50%):**

**High bias** because the training set is smaller, leading to underfitting.

**LOOCV (n-1):**

**Low bias** because each training set uses almost the entire dataset (n - 1 observations).

**k-Fold Cross-Validation (k-1/k):**

**Intermediate bias** (lower than the validation set approach but higher than LOOCV).

| Method | Training data | Bias |
|---|---|---|
| Validation Set Approach | 50% | High |
| LOOCV | n-1 | Low |
| k-Fold CV | (k-1)/k | Medium |

# Bias-variance trade-off k-fold CV

Therefore, from the perspective of bias reduction it is clear that Loocv is to be preferred to k-fold.