Introduction to Data Science

# Basics of Modeling

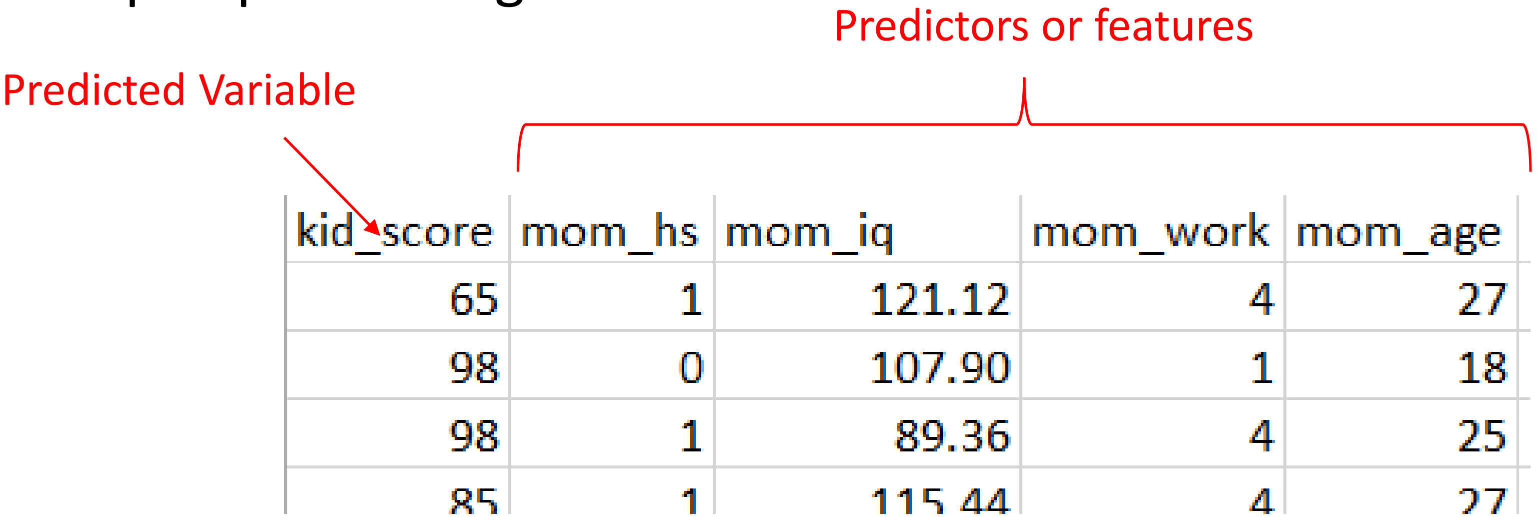
Gordon Anderson

# What is a Model?

- A model is a generalization that captures a fundamental trend that occurs in a set of observations.
- The model is useful because it can tell us interesting aspects about the trend.
- It is also useful in that it can predict something about new, previously unseen observations.
- Simpler models are preferred over complex models as long as they have similar performance.

# Data Analysis and Modeling

- Data: a set of observations.
- Each row is called a "tuple" of observations.
- One variable is the dependent variable, or the predicted variable.
- The independent variables are the "predictors".
- Example: predict cognitive scores of children based on mother's characteristics.



