Assessing Model Accuracy

Nate Wells

Math 243: Stat Learning

September 8th, 2021

Outline

In today's class, we will...

Outline

In today's class, we will...

- Analyze data from the 'guess my age' activity

Outline

In today's class, we will...

- Analyze data from the 'guess my age' activity
- Discuss the Mean Squared Error as measure of model accuracy

Outline

In today's class, we will...

- Analyze data from the 'guess my age' activity
- Discuss the Mean Squared Error as measure of model accuracy
- Investigate the Bias-Variance trade-off

Section 1

How Old?

Reflection

The task: Consider photos for 8 math and stats faculty at Reed. Estimate the age of each faculty member (at the time photo was taken).








Reflection

The task: Consider photos for 8 math and stats faculty at Reed. Estimate the age of each faculty member (at the time photo was taken).

- Was the *How Old?* activity supervised or unsupervised?

Reflection

The task: Consider photos for 8 math and stats faculty at Reed. Estimate the age of each faculty member (at the time photo was taken).








- Was the *How Old?* activity supervised or unsupervised?
- Did it represent a classification or regression problem?

Reflection

The task: Consider photos for 8 math and stats faculty at Reed. Estimate the age of each faculty member (at the time photo was taken).

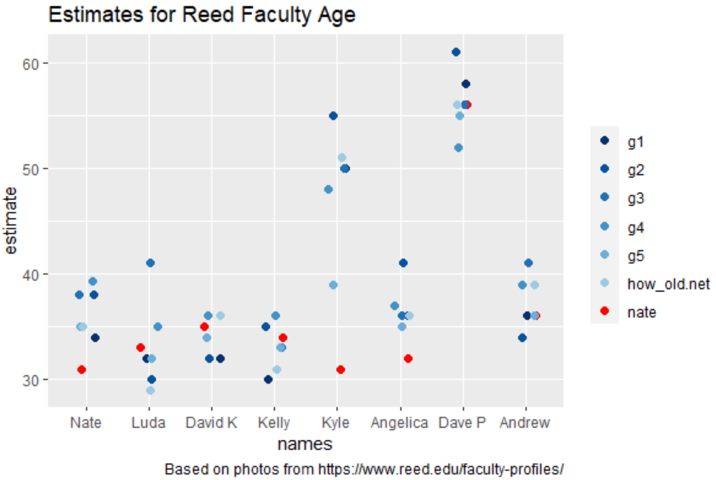
- Was the *How Old?* activity supervised or unsupervised?
- Did it represent a classification or regression problem?
- Were you interested primarily in prediction or inference?

Reflection

The task: Consider photos for 8 math and stats faculty at Reed. Estimate the age of each faculty member (at the time photo was taken).

- Was the *How Old?* activity supervised or unsupervised?
- Did it represent a classification or regression problem?
- Were you interested primarily in prediction or inference?
- Did you use parametric or non-parametric methods?

The Results



Estimates for Reed Faculty Age
estimate
60
50
40
30
Nate
Luda
David K
Kelly
Kyle
Angelica
Dave P
Andrew
names
g1
g2
g3
g4
g5
how_old.net
nate
Based on photos from https://www.reed.edu/faculty-profiles/

Debrief

- How should we quantify error?
- What are some sources for error in our estimates?
- How should we assess the overall accuracy of a group's predictions?
- Did any groups seem to consistently over- or under-estimate ages? By how much?
- Do any faculty member ages seem to consistently be over- or under-estimated?
- Are there any faculty members where the guesses seem to be in a particularly large or small range?

Section 2

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.

- Under what circumstances is MSE small?

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.

- Under what circumstances is MSE small?
- What are the problems with trying to 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

- **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

- **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

- **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

- **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

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

Training and Test Data

- **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

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

- If we have training and test data, 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 wish to predict students' final exam scores Y based on their first midterm scores X. We have data from two previous classes.

An Example

- Suppose we wish to predict students' final exam scores Y based on their first midterm scores X. We have data from two previous classes.
- Suppose e don't care about how well our model predicts exam scores for the previous classes. We want to know how well it predicts future scores.

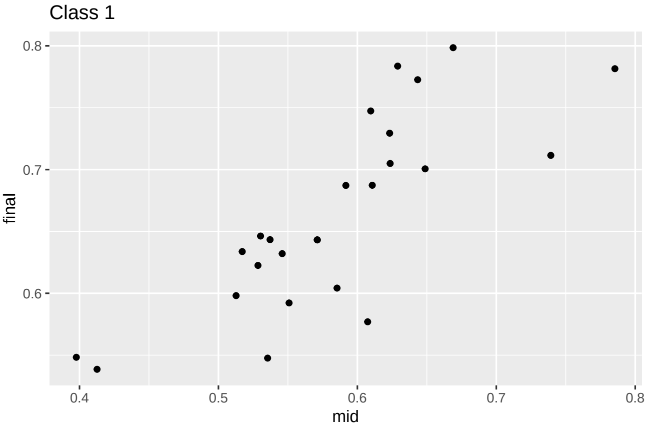
An Example

- Suppose we wish to predict students' final exam scores Y based on their first midterm scores X. We have data from two previous classes.

- Suppose e don't care about how well our model predicts exam scores for the previous classes. We want to know how well it predicts future scores.
  - Use the first class as training data
  - Use the second class as test data

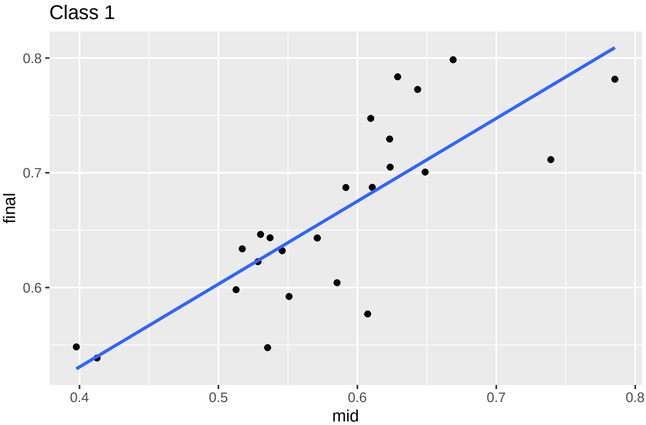
Training Set

```
##
##
scores %>% ggplot( aes(x = mid, y = final)) +
  geom_point()+labs(title = "Class 1")
```



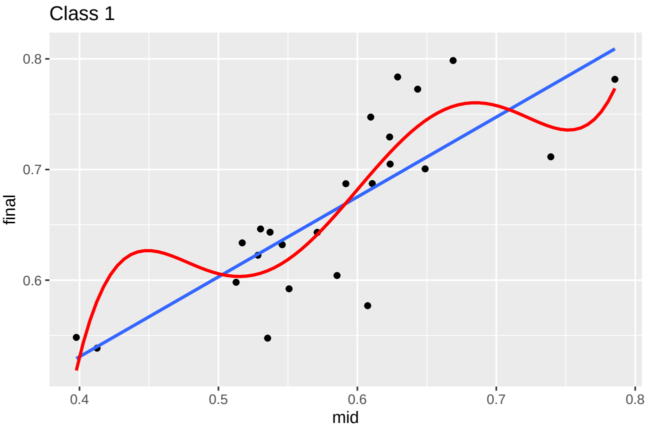
Model 1

```
##
scores %>% ggplot( aes(x = mid, y = final)) + geom_point()+
  labs(title = "Class 1") +
  geom_smooth( method = "lm" , se = FALSE)
```

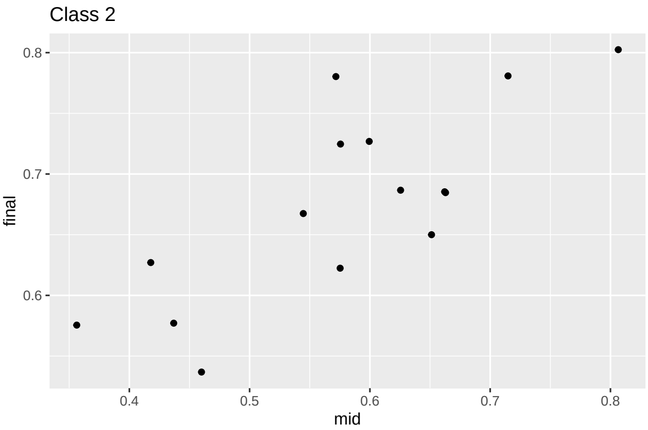


Model 1 and 2

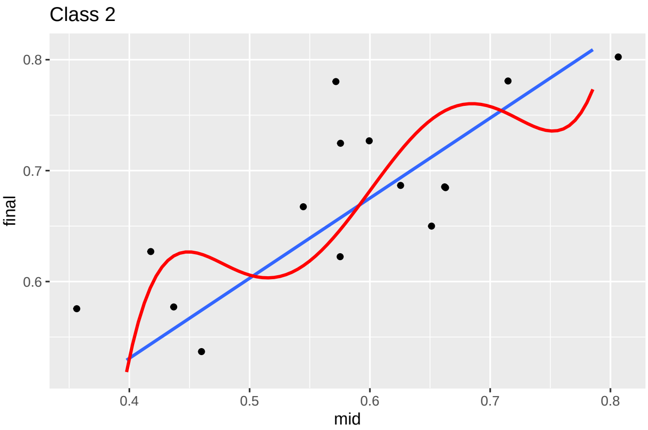
```
scores %>% ggplot( aes(x = mid, y = final)) + geom_point() +
  labs(title = "Class 1") +
  geom_smooth( method = "lm" , se = FALSE) +
  geom_smooth( method = "lm" ,formula = y ~ poly(x, 5), se = FALSE, color = "red")
```



Test Set



Test Set with models



MSE

Prediction accuracy

```
## # A tibble: 15 x 5
##    actual lin_pred poly_pred lin_sq_error poly_sq_error
##     <dbl>    <dbl>     <dbl>        <dbl>         <dbl>
##  1  0.537    0.574    0.625     0.00139       0.00771
##  2  0.687    0.694    0.718     0.0000487     0.000988
##  3  0.576    0.499    0.0801    0.00582       0.245
##  4  0.727    0.675    0.681     0.00271       0.00211
##  5  0.685    0.720    0.754     0.00121       0.00469
##  6  0.781    0.758    0.751     0.000515      0.000871
##  7  0.627    0.544    0.595     0.00695       0.00101
##  8  0.622    0.657    0.647     0.00122       0.000585
##  9  0.725    0.658    0.647     0.00450       0.00603
## 10  0.780    0.655    0.642     0.0157        0.0191
## 11  0.667    0.635    0.614     0.00104       0.00283
## 12  0.685    0.721    0.754     0.00129       0.00485
## 13  0.802    0.824    0.864     0.000478      0.00381
## 14  0.577    0.557    0.623     0.000387      0.00211
## 15  0.650    0.712    0.746     0.00387       0.00925
```

MSE

Prediction accuracy

```
## # A tibble: 15 x 5
##    actual lin_pred poly_pred lin_sq_error poly_sq_error
##     <dbl>    <dbl>     <dbl>        <dbl>         <dbl>
##  1  0.537    0.574    0.625     0.00139       0.00771
##  2  0.687    0.694    0.718     0.0000487     0.000988
##  3  0.576    0.499    0.0801    0.00582       0.245
##  4  0.727    0.675    0.681     0.00271       0.00211
##  5  0.685    0.720    0.754     0.00121       0.00469
##  6  0.781    0.758    0.751     0.000515      0.000871
##  7  0.627    0.544    0.595     0.00695       0.00101
##  8  0.622    0.657    0.647     0.00122       0.000585
##  9  0.725    0.658    0.647     0.00450       0.00603
## 10  0.780    0.655    0.642     0.0157        0.0191
## 11  0.667    0.635    0.614     0.00104       0.00283
## 12  0.685    0.721    0.754     0.00129       0.00485
## 13  0.802    0.824    0.864     0.000478      0.00381
## 14  0.577    0.557    0.623     0.000387      0.00211
## 15  0.650    0.712    0.746     0.00387       0.00925
```

Overall MSE

```
## # A tibble: 1 x 2
##   lin_mse poly_mse
##     <dbl>    <dbl>
## 1 0.00315   0.0208
```

Minimize MSE subject to model shape

What if no test data is available?

Minimize MSE subject to model shape

What if no test data is available?

- Recall the setting of simple linear regression from Math 141.

Minimize MSE subject to model shape

What if no test data is available?

- Recall the setting of simple linear regression from Math 141.

We can choose a model that minimizes MSE on the training set, subject to constraints (i.e. restricting to linear, quadratic, exponential models)

Minimize MSE subject to model shape

What if no test data is available?

- Recall the setting of simple linear regression from Math 141.

We can choose a model that minimizes MSE on the training set, subject to constraints (i.e. restricting to linear, quadratic, exponential models)

But no guarantee that model which minimizes MSE on training data will also do so on test data.

Minimize MSE subject to model shape

What if no test data is available?

- Recall the setting of simple linear regression from Math 141.

We can choose a model that minimizes MSE on the training set, subject to constraints (i.e. restricting to linear, quadratic, exponential models)

But no guarantee that model which minimizes MSE on training data will also do so on test data.

In fact, when selecting a complex model that minimizes MSE on the training data, the test MSE will often be very large!

Demo in RStudio

See .Rmd file (Wednesday 9-9 Demo) on the schedule page of the course website

Section 3

Bias-Variance Trade-off

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing complexity. What happens to the training MSE and the test MSE as model complexity increases?

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:

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

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

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:

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

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

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

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

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:

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

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

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 training sets

Bias and Variance

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

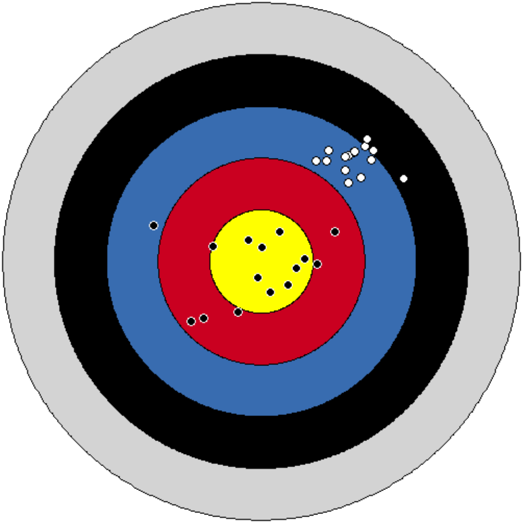
- **Bias** refers to amount by which the average of estimates $\hat{f}(x_0)$ differs from the true value of $f(x_0)$
  - Bias is produced by the difference between model shape assumptions and reality

Bias and Variance

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

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

Target Practice



The Trade-off

What is the problem?

The Trade-off

What is the problem?

How do we solve it?