Foundations of Statistical Learning II

Prof Wells

STA 295: Stat Learning

February 1st, 2024

Outline

In today's class, we will...

Outline

In today's class, we will...

- Discuss the Mean Squared Error as measure of model accuracy
- Investigate the Bias-Variance trade-off

Section 1

Mean Squared Error

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

- For regression, the most common measure of error is the **Mean Squared Error** (MSE):

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

  where $\hat{f}$ is the model, the x_i are the observed predictor values, and the y_i are the corresponding observed response values.

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

- For regression, the most common measure of error is the **Mean Squared Error** (MSE):

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

  where $\hat{f}$ is the model, the x_i are the observed predictor values, and the y_i are the corresponding observed response values.

- We also often work with **root mean squared error** (RMSE):

$$\mathrm{RMSE}(\hat{f}) = \sqrt{\mathrm{MSE}(\hat{f})}$$

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

- For regression, the most common measure of error is the **Mean Squared Error** (MSE):

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

  where $\hat{f}$ is the model, the x_i are the observed predictor values, and the y_i are the corresponding observed response values.

- We also often work with **root mean squared error** (RMSE):

$$\mathrm{RMSE}(\hat{f}) = \sqrt{\mathrm{MSE}(\hat{f})}$$

  - What is one advantage of RMSE over MSE?

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

- For regression, the most common measure of error is the **Mean Squared Error** (MSE):

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

  where $\hat{f}$ is the model, the x_i are the observed predictor values, and the y_i are the corresponding observed response values.

- We also often work with **root mean squared error** (RMSE):

$$\mathrm{RMSE}(\hat{f}) = \sqrt{\mathrm{MSE}(\hat{f})}$$

  - What is one advantage of RMSE over MSE?

- Under what circumstances is MSE small?

Training and Test Data

- What are the problems with finding a function f which minimize MSE on the set of observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$?

Training and Test Data

- What are the problems with finding a function f which minimize MSE on the set of observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$?
- **Training Data** is the collection of data we use to build our model. Often, it is a subset of all data we have available.

Training and Test Data

- What are the problems with finding a function f which minimize MSE on the set of observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$?
- **Training Data** is the collection of data we use to build our model. Often, it is a subset of all data we have available.
- **Test Data** is the collection of data on which we assess the accuracy of our model. It should be distinct from the training data.

Training and Test Data

- What are the problems with finding a function f which minimize MSE on the set of observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$?
- **Training Data** is the collection of data we use to build our model. Often, it is a subset of all data we have available.
- **Test Data** is the collection of data on which we assess the accuracy of our model. It should be distinct from the training data.
- *Goal*: Use a model-building algorithm that builds model on **training data** in order to minimize MSE on a large number of unobserved **test data** points (x_0, y_0)

Training and Test Data

- What are the problems with finding a function f which minimize MSE on the set of observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$?
- **Training Data** is the collection of data we use to build our model. Often, it is a subset of all data we have available.
- **Test Data** is the collection of data on which we assess the accuracy of our model. It should be distinct from the training data.
- *Goal*: Use a model-building algorithm that builds model on **training data** in order to minimize MSE on a large number of unobserved **test data** points (x_0, y_0)
  - i.e. minimize

$$\text{test MSE} = \mathrm{Ave}\left(y_0 - \hat{f}(x_0)\right)^2$$

Training and Test Data

- What are the problems with finding a function f which minimize MSE on the set of observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$?
- **Training Data** is the collection of data we use to build our model. Often, it is a subset of all data we have available.
- **Test Data** is the collection of data on which we assess the accuracy of our model. It should be distinct from the training data.
- *Goal*: Use a model-building algorithm that builds model on **training data** in order to minimize MSE on a large number of unobserved **test data** points (x_0, y_0)
  - i.e. minimize

$$\text{test MSE} = \mathrm{Ave}\Big(y_0 - \hat{f}(x_0)\Big)^2$$

- Additionally, we can construct a number of models on the training data, and compare their performance on the test data in order to select the best model

An Example

- Suppose we have 50 observations on a quantitative response Y and quantitative predictor X

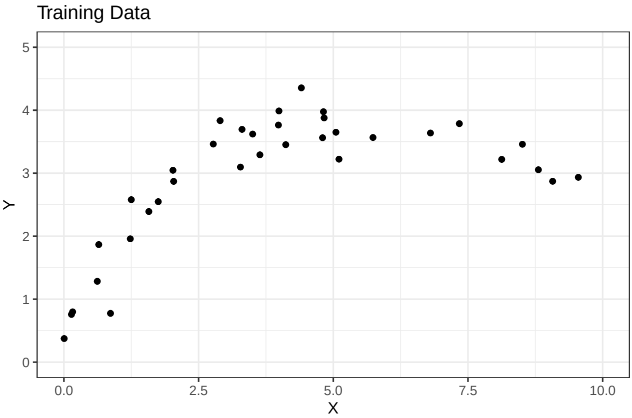
An Example

- Suppose we have 50 observations on a quantitative response Y and quantitative predictor X
    - We plan to use 70% of our data (35 observations) as a training set.
    - We use the remaining 30% of the data (15 observations) as a test set.

An Example

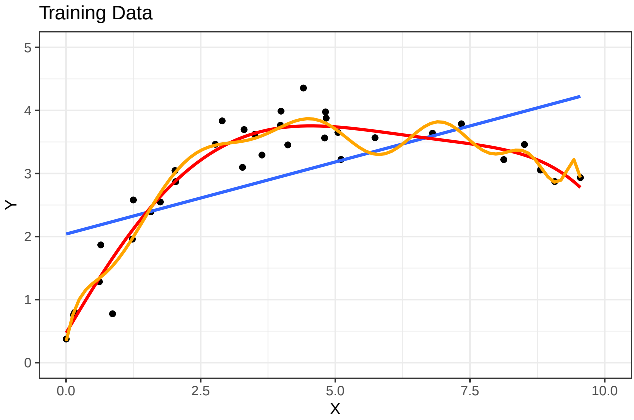
- Suppose we have 50 observations on a quantitative response Y and quantitative predictor X
  - We plan to use 70% of our data (35 observations) as a training set.
  - We use the remaining 30% of the data (15 observations) as a test set.
- We will fit three models:
  1. A linear model; low flexibility)
  2. A quintic model; medium flexibility
  3. A degree 15 model; high flexibility

Training Set



- Data follows a non-linear trend

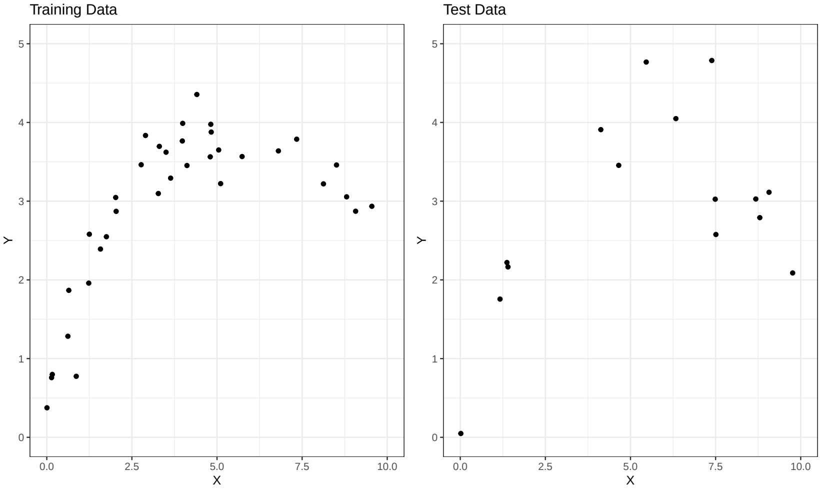
Model 1, 2, and 3



model	Train.MSE
Linear	0.677
Quintic	0.086
Poly	0.071

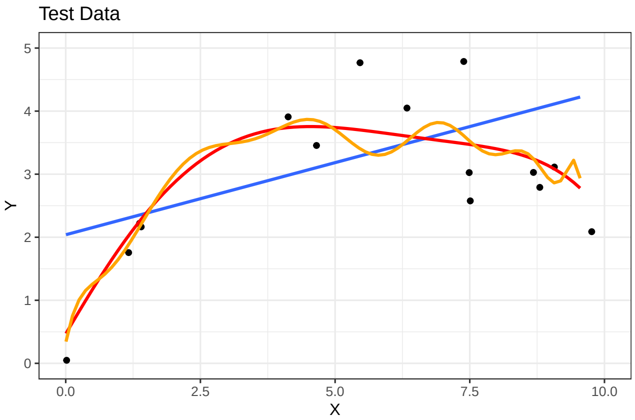
- We build a linear, quintic, and 17th degree polynomial model

Test Set



- Test data generated from same model as training data

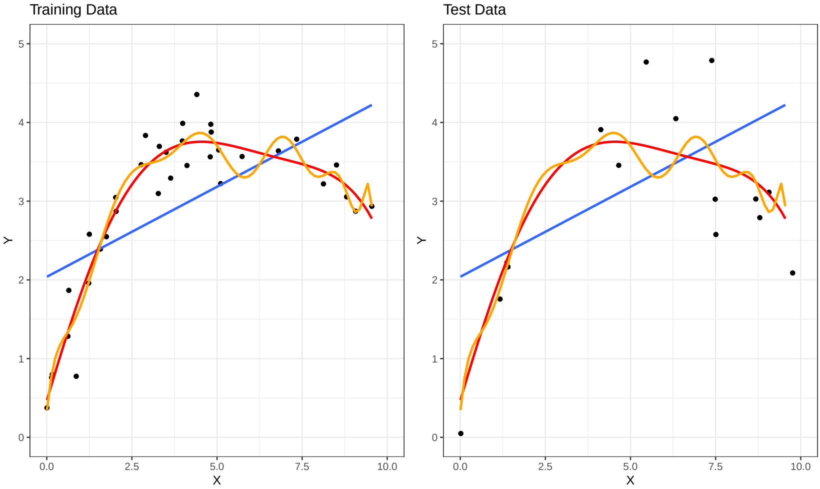
Test Set with Models



model	Test.MSE
Linear	1.281
Quintic	0.326
Poly	1.822

- Models built on training data are plotted on test data

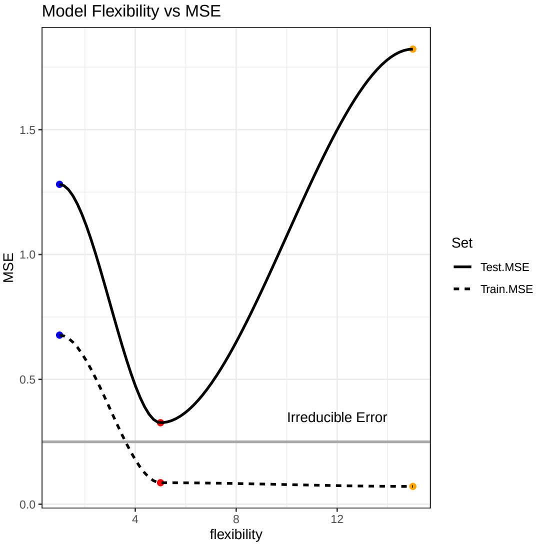
Test vs Train



- The 15th degree poly model fits the training data well. But doesn't do as well on test data.

Train vs Test MSE

model	Train.MSE	Test.MSE
Linear	0.677	1.281
Quintic	0.086	0.326
Poly	0.071	1.822



Section 2

Bias-Variance Trade-off

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing flexibility / complexity.

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing flexibility / complexity.

- What happens to the training MSE and the test MSE as model flexibility / complexity increases?

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing flexibility / complexity.

- What happens to the training MSE and the test MSE as model flexibility / complexity increases?
- As model flexibility / complexity increases, training MSE will decrease, but test MSE might not.

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing flexibility / complexity.

- What happens to the training MSE and the test MSE as model flexibility / complexity increases?
- As model flexibility / complexity increases, training MSE will decrease, but test MSE might not.
- Flexible / complex models may **overfit** data, meaning they fit patterns from the random error (noise), rather than the true model (signal)

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing flexibility / complexity.

- What happens to the training MSE and the test MSE as model flexibility / complexity increases?
- As model flexibility / complexity increases, training MSE will decrease, but test MSE might not.
- Flexible / complex models may **overfit** data, meaning they fit patterns from the random error (noise), rather than the true model (signal)
    - This leads to low train MSE, but high test MSE

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing flexibility / complexity.

- What happens to the training MSE and the test MSE as model flexibility / complexity increases?
- As model flexibility / complexity increases, training MSE will decrease, but test MSE might not.
- Flexible / complex models may **overfit** data, meaning they fit patterns from the random error (noise), rather than the true model (signal)
  - This leads to low train MSE, but high test MSE
- On the other hand, inflexible / simple models may be too rigid to fit the true pattern (lack the fidelity to convey signal)

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing flexibility / complexity.

- What happens to the training MSE and the test MSE as model flexibility / complexity increases?
- As model flexibility / complexity increases, training MSE will decrease, but test MSE might not.
- Flexible / complex models may **overfit** data, meaning they fit patterns from the random error (noise), rather than the true model (signal)
  - This leads to low train MSE, but high test MSE
- On the other hand, inflexible / simple models may be too rigid to fit the true pattern (lack the fidelity to convey signal)
  - This may lead to high train MSE and high test MSE

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

Expected test MSE can be decomposed as the sum of 3 quantities:

$$\mathrm{E}(y_0 - \hat{f}(x_0)) = \mathrm{Var}(\hat{f}(x_0)) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\epsilon)$$

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

Expected test MSE can be decomposed as the sum of 3 quantities:

$$\mathrm{E}(y_0 - \hat{f}(x_0)) = \mathrm{Var}(\hat{f}(x_0)) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\epsilon)$$

- Here $\mathrm{E}(y_0 - \hat{f}(x_0))$ denotes expected test MSE **at** x_0, if many models for f were built using a variety of random training data sets containing x_0

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

Expected test MSE can be decomposed as the sum of 3 quantities:

$$\mathrm{E}(y_0 - \hat{f}(x_0)) = \mathrm{Var}(\hat{f}(x_0)) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\epsilon)$$

- Here $\mathrm{E}(y_0 - \hat{f}(x_0))$ denotes expected test MSE **at** x_0, if many models for f were built using a variety of random training data sets containing x_0
- Total expected test MSE is obtained by averaging across all possible x_0 in the test set.

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

Expected test MSE can be decomposed as the sum of 3 quantities:

$$\mathrm{E}(y_0 - \hat{f}(x_0)) = \mathrm{Var}(\hat{f}(x_0)) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\epsilon)$$

- Here $\mathrm{E}(y_0 - \hat{f}(x_0))$ denotes expected test MSE **at** x_0, if many models for f were built using a variety of random training data sets containing x_0
- Total expected test MSE is obtained by averaging across all possible x_0 in the test set.
- A proof is given in Section 7.3 of *The Elements of Statistical Learning* (or STA 336)

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

Expected test MSE can be decomposed as the sum of 3 quantities:

$$\mathrm{E}(y_0 - \hat{f}(x_0)) = \mathrm{Var}(\hat{f}(x_0)) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\epsilon)$$

- Here $\mathrm{E}(y_0 - \hat{f}(x_0))$ denotes expected test MSE **at** x_0, if many models for f were built using a variety of random training data sets containing x_0
- Total expected test MSE is obtained by averaging across all possible x_0 in the test set.
- A proof is given in Section 7.3 of *The Elements of Statistical Learning* (or STA 336)
- To minimize MSE, we need to *simultaneously* minimize both variance and bias.

Bias and Variance

- **Variance** refers to the amount of variability in $\hat{f}(x_0)$ across random training sets containing x_0

Bias and Variance

- **Variance** refers to the amount of variability in $\hat{f}(x_0)$ across random training sets containing x_0
  - What type of models tend to have low/high variance?

Bias and Variance

- **Variance** refers to the amount of variability in $\hat{f}(x_0)$ across random training sets containing x_0
  - What type of models tend to have low/high variance?

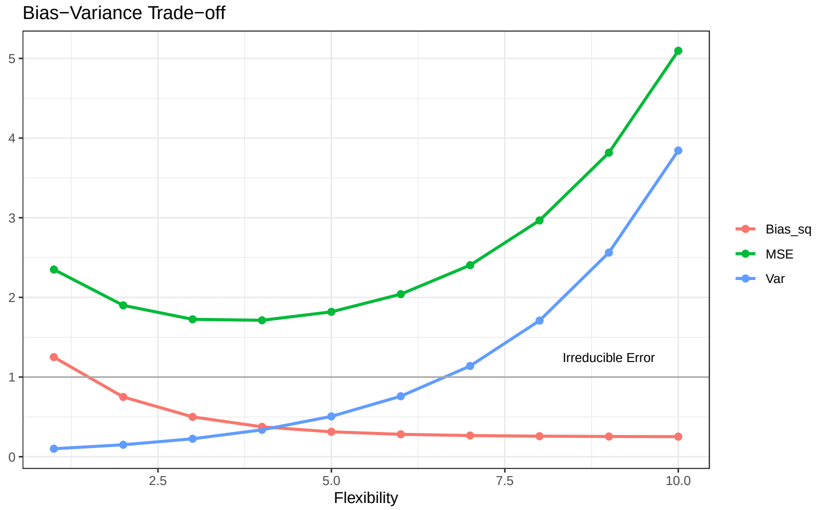
- **Bias** refers to amount by which $\hat{f}(x_0)$ differs from the true value of $f(x_0)$, on average across random training sets.
  - Bias is produced by the difference between model shape assumptions and reality

Bias and Variance

- **Variance** refers to the amount of variability in $\hat{f}(x_0)$ across random training sets containing x_0
  - What type of models tend to have low/high variance?

- **Bias** refers to amount by which $\hat{f}(x_0)$ differs from the true value of $f(x_0)$, on average across random training sets.
  - Bias is produced by the difference between model shape assumptions and reality
  - What type of models tend to have low/high bias?

Bias-Variance Trade-off



Target Practice