

AI Bootcamp

Advanced Regression Techniques

Module 12 Day 2



Class Objectives

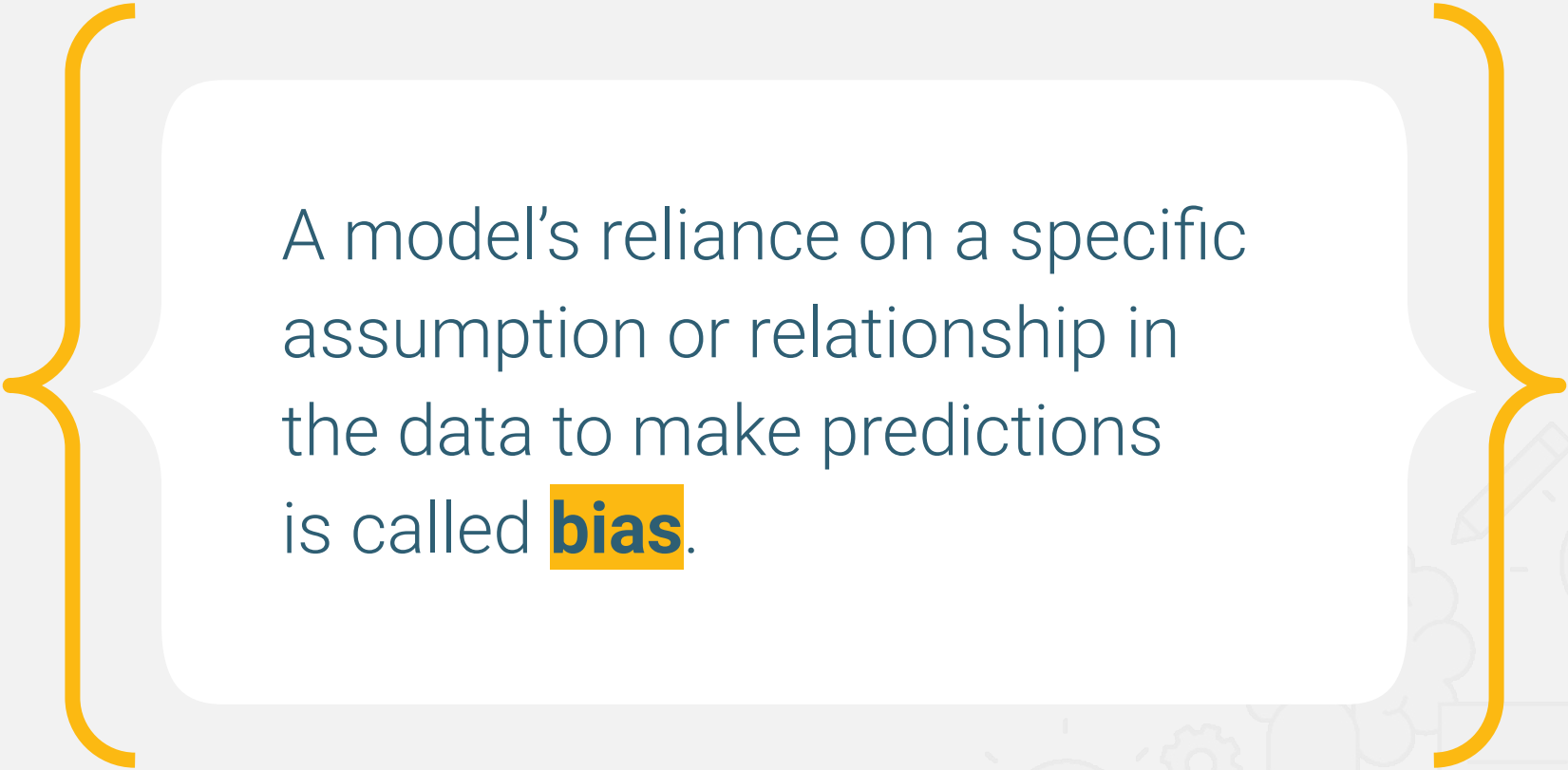
By the end of class, you will be able to:

- 1 Define bias and variance.
- 2 Describe the need for the adjusted R-squared value.
- 3 Select features for a model using p-values from OLS in the statsmodels library.
- 4 Define multicollinearity.
- 5 Test for multicollinearity using VIF.
- 6 Define regularization.
- 7 Apply regularization using ridge and lasso regression.




Instructor **Demonstration**

Bias, Variance, and Evaluation

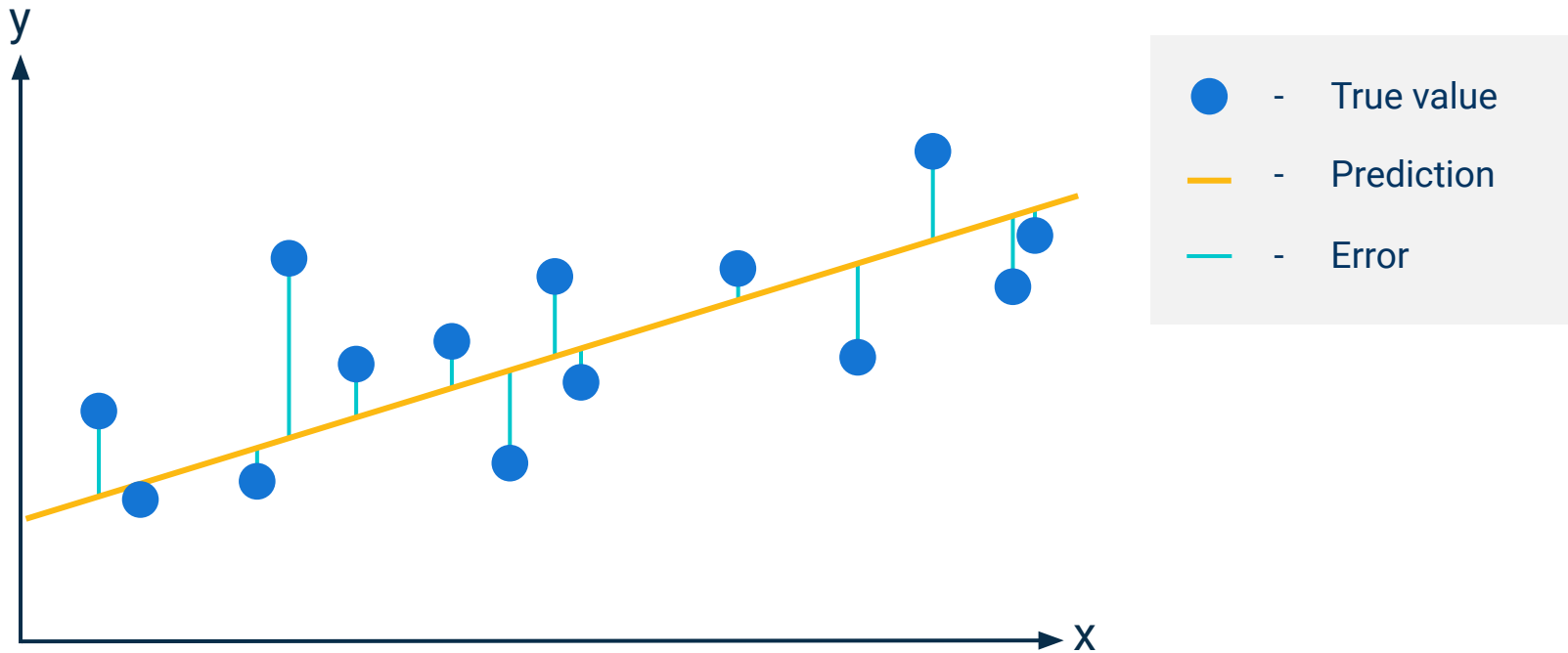


A model's reliance on a specific assumption or relationship in the data to make predictions is called **bias**.