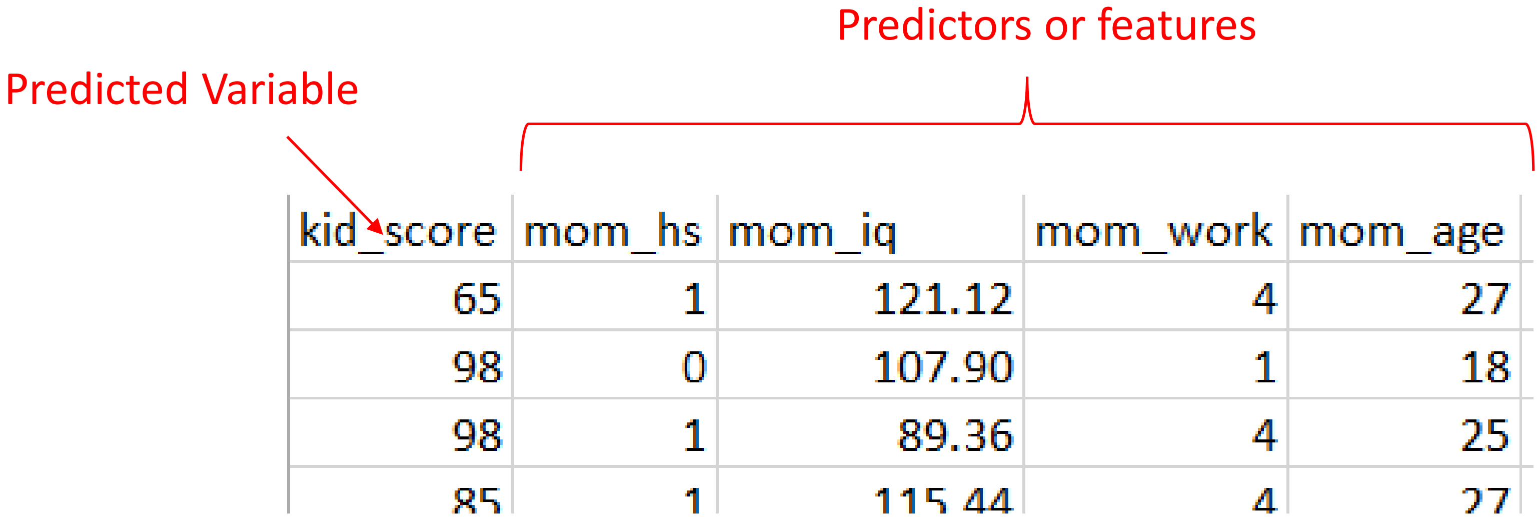
Predictors or features

Predicted Variable

| kid_score | mom_hs | mom_iq | mom_work | mom_age |
|---|---|---|---|---|
| 65 | 1 | 121.12 | 4 | 27 |
| 98 | 0 | 107.90 | 1 | 18 |
| 98 | 1 | 89.36 | 4 | 25 |
| 85 | 1 | 115.44 | 4 | 27 |

# Modeling tasks

- Regression: Given predictors, predict a continuous (numeric) value.



Predictors or features

Predicted Variable

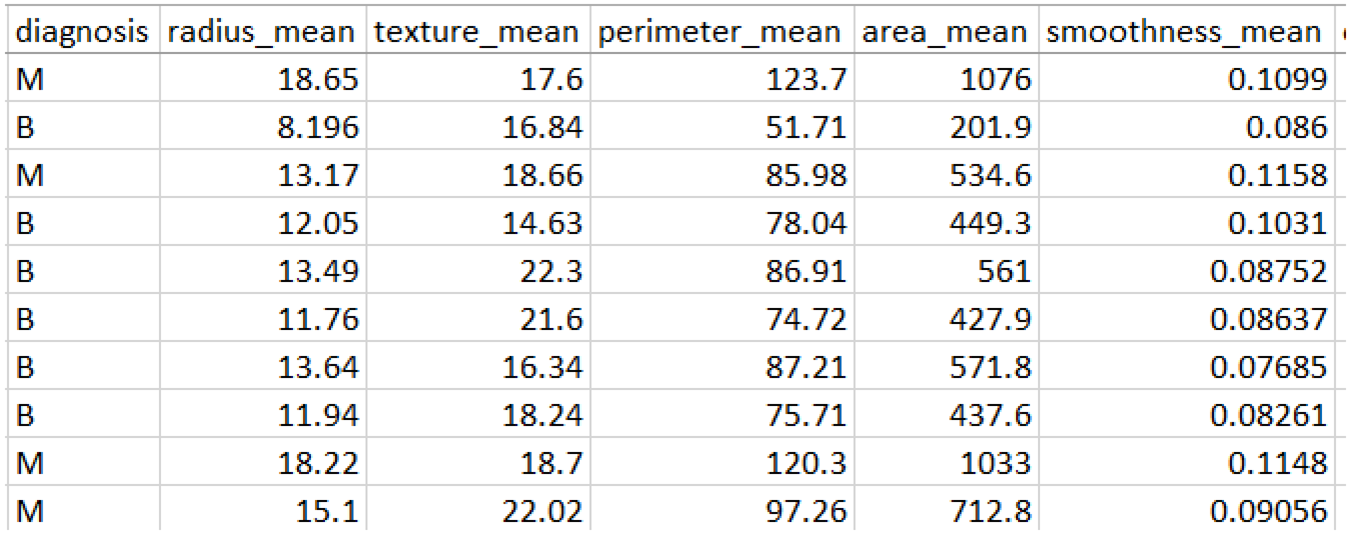
| kid_score | mom_hs | mom_iq | mom_work | mom_age |
|---|---|---|---|---|
| 65 | 1 | 121.12 | 4 | 27 |
| 98 | 0 | 107.90 | 1 | 18 |
| 98 | 1 | 89.36 | 4 | 25 |
| 85 | 1 | 115.44 | 4 | 27 |

# Modeling tasks

- Classification: Given predictors, predict a categorical value (label).

Predictors

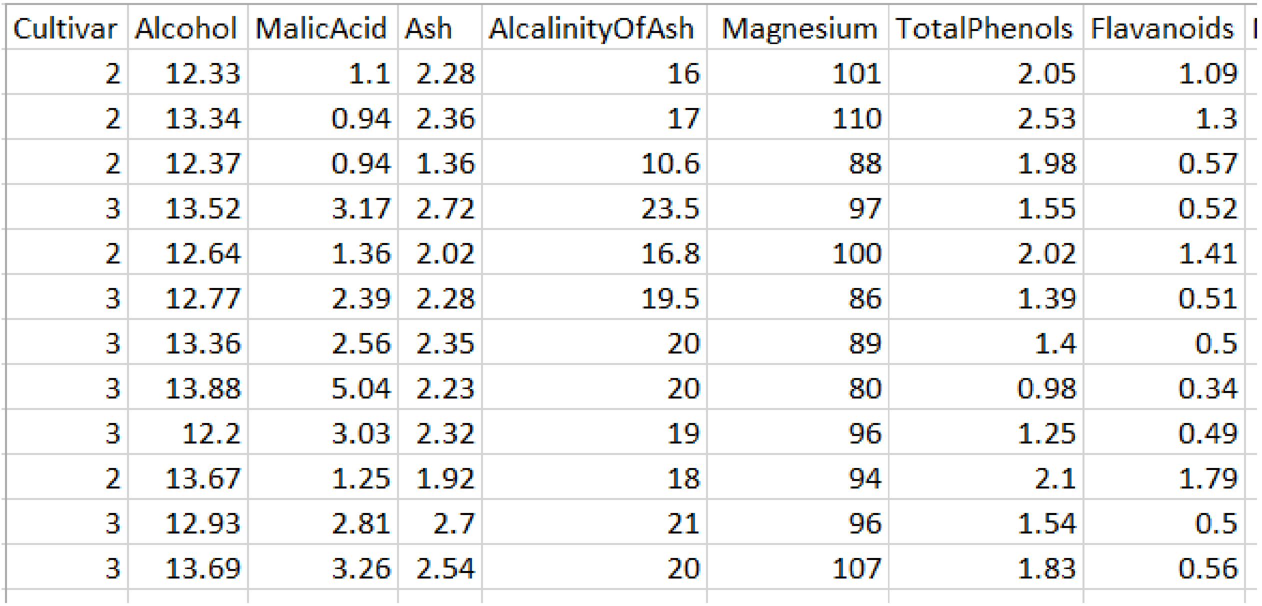
Predicted Variable
or “label”



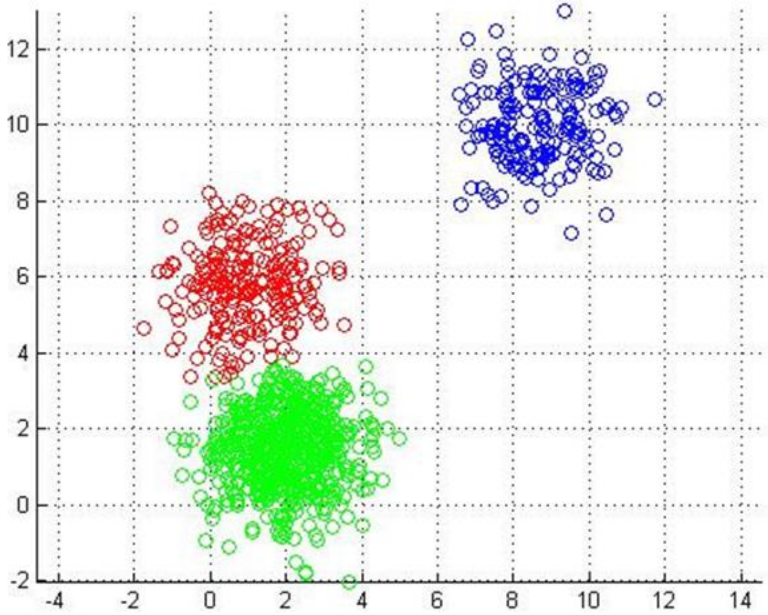
| diagnosis | radius_mean | texture_mean | perimeter_mean | area_mean | smoothness_mean |
|---|---|---|---|---|---|
| M | 18.65 | 17.6 | 123.7 | 1076 | 0.1099 |
| B | 8.196 | 16.84 | 51.71 | 201.9 | 0.086 |
| M | 13.17 | 18.66 | 85.98 | 534.6 | 0.1158 |
| B | 12.05 | 14.63 | 78.04 | 449.3 | 0.1031 |
| B | 13.49 | 22.3 | 86.91 | 561 | 0.08752 |
| B | 11.76 | 21.6 | 74.72 | 427.9 | 0.08637 |
| B | 13.64 | 16.34 | 87.21 | 571.8 | 0.07685 |
| B | 11.94 | 18.24 | 75.71 | 437.6 | 0.08261 |
| M | 18.22 | 18.7 | 120.3 | 1033 | 0.1148 |
| M | 15.1 | 22.02 | 97.26 | 712.8 | 0.09056 |

# Modeling tasks

- Clustering: Given data, find “natural” groupings or clusters.



| Cultivar | Alcohol | MalicAcid | Ash | AlcalinityOfAsh | Magnesium | TotalPhenols | Flavanoids | I |
|---|---|---|---|---|---|---|---|---|
| 2 | 12.33 | 1.1 | 2.28 | 16 | 101 | 2.05 | 1.09 | |
| 2 | 13.34 | 0.94 | 2.36 | 17 | 110 | 2.53 | 1.3 | |
| 2 | 12.37 | 0.94 | 1.36 | 10.6 | 88 | 1.98 | 0.57 | |
| 3 | 13.52 | 3.17 | 2.72 | 23.5 | 97 | 1.55 | 0.52 | |
| 2 | 12.64 | 1.36 | 2.02 | 16.8 | 100 | 2.02 | 1.41 | |
| 3 | 12.77 | 2.39 | 2.28 | 19.5 | 86 | 1.39 | 0.51 | |
| 3 | 13.36 | 2.56 | 2.35 | 20 | 89 | 1.4 | 0.5 | |
| 3 | 13.88 | 5.04 | 2.23 | 20 | 80 | 0.98 | 0.34 | |
| 3 | 12.2 | 3.03 | 2.32 | 19 | 96 | 1.25 | 0.49 | |
| 2 | 13.67 | 1.25 | 1.92 | 18 | 94 | 2.1 | 1.79 | |
| 3 | 12.93 | 2.81 | 2.7 | 21 | 96 | 1.54 | 0.5 | |
| 3 | 13.69 | 3.26 | 2.54 | 20 | 107 | 1.83 | 0.56 | |



# A review of the inference process:

- Given: the question you want answered has been defined.

Do:

- Step 1: make observations- collect data.
- Step 2: analysis- make ***models***, evaluate them.
- Step 3: draw conclusion, ask: has the question been answered?
- Step 4: communicate results- graphs, charts.

We’ll look into step 2: modeling.

# The Modeling Process

Given a set of data:

1. Choose a type of model.
2. Fit the model to the data (algorithm “learns” model parameters).
3. Evaluate the quality of the model fit.
4. If necessary, go to step 2, else go to next step.
5. Evaluate model- either for prediction or exploring variable interactions.

# Machine Learning- ML

- ML occurs when we use an algorithm to "learn" the parameters of a model we are using.

- There three three basic types of ML:
    1. Supervised: the algorithm gets "training" data in the form of correct outcome and its predictors. (Called "labeled" data). The resulting model learns to predict the outcome based on the predictors.
    2. Unsupervised: the algorithm does not get examples of correct outcomes. (Called "unlabeled" data).
    3. Semi-supervised: Some of the data is labeled. The algorithm has to deal with incompletely labeled data.

# Modeling tasks and ML Modeling steps:

1. Regression- Supervised
2. Classification- Supervised
3. Clustering- Unsupervised

Supervised ML analysis creates training and testing subsets of the data set.

The training set contains examples of predictors and outcomes.

The ML algorithm learns to predict the outcome based on the predictors.

This is called “fitting” the model to the data.

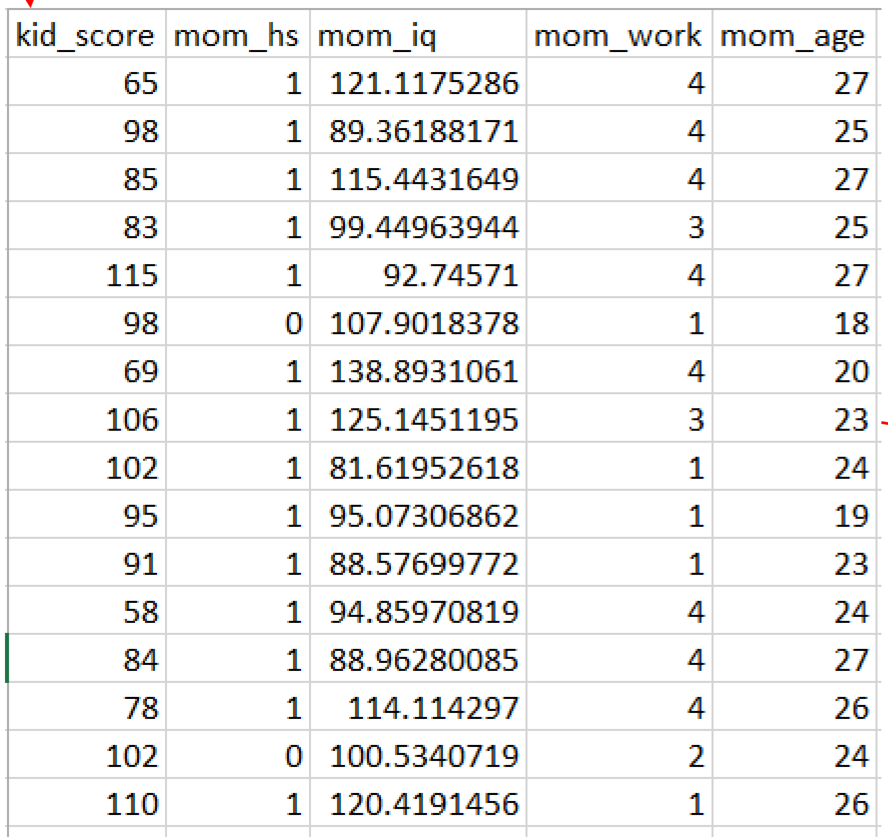
The resulting model is then given the predictors from the test set and its predicted outcomes are compared with the actual outcomes.

# ML Modeling Data

Entire data set:

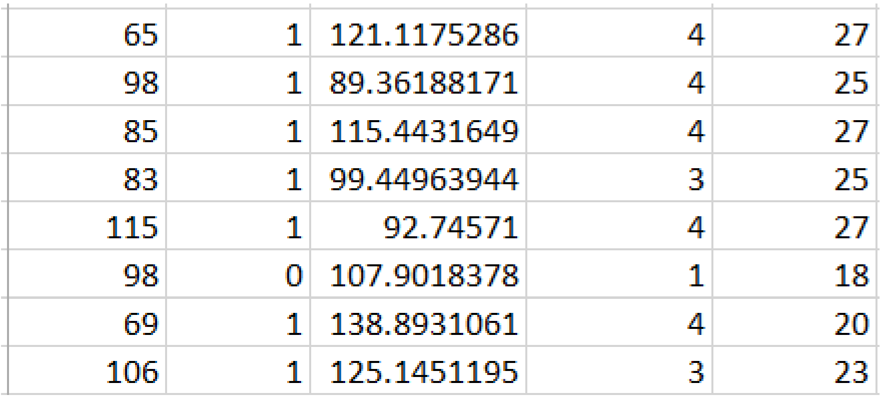
Predictors

Outcome



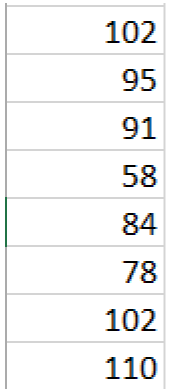
| kid_score | mom_hs | mom_iq | mom_work | mom_age |
|---|---|---|---|---|
| 65 | 1 | 121.1175286 | 4 | 27 |
| 98 | 1 | 89.36188171 | 4 | 25 |
| 85 | 1 | 115.4431649 | 4 | 27 |
| 83 | 1 | 99.44963944 | 3 | 25 |
| 115 | 1 | 92.74571 | 4 | 27 |
| 98 | 0 | 107.9018378 | 1 | 18 |
| 69 | 1 | 138.8931061 | 4 | 20 |
| 106 | 1 | 125.1451195 | 3 | 23 |
| 102 | 1 | 81.61952618 | 1 | 24 |
| 95 | 1 | 95.07306862 | 1 | 19 |
| 91 | 1 | 88.57699772 | 1 | 23 |
| 58 | 1 | 94.85970819 | 4 | 24 |
| 84 | 1 | 88.96280085 | 4 | 27 |
| 78 | 1 | 114.114297 | 4 | 26 |
| 102 | 0 | 100.5340719 | 2 | 24 |
| 110 | 1 | 120.4191456 | 1 | 26 |

Training set:
<outcome, predictors>



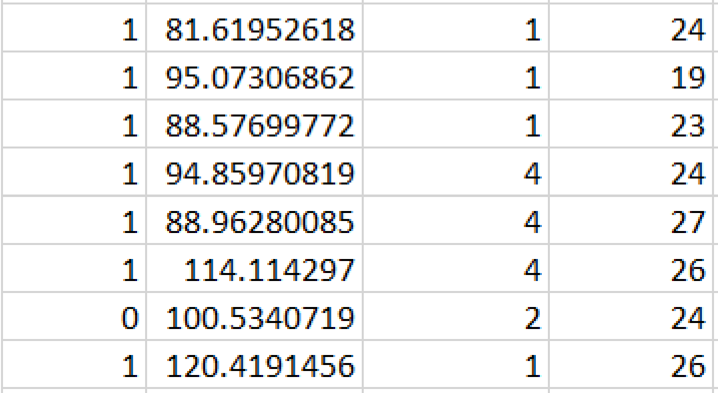
| | | | | |
|---|---|---|---|---|
| 65 | 1 | 121.1175286 | 4 | 27 |
| 98 | 1 | 89.36188171 | 4 | 25 |
| 85 | 1 | 115.4431649 | 4 | 27 |
| 83 | 1 | 99.44963944 | 3 | 25 |
| 115 | 1 | 92.74571 | 4 | 27 |
| 98 | 0 | 107.9018378 | 1 | 18 |
| 69 | 1 | 138.8931061 | 4 | 20 |
| 106 | 1 | 125.1451195 | 3 | 23 |

“True” outcomes:



| |
|---|
| 102 |
| 95 |
| 91 |
| 58 |
| 84 |
| 78 |
| 102 |
| 110 |

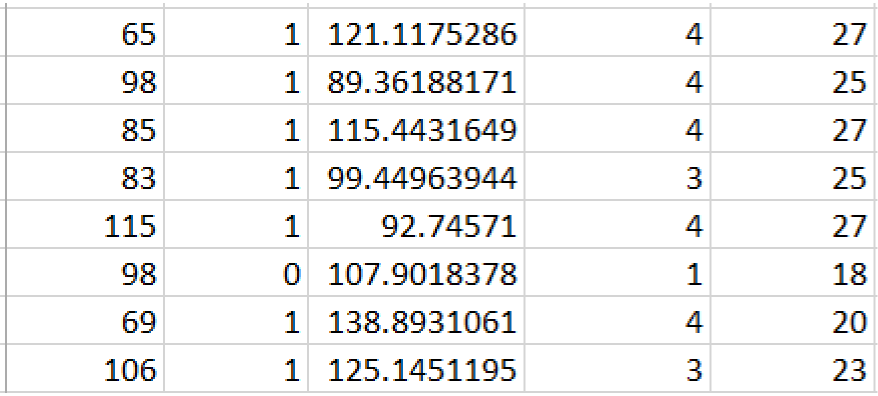
Testing set:
<predictors>, model outputs predictions



| | | | |
|---|---|---|---|
| 1 | 81.61952618 | 1 | 24 |
| 1 | 95.07306862 | 1 | 19 |
| 1 | 88.57699772 | 1 | 23 |
| 1 | 94.85970819 | 4 | 24 |
| 1 | 88.96280085 | 4 | 27 |
| 1 | 114.114297 | 4 | 26 |
| 0 | 100.5340719 | 2 | 24 |
| 1 | 120.4191456 | 1 | 26 |

# ML Modeling Steