Foundations of Statistical Learning

Nate Wells

Math 243: Stat Learning

September 3rd, 2021

Outline

In today's class, we will...

Outline

In today's class, we will...

- Discuss the goals of statistical learning algorithms

Outline

In today's class, we will...

- Discuss the goals of statistical learning algorithms
- Survey some of the most common methods for statistical learning

Outline

In today's class, we will...

- Discuss the goals of statistical learning algorithms
- Survey some of the most common methods for statistical learning
- Analyze data from the 'guess my age' activity

Section 1

What is Stat Learning?

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ for a population, and one or more response variables $Y_1, Y_2, \ldots$.

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ for a population, and one or more response variables $Y_1, Y_2, \ldots$.
  - Sometimes, we'll study the relationship among predictor variables in isolation (no response)

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ for a population, and one or more response variables $Y_1, Y_2, \ldots$.
  - Sometimes, we'll study the relationship among predictor variables in isolation (no response)
- In the simplest case, we observe the values of one quantitative response Y, as well as p many predictors $X_1, \ldots, X_p$.

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ for a population, and one or more response variables $Y_1, Y_2, \ldots$.
  - Sometimes, we'll study the relationship among predictor variables in isolation (no response)
- In the simplest case, we observe the values of one quantitative response Y, as well as p many predictors $X_1, \ldots, X_p$.
- We assume there is a (usually unknown) relationship between these observed values:

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

The Setting

- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ for a population, and one or more response variables $Y_1, Y_2, \ldots$.
  - Sometimes, we'll study the relationship among predictor variables in isolation (no response)
- In the simplest case, we observe the values of one quantitative response Y, as well as p many predictors $X_1, \ldots, X_p$.
- We assume there is a (usually unknown) relationship between these observed values:

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

- Here, ϵ denotes a random or unobserved error term **independent** of $X_1, \ldots, X_p$

The Setting

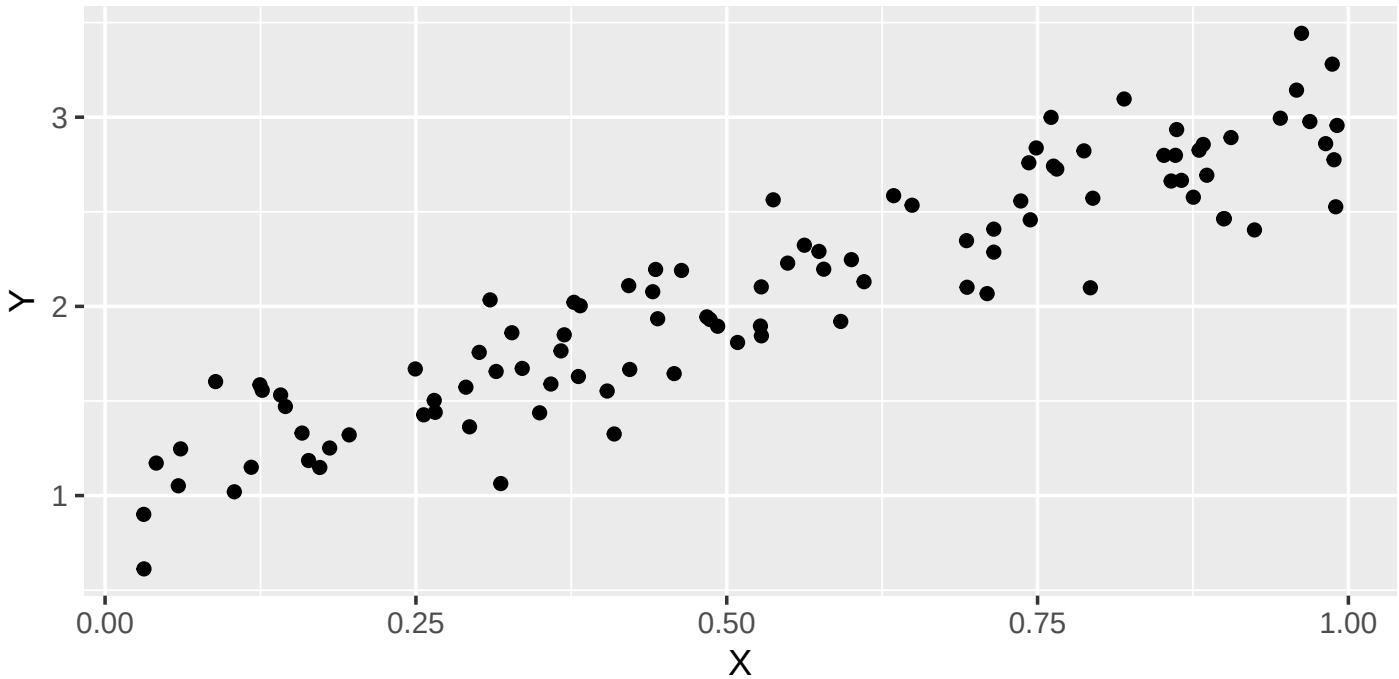
- Fundamentally, stat learning is the study of the relationships between predictor variables $X_1, \ldots, X_p$ for a population, and one or more response variables $Y_1, Y_2, \ldots$.
  - Sometimes, we'll study the relationship among predictor variables in isolation (no response)
- In the simplest case, we observe the values of one quantitative response Y, as well as p many predictors $X_1, \ldots, X_p$.
- We assume there is a (usually unknown) relationship between these observed values:

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

- Here, ϵ denotes a random or unobserved error term **independent** of $X_1, \ldots, X_p$
- The overarching goal of stat learning is to estimate f, given data on X and Y.

An Example

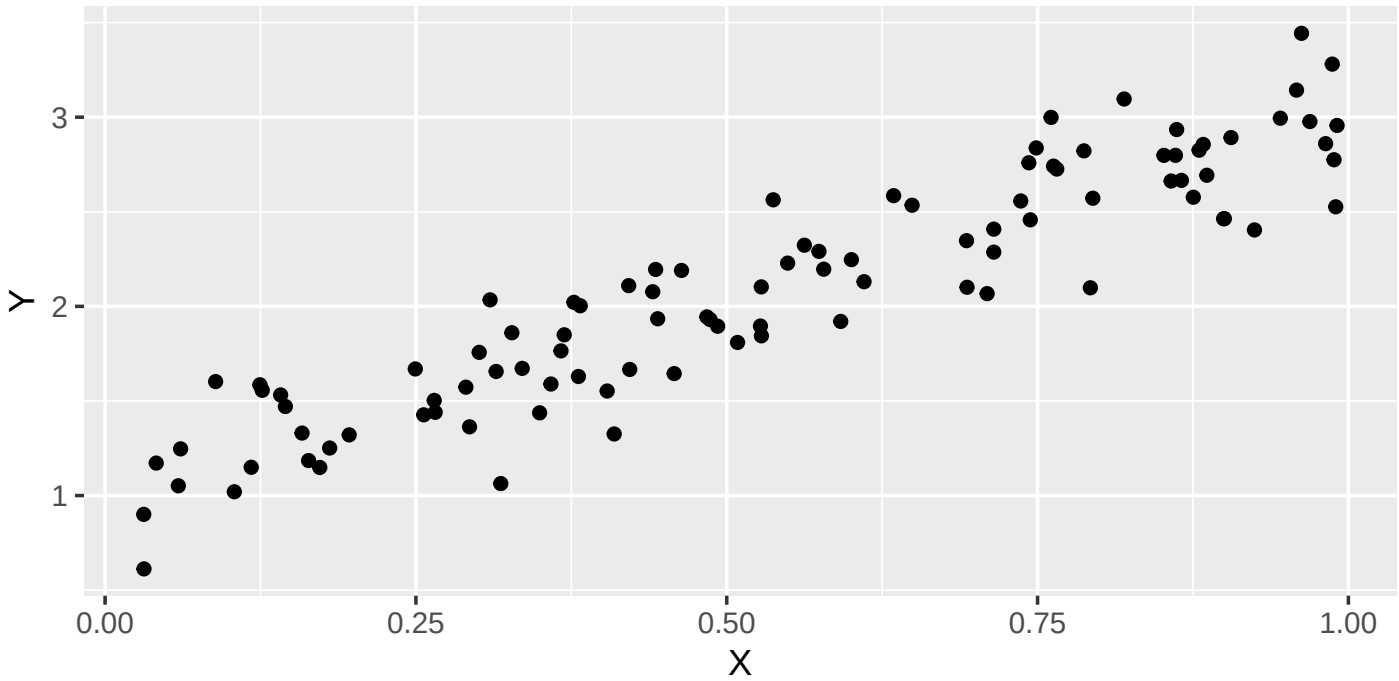
Consider the following observations for variables X and Y



What is the relationship between X and Y?

An Example

Consider the following observations for variables X and Y



What is the relationship between X and Y?

```
X = runif(100, 0,1 )
E = rnorm(100, 0, .25)
Y = 2*X + 1 +  E
```

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

- Suppose for each Reed faculty, we know the number of years X between when their undergrad degree was awarded and when their faculty picture was taken.

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

- Suppose for each Reed faculty, we know the number of years X between when their undergrad degree was awarded and when their faculty picture was taken.
- Ultimately, we want to estimate the age Y of each faculty member.

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

- Suppose for each Reed faculty, we know the number of years X between when their undergrad degree was awarded and when their faculty picture was taken.
- Ultimately, we want to estimate the age Y of each faculty member.
- To do so, we theorize a model that takes in X as input and outputs our best guess $\hat{Y}$ for Y.

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

- Suppose for each Reed faculty, we know the number of years X between when their undergrad degree was awarded and when their faculty picture was taken.
- Ultimately, we want to estimate the age Y of each faculty member.
- To do so, we theorize a model that takes in X as input and outputs our best guess $\hat{Y}$ for Y.
    - What is one such possible model f?

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

- Suppose for each Reed faculty, we know the number of years X between when their undergrad degree was awarded and when their faculty picture was taken.
- Ultimately, we want to estimate the age Y of each faculty member.
- To do so, we theorize a model that takes in X as input and outputs our best guess $\hat{Y}$ for Y.
  - What is one such possible model f?
- But even if we have a perfect estimate for f in $Y = f(X) + \epsilon$, the predicted value $\hat{Y} = f(X)$ of Y may not equal Y, since Y also depends on ϵ.

Estimating f for Prediction

Prediction is useful in settings where X can be observed, but Y cannot. Ex:

- Suppose for each Reed faculty, we know the number of years X between when their undergrad degree was awarded and when their faculty picture was taken.
- Ultimately, we want to estimate the age Y of each faculty member.
- To do so, we theorize a model that takes in X as input and outputs our best guess $\hat{Y}$ for Y.
  - What is one such possible model f?
- But even if we have a perfect estimate for f in $Y = f(X) + \epsilon$, the predicted value $\hat{Y} = f(X)$ of Y may not equal Y, since Y also depends on ϵ.
  - What is one source of error ϵ in the previous model?

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

1. Reducible error, in the form of our estimate $\hat{f}$ for f.

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

1. Reducible error, in the form of our estimate $\hat{f}$ for f.
2. Irreducible error, in the form of ϵ

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

1. Reducible error, in the form of our estimate $\hat{f}$ for f.
2. Irreducible error, in the form of ϵ

What steps can be taken to improve reducible error?

Types of Error

In general, there are two sources of error in a model $\hat{Y} = \hat{f}(X_1, \ldots, X_p) + \epsilon$ for the relationship

$$Y = f(X_1, \ldots, X_p) + \epsilon$$

1. Reducible error, in the form of our estimate $\hat{f}$ for f.
2. Irreducible error, in the form of ϵ

What steps can be taken to improve reducible error?

What about irreducible error?

Inference

In many settings, we are interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response Y.

Inference

In many settings, we are interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response Y.

1. Which predictors are likely to be associated with response?

Inference

In many settings, we are interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response Y.

1. Which predictors are likely to be associated with response?
2. What is the degree and strength of the relationship between significant predictors and the response?

Inference

In many settings, we are interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response Y.

1. Which predictors are likely to be associated with response?
2. What is the degree and strength of the relationship between significant predictors and the response?
3. What type of relationship exists between the predictors and the response? (Linear? Exponential? Something more complicated?)

Inference

In many settings, we are interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response Y.

1. Which predictors are likely to be associated with response?
2. What is the degree and strength of the relationship between significant predictors and the response?
3. What type of relationship exists between the predictors and the response? (Linear? Exponential? Something more complicated?)

Ex:

A data set contains information on a professor's age, gender, tenure-status, ethnicity, and department. Which of these predictors are associated with course evaluation scores, and how?

Inference

In many settings, we are interested in the relationship between each predictor $X_1, \ldots, X_p$ and the response Y.

1. Which predictors are likely to be associated with response?
2. What is the degree and strength of the relationship between significant predictors and the response?
3. What type of relationship exists between the predictors and the response? (Linear? Exponential? Something more complicated?)

Ex:

A data set contains information on a professor's age, gender, tenure-status, ethnicity, and department. Which of these predictors are associated with course evaluation scores, and how?

Here, we are trying to **infer** information about the factors which contribute to course eval score.

Section 2

Methods of Stat Learning

Parametric Methods

Parametric methods for estimating f involve two steps:

1. Based on domain knowledge, make assumptions about the functional form or shape of f.

Parametric Methods

Parametric methods for estimating f involve two steps:

1. Based on domain knowledge, make assumptions about the functional form or shape of f.
- The linear model is a common choice for the shape of f:

$$f(X) = \beta_0 + \beta_1 X_1 \quad \text{simple linear}$$
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{multilinear}$$

Parametric Methods

Parametric methods for estimating f involve two steps:

1. Based on domain knowledge, make assumptions about the functional form or shape of f.
   - The linear model is a common choice for the shape of f:

$$f(X) = \beta_0 + \beta_1 X_1 \quad \text{simple linear}$$
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{multilinear}$$

2. After a model has been chosen, we implement a procedure for estimating the **parameters** of the model that minimizes the reducible error.

Parametric Methods

Parametric methods for estimating f involve two steps:

1. Based on domain knowledge, make assumptions about the functional form or shape of f.
   - The linear model is a common choice for the shape of f:

$$f(X) = \beta_0 + \beta_1 X_1 \quad \text{simple linear}$$
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{multilinear}$$

2. After a model has been chosen, we implement a procedure for estimating the **parameters** of the model that minimizes the reducible error.
   - In the case of the linear model, we estimate the values of $\beta_0, \ldots, \beta_p$ using the *method of least squares*.

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions.

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions.

- In doing so, non-parametric models avoid the problem of mischaracterizing the relationship between predictors and response

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions.

- In doing so, non-parametric models avoid the problem of mischaracterizing the relationship between predictors and response
- However, non-parametric models run the risk of **overfitting**, where the model closely matches the observed data, but does not represent the true unobserved relationship between the variables

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions.

- In doing so, non-parametric models avoid the problem of mischaracterizing the relationship between predictors and response
- However, non-parametric models run the risk of **overfitting**, where the model closely matches the observed data, but does not represent the true unobserved relationship between the variables
  - How is this possible?

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions.

- In doing so, non-parametric models avoid the problem of mischaracterizing the relationship between predictors and response
- However, non-parametric models run the risk of **overfitting**, where the model closely matches the observed data, but does not represent the true unobserved relationship between the variables
  - How is this possible?
- Non-parametric models often require orders of magnitude more data to make accurate predictions, compared to parametric models

Non-parametric Methods

Non-parametric methods forgo assumptions on the shape of f, working instead in a very general class of functions.

- In doing so, non-parametric models avoid the problem of mischaracterizing the relationship between predictors and response
- However, non-parametric models run the risk of **overfitting**, where the model closely matches the observed data, but does not represent the true unobserved relationship between the variables
  - How is this possible?
- Non-parametric models often require orders of magnitude more data to make accurate predictions, compared to parametric models
- Some examples of non-parametric models include: Spline Regression, Support Vector Machines, and Neural Networks

Techniques and Problems

Most statistical learning **techniques** fall into one of two categories:

Techniques and Problems

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables.
   - Because we have both predicted and actual values of response, we can assess the accuracy of the model.
2. Unsupervised learning, in which patterns and trends are detected in the predictors without reference to a response variable.
   - There is no available metric to determine when the model is performing "well"

Techniques and Problems

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables.
    - Because we have both predicted and actual values of response, we can assess the accuracy of the model.
2. Unsupervised learning, in which patterns and trends are detected in the predictors without reference to a response variable.
    - There is no available metric to determine when the model is performing "well"

Statistical learning **problems** also fall into a pair of categories:

Techniques and Problems

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables.
    - Because we have both predicted and actual values of response, we can assess the accuracy of the model.
2. Unsupervised learning, in which patterns and trends are detected in the predictors without reference to a response variable.
    - There is no available metric to determine when the model is performing "well"

Statistical learning **problems** also fall into a pair of categories:

1. Regression problems, wherein we measure the magnitude of a **quantitative** response variable

Techniques and Problems

Most statistical learning **techniques** fall into one of two categories:

1. Supervised learning, in which predictors are compared with one or more response variables.
    - Because we have both predicted and actual values of response, we can assess the accuracy of the model.
2. Unsupervised learning, in which patterns and trends are detected in the predictors without reference to a response variable.
    - There is no available metric to determine when the model is performing "well"

Statistical learning **problems** also fall into a pair of categories:

1. Regression problems, wherein we measure the magnitude of a **quantitative** response variable
2. Classification problems, wherein we sort a **qualitative** response variable into several discrete classes.

Section 3

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).

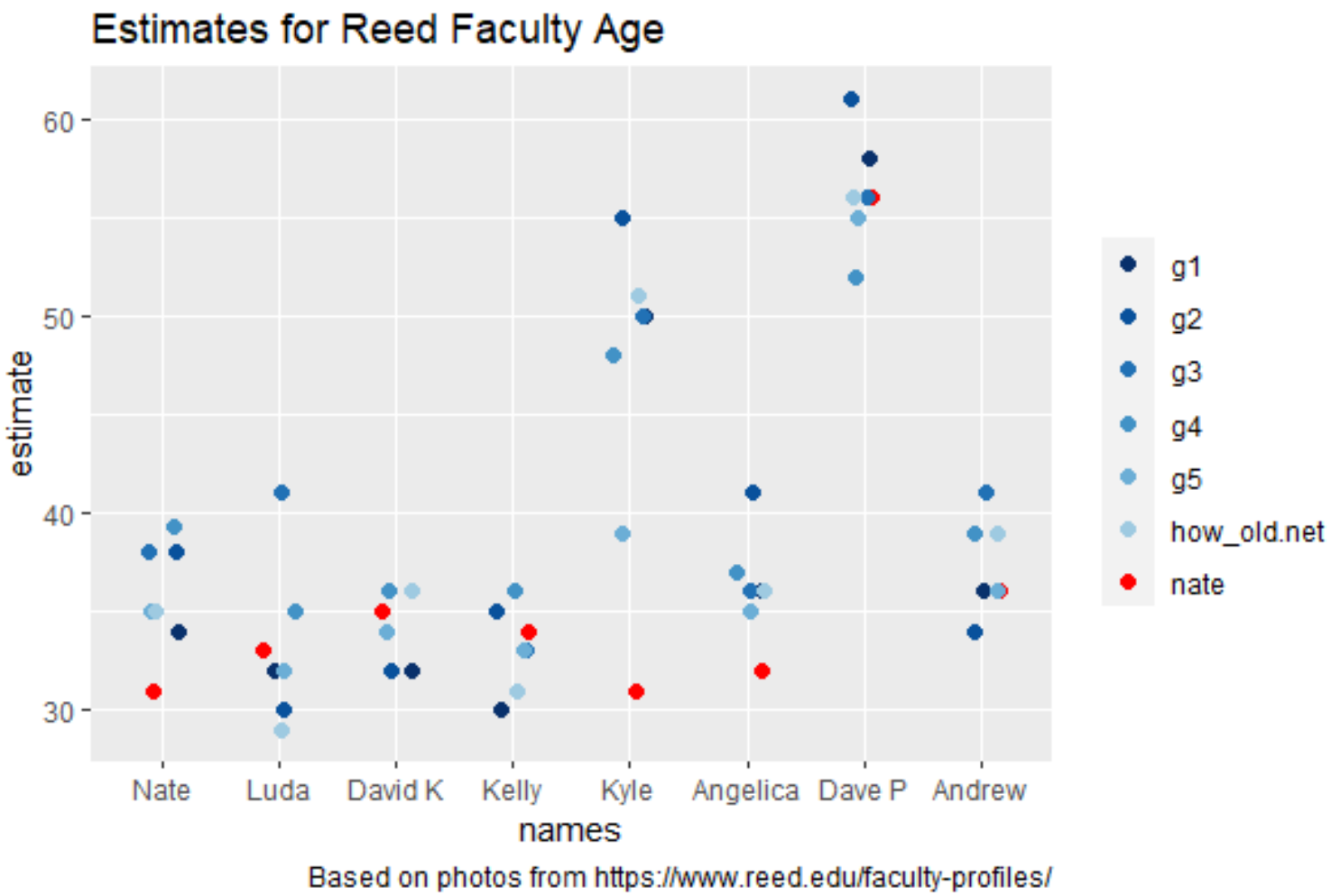
- Was the *How Old?* activity supervised or unsupervised?
- Did it represent a classification or regression problem?
- Were you interested primarily in prediction or inference?

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

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?