Bias

This model is biased toward a linear interpretation of the data.



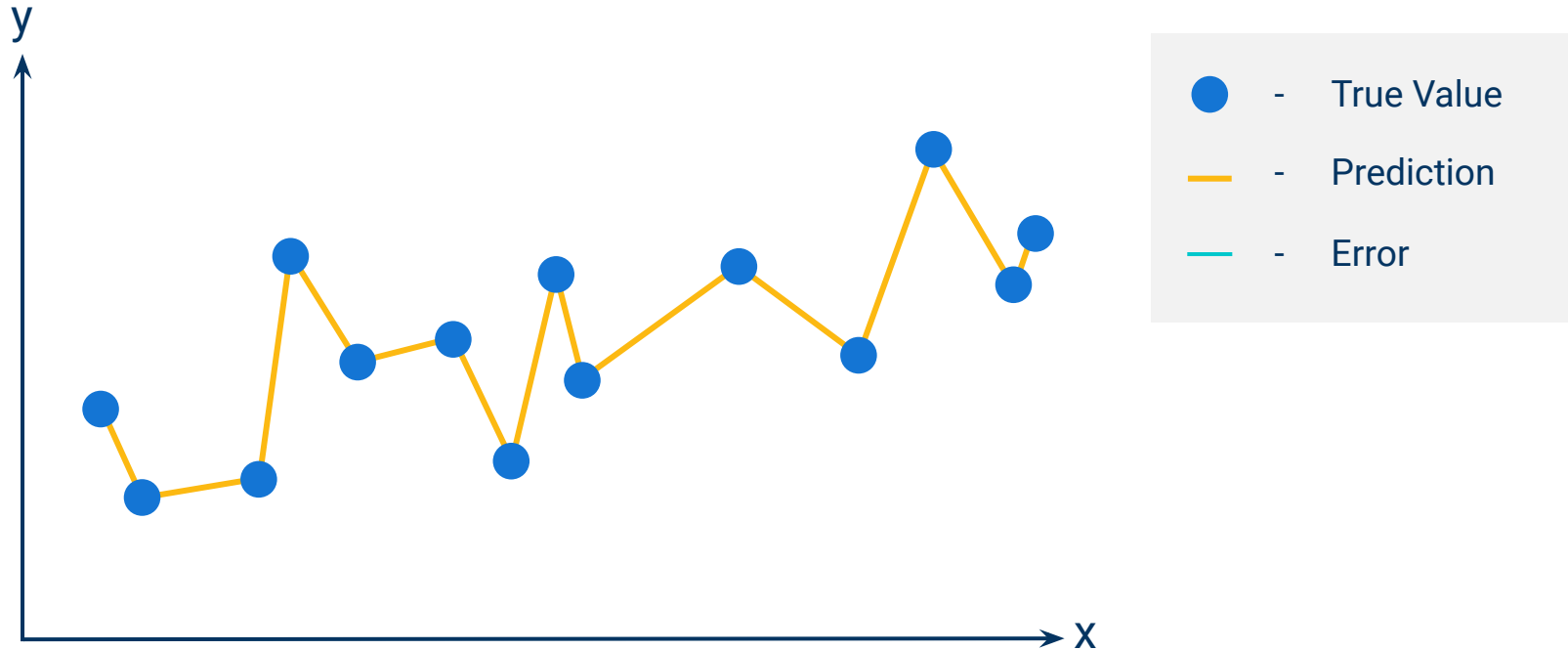


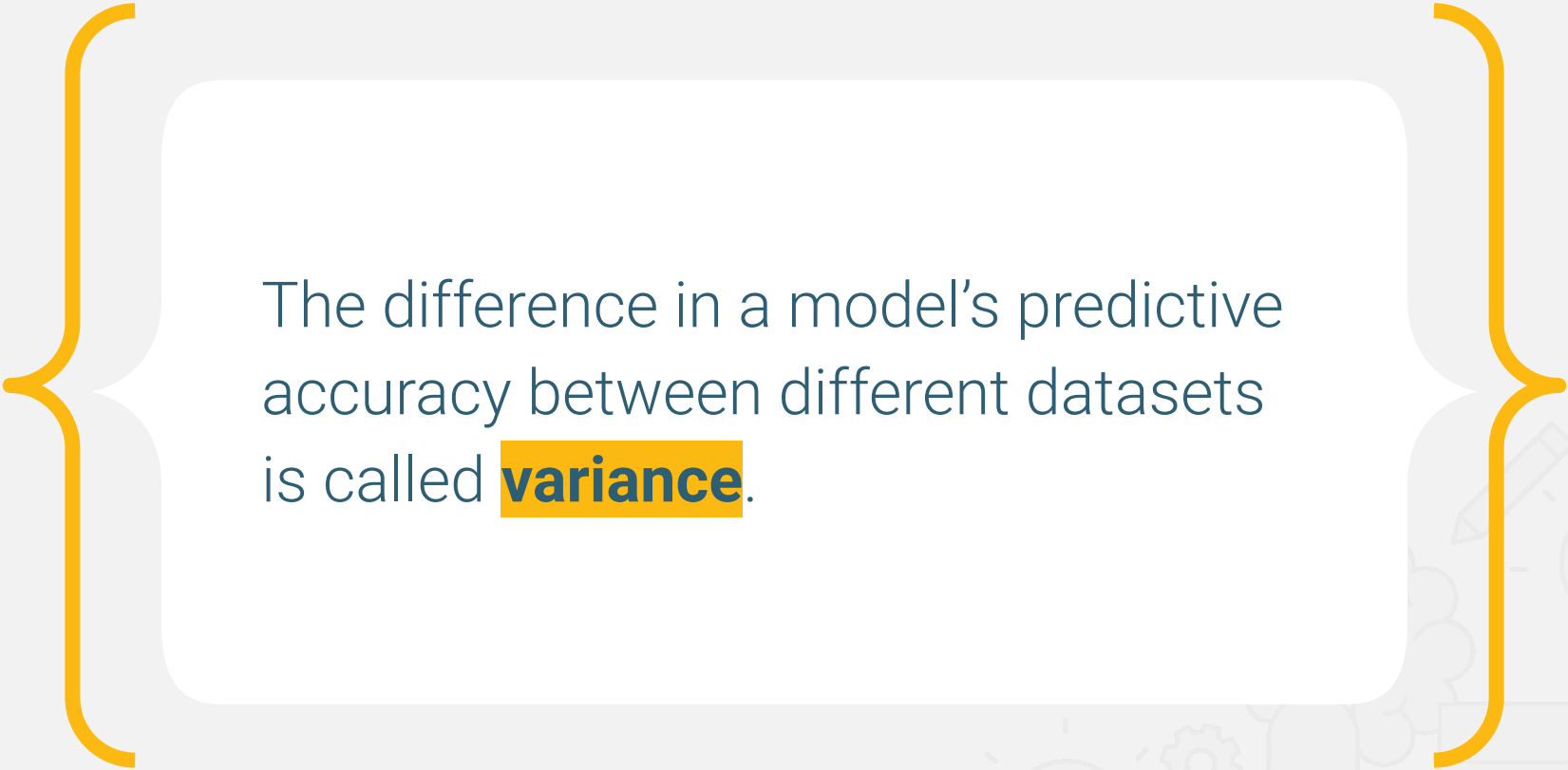
What would the model
look like with **zero bias**?




Zero Bias Model

Every value is perfectly predicted by the model.



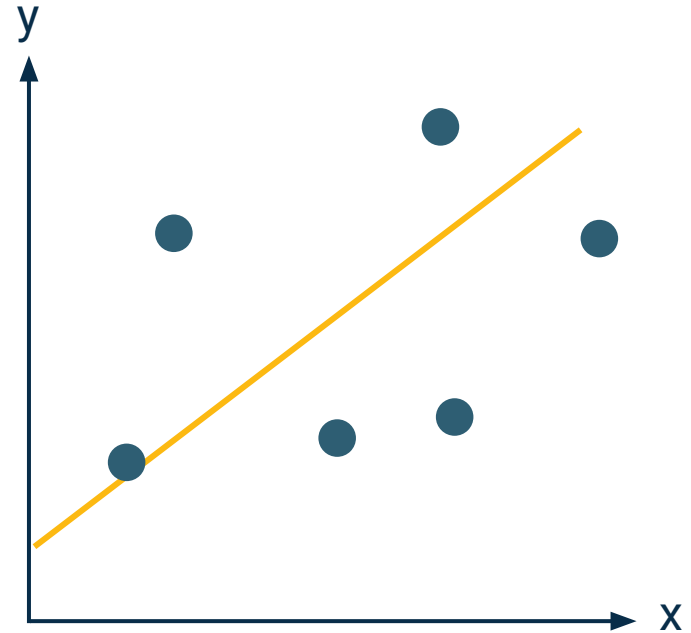
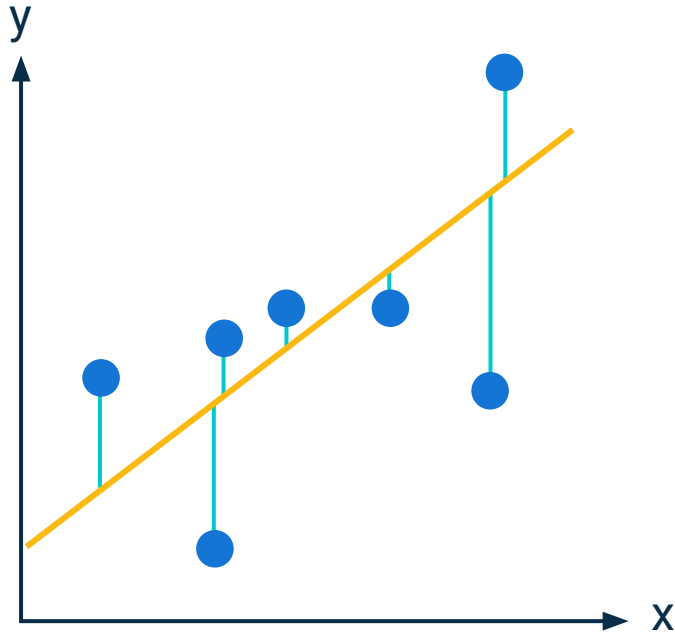


The difference in a model's predictive accuracy between different datasets is called **variance**.



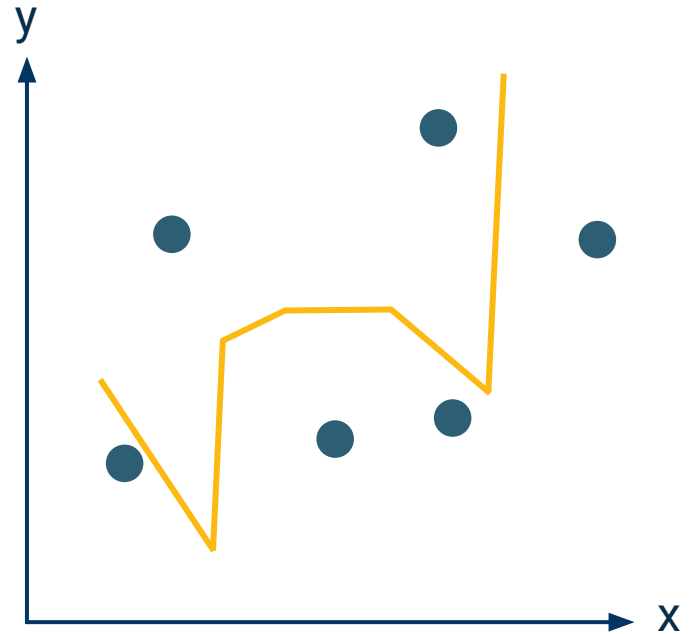
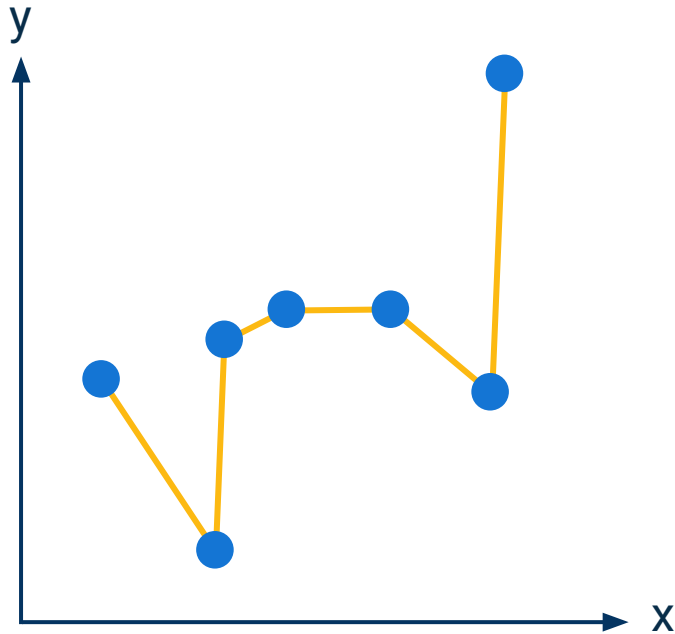
Variance

Change in the predictive accuracy of a linear model with two different datasets.



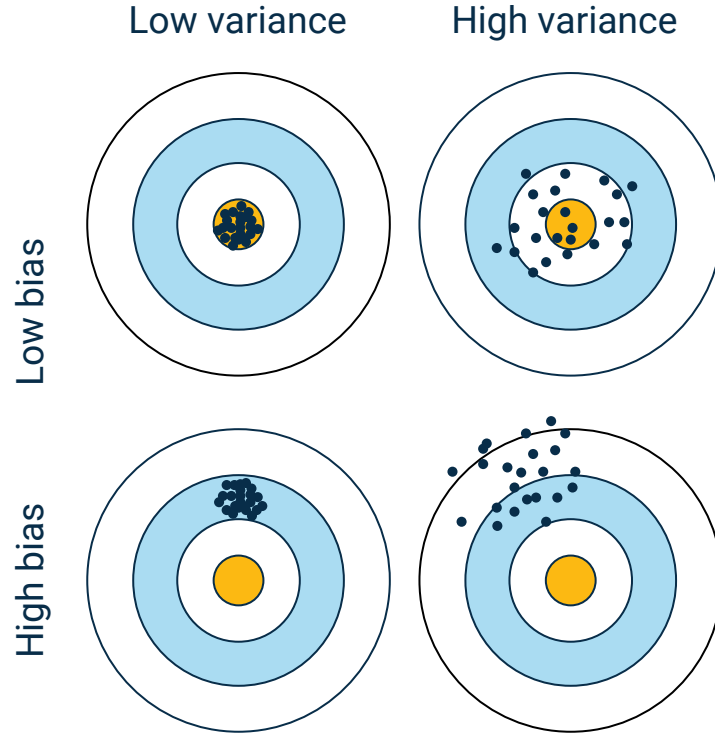
Variance

Change in the predictive accuracy of a non-linear model with two different datasets.



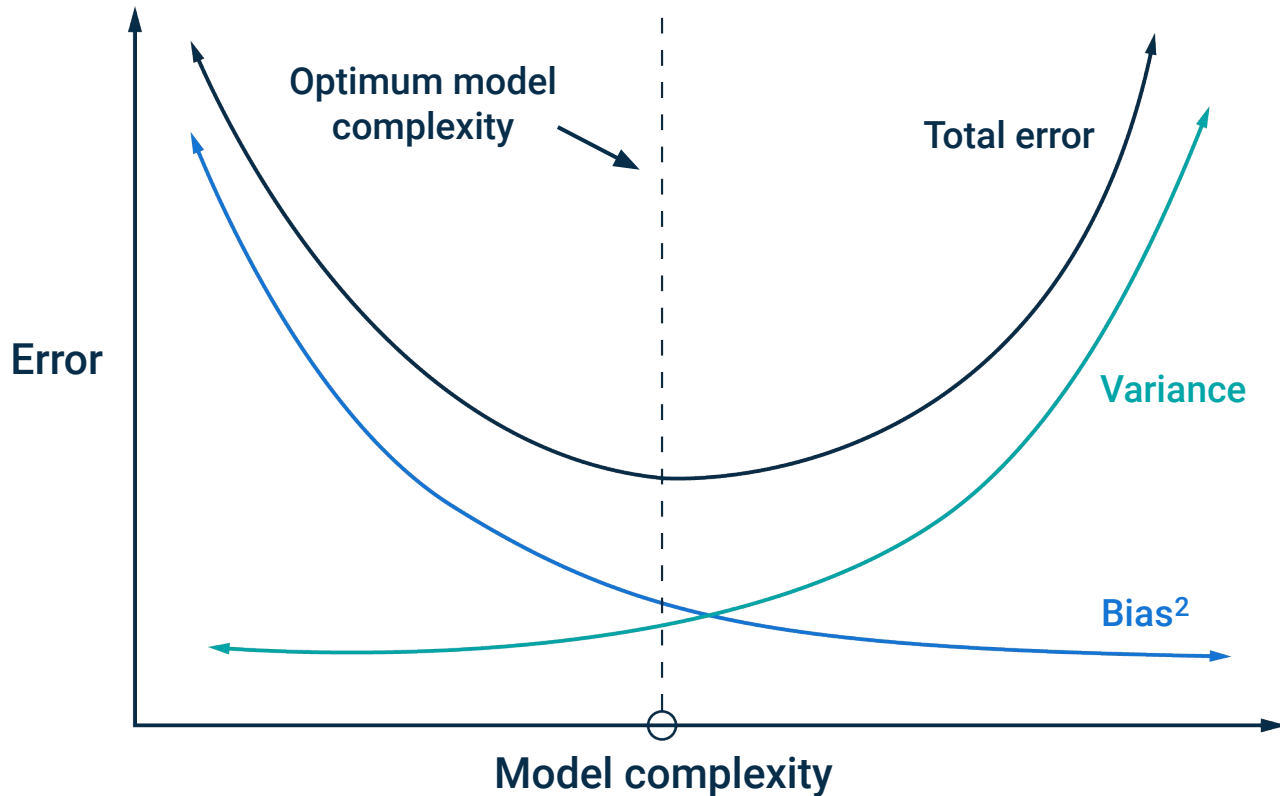
Bias and Variance

Degrees of low and high bias and variance.



Bias–Variance Trade-off

There is a trade-off between bias and variance that is tied to model complexity. This relationship is referred to as a model's "fit."

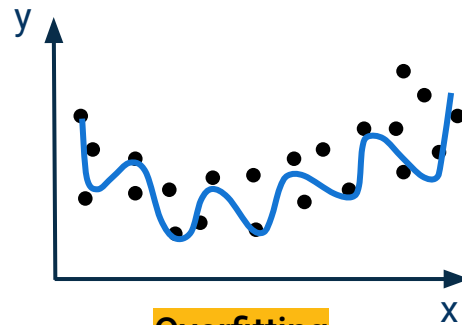
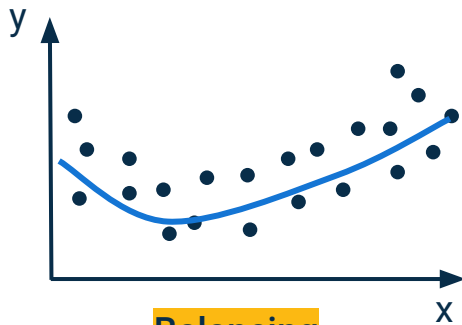
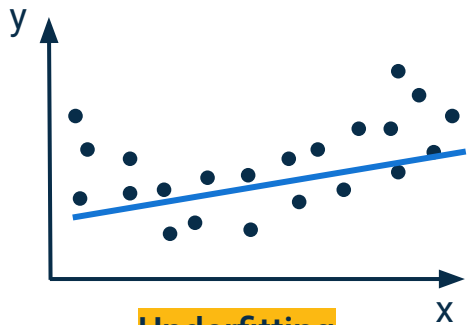


Overfitting and Underfitting

Degrees of low and high bias and variance.

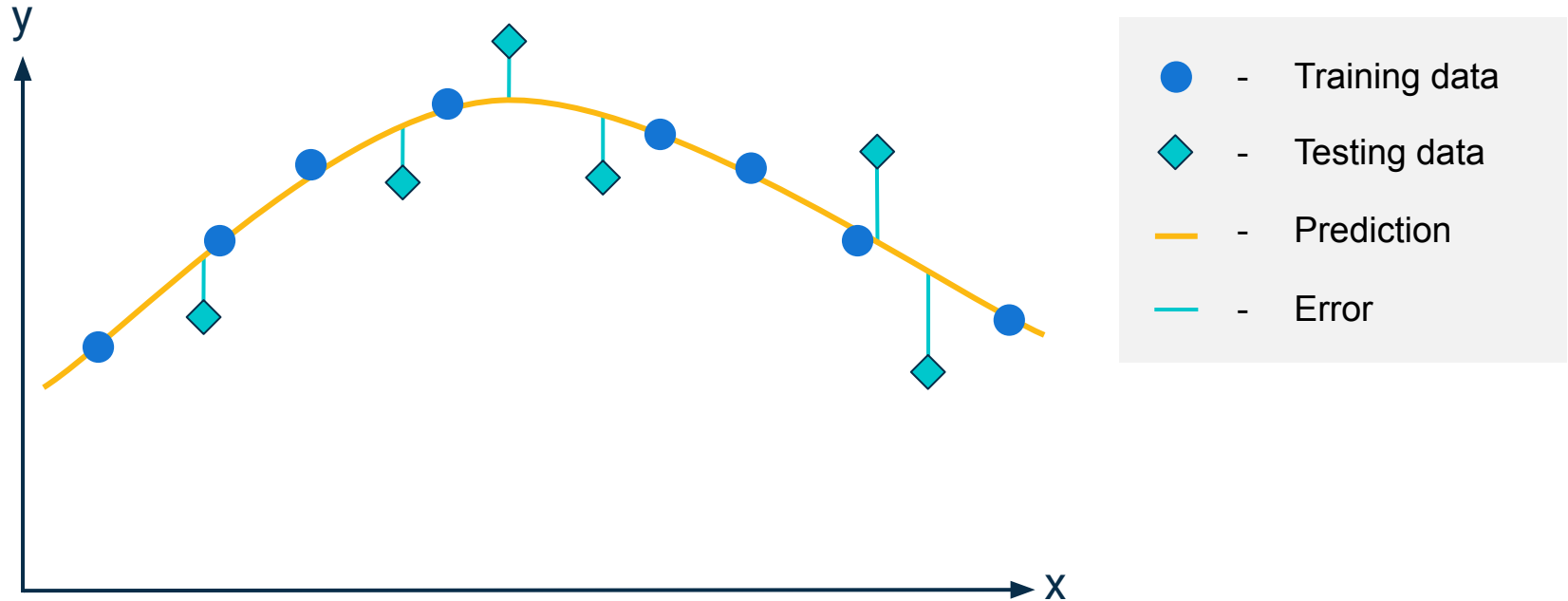
Overfitting: The model is too complex. There is high variance. This is similar to memorizing the answers to a practice exam.

Underfitting: The model is too simple and does not adequately capture the patterns in the data.



Validation

There are many ways to validate a model's bias and variance. **train_test_split** allows us to validate a model's ability to find relationships in the training data that describe the dataset as a whole.



Examples: Methods to Reduce Bias

1 Add more features (columns) to train the model.

2 Try limiting the complexity of the model.

3 Choose a different algorithm.

4 Use an ensemble of weaker models.

5 Scale the data.

Examples: Methods to Reduce Variance

- 1 Add more training data.
- 2 Validate the model before deploying.
- 3 Use multiple types of validation.
- 4 Remove features that are under or overvalued by the model.
- 5 Retrain the model regularly while in use in production.

Questions

1

A model performs well on the training data but poorly on the testing data. Which is likely more responsible for this problem: high bias or high variance?

2

Is a model that performs well on training data but poorly on testing data underfitted or overfitted?

3

To address overfitting, is it better to increase or decrease the complexity of a model?



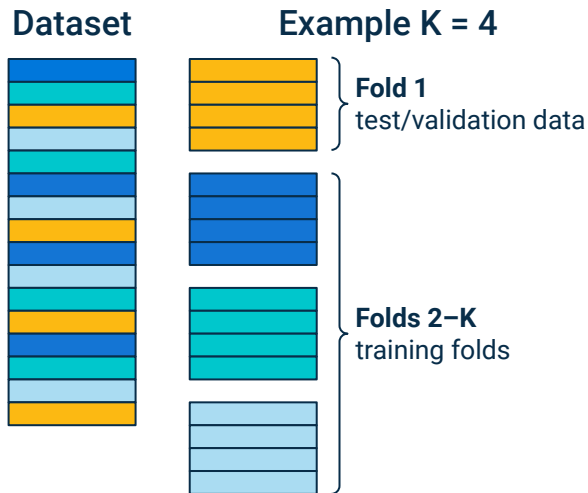
K-fold Cross-validation

Cross-validation ensures the model is trained in the most reliable way possible. The process is as follows:

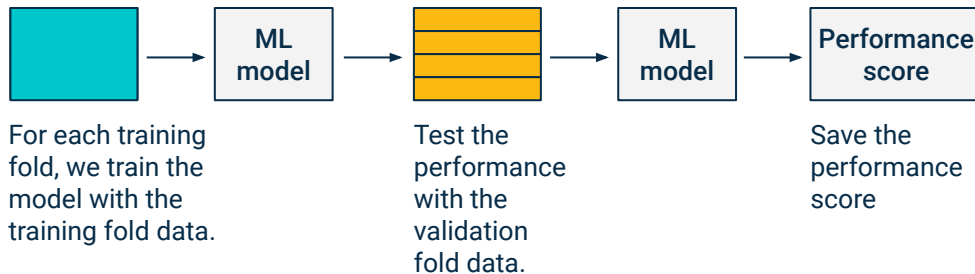
- 1 Split the dataset into a training and testing dataset as before.
- 2 Split the training dataset into **k** number of partitions or **folds**.
- 3 The model is trained **k** times. In each iteration, set aside one fold as a **validation fold** (or what we've been using as a testing set) and use the rest of the folds to train the model. Then, test the model on the validation fold.
- 4 With every subsequent iteration, select a different fold to be used as the **validation** fold.
- 5 After you've evaluated the performance of the model in each iteration, calculate the average values of the model's performance metrics.
- 5 Finally, test the model on the original testing dataset from Step 1.

K-fold Cross-validation

1 Split the data

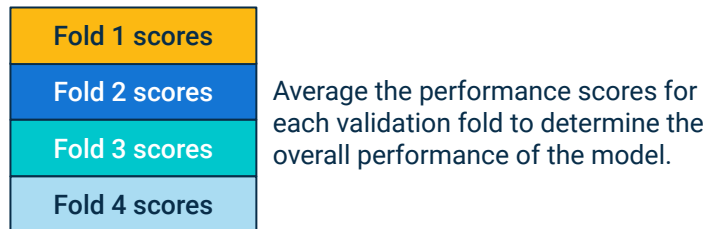


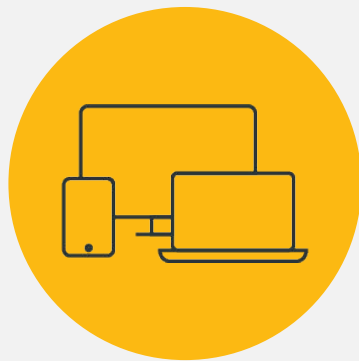
2 Train the data and score the model for k-folds



Step 2 is repeated until each fold has been used as validation data.

3 Interpret the overall performance





Instructor **Demonstration**

Measuring Bias and Variance



Activity:

Adjusted R^2 and Cross-validation

In this activity, you will use R^2 and adjusted R^2 to evaluate two linear regressions.

Suggested Time:

15 Minutes



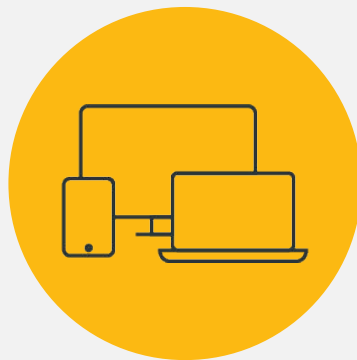


Time's up!
Let's review



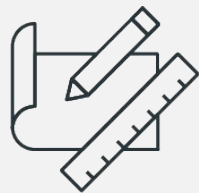
Questions?





Instructor **Demonstration**

P-values for Regression



Introduction to **Hypothesis Testing**





Hypothesis testing can be confusing at times, mostly because you must create your null and alternative hypotheses before performing any analysis.



Hypothesis Testing

A hypothesis statement is an educated guess.

The hypothesis is often expressed as an **if-then** statement.

01

Hypothesis testing is a way to test if the results of a survey or experiment are meaningful.

02

We test for two mutually exclusive outcomes:

- The null hypothesis
- The alternative hypothesis

Hypothesis Testing

Null and alternative hypotheses.

Null hypothesis (H_0)

- The hypothesis to disprove; it states that no statistical significance exists between the two variables.
- The null hypothesis assumes that the results happened by chance.

Alternative hypothesis (H_a)

- The opposite of the null hypothesis; it assumes that some factor influenced the results—meaning that they did not happen by chance.
- Moving forward, we will refer to the alternative hypothesis as just the hypothesis.

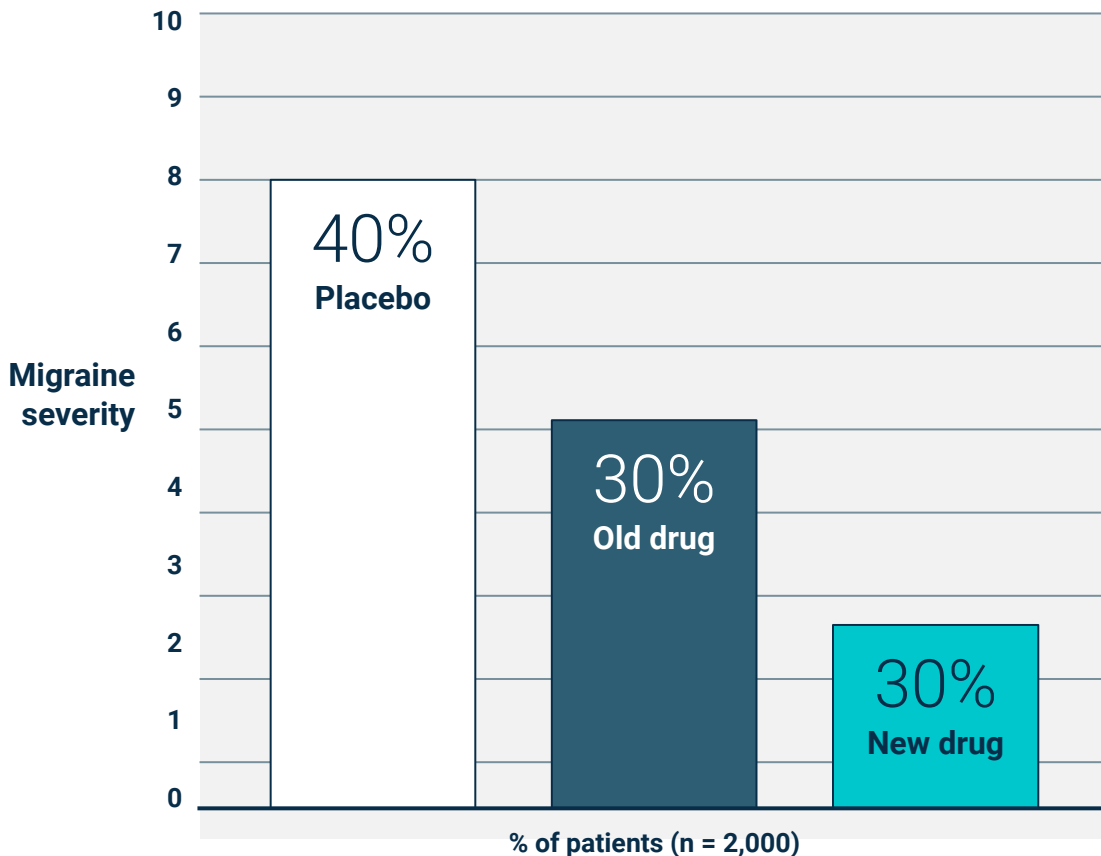
Null and Alternative Hypotheses: Example

Null hypothesis:

Migraine severity when taking the new drug is the same when taking the old drug.

Alternative hypothesis:

Migraine severity when taking the new drug is statistically significantly better or worse than when taking the old drug.





Hypothesis Testing

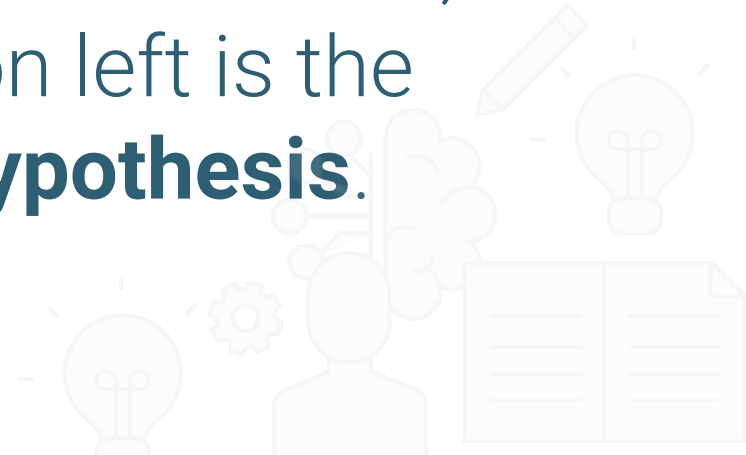
Steps for hypothesis testing:

- 1 Determine the hypothesis and null hypothesis.
- 2 Identify the appropriate statistical test.
- 3 Determine the acceptable significance value.
- 4 Compute the p-value.
- 4 Determine if the p-value rejects the null hypothesis by comparing it to the significance value (typically, $p < 0.05$).





In an analysis, we test the hypothesis and decide whether we can reject the null hypothesis. If we can, the only option left is the **alternative hypothesis.**





Activity:

P-values with Regression

- In this activity, you will use apartment rent data to predict the price of a rental.
- You will use p-values to choose only the most relevant columns in a model's training.
- And you will compare the adjusted R-squared value to that of a model trained with all columns.

Suggested Time:

10 Minutes





Time's up!
Let's review



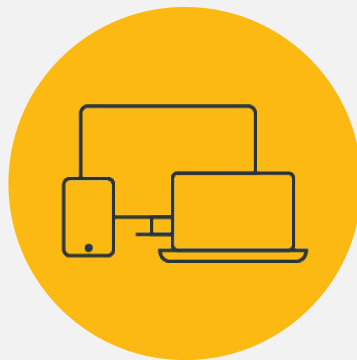
Questions?





Break

15 mins



Instructor **Demonstration**

Introducing Multicollinearity and VIF

Reasons multicollinearity can occur:

01

Inaccurate use of dummy variables

02

Poorly designed dataset

03

Multiple features that contain the same data in a different format

04

Variables that are created from other features in the dataset

05

Insufficient data

Variance Inflation Factor (VIF)

The VIF is one of the ways we check for multicollinearity between features.



The lowest value of VIF is 1: no correlation between that feature and any of the other variables.



The higher the VIF for a feature, the more likely that it is contributing to multicollinearity.



Calculate the VIF for each feature and keep an eye out for any features with a VIF more than 1.5, which indicates multicollinearity.

What to do once the VIF has been calculated for each variable:



Since higher VIF indicates stronger multicollinearity, the feature with the highest VIF can be dropped.



VIF should be recalculated every time a feature is dropped from a dataset. This is because the dropped feature may have impacted the VIF of the other features, so the score may change depending on how much that feature correlated with the others.



Features may be dropped and VIF recalculated until the VIF scores appear to be sufficiently reduced. This may be when the VIF scores are under 1.5, 5, or 10, depending on the context. Or we may stop dropping features when we see a substantial drop in adjusted R^2 .



Activity:

Detecting Multicollinearity Using VIF

In this activity, you will test for multicollinearity by calculating VIF in a car price prediction model.

Suggested Time:

15 Minutes



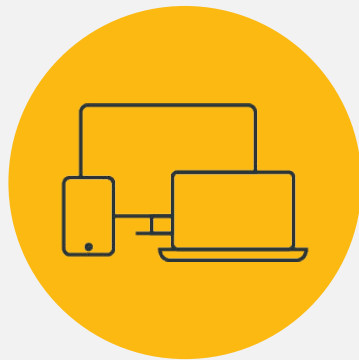


Time's up!
Let's review



Questions?





Instructor **Demonstration**

Introduction to Regularization

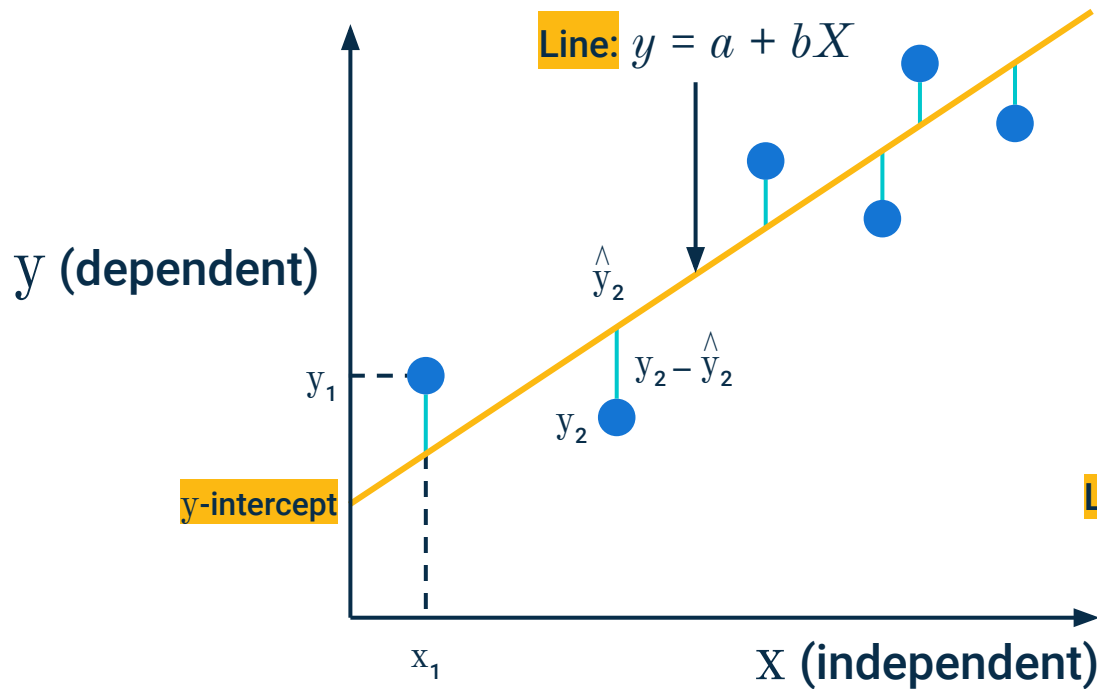


Regularization is a technique used to address overfitting or variance in regression models.



Linear Regression Model

The goal is to minimize the sum of squared residuals.



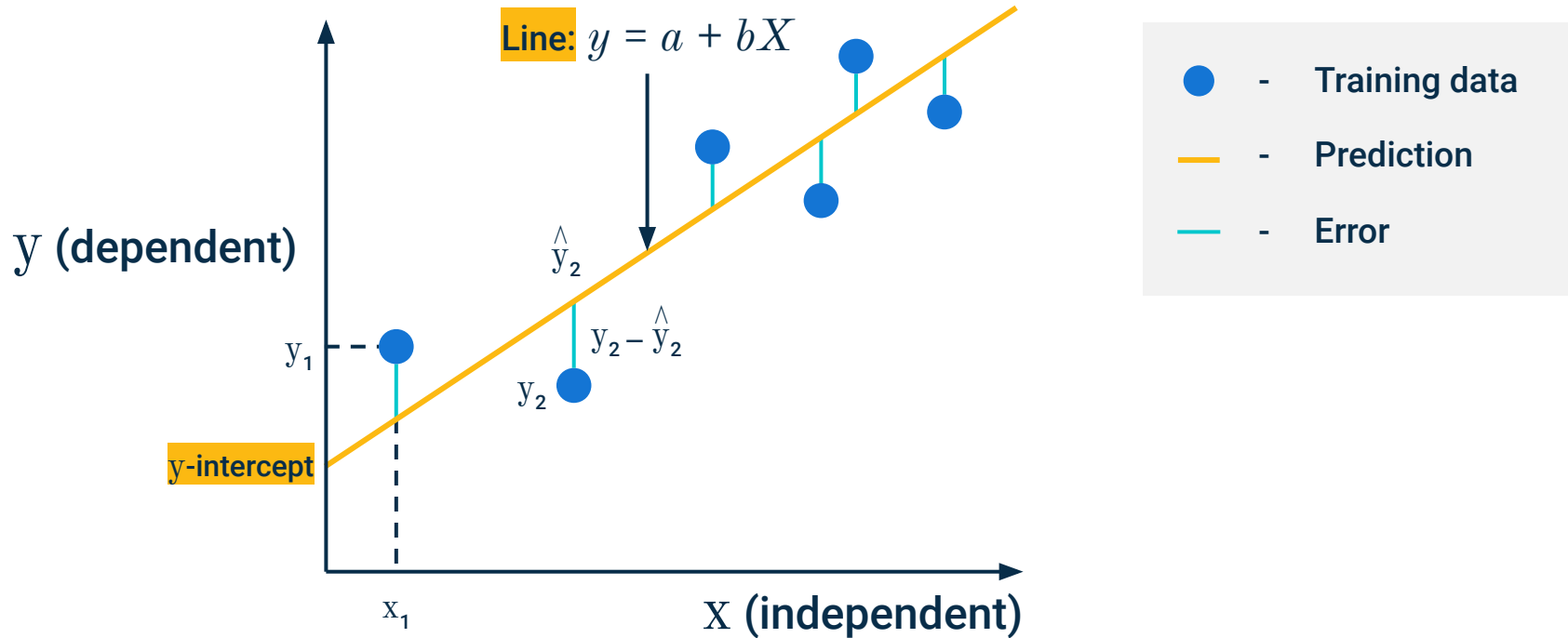
- - Training data
- - Prediction
- - Error

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

Least squares method $i = 1$

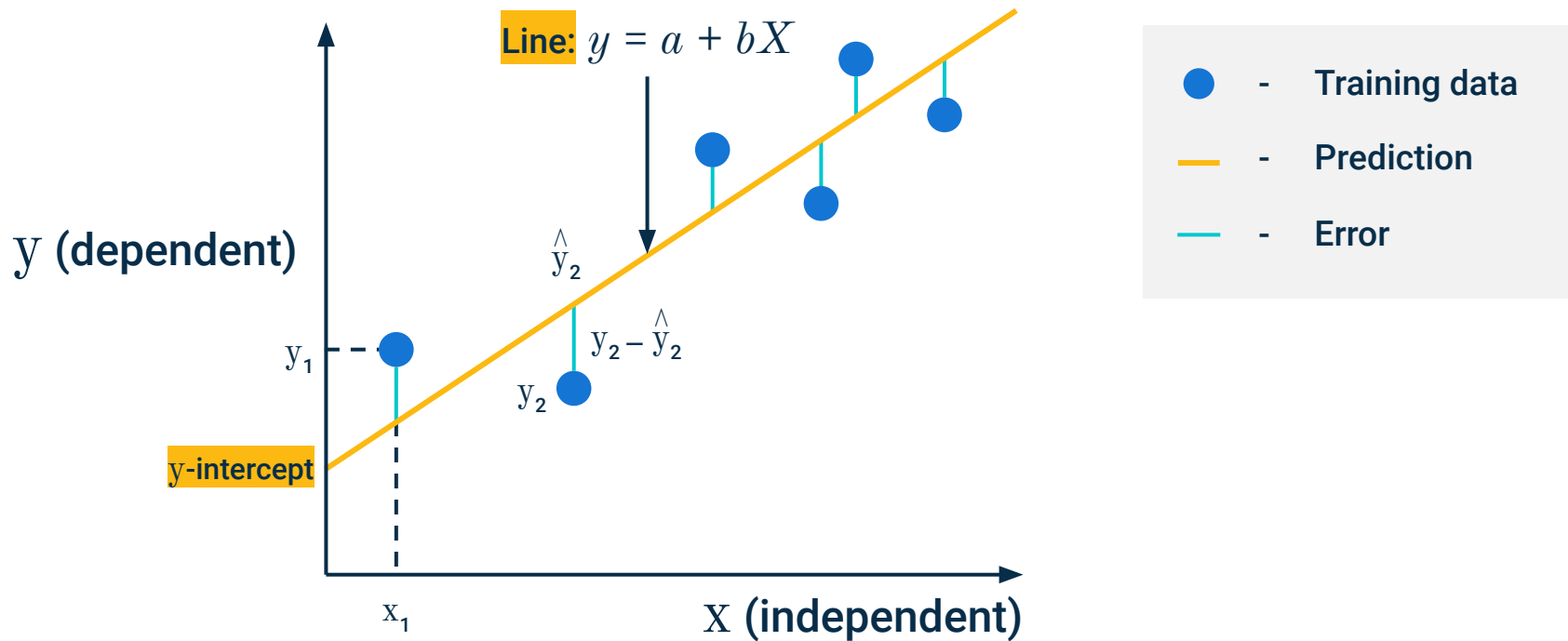
Ridge Regression Model

The goal is to minimize the sum of squared residuals + penalty for large coefficients (ridge penalty = $\alpha \times \text{sum of squared coefficients}$).



Lasso Regression Model

The goal is to minimize the sum of squared residuals + penalty for large coefficients (lasso penalty = $\frac{1}{2} \alpha * \text{sum of absolute value of coefficients}$).





Activity:

Ridge and Lasso Regression

In this activity, you will implement ridge regression on housing data and compare its results to linear regression. Then, you'll also implement lasso regression.

Suggested Time:

15 Minutes





Time's up!
Let's review



Questions?





Review the Class Objectives

In this lesson, you learned how to:

- 1 Define bias and variance.
- 2 Describe the need for the adjusted R-squared value.
- 3 Select features for a model using p-values from OLS in the statsmodels library.
- 4 Define multicollinearity.
- 5 Test for multicollinearity using VIF.
- 6 Define regularization.
- 7 Apply regularization using ridge and lasso regression.



Next

In the next lesson, you will learn about machine learning workflows and work in groups on a mini project.



Questions?





The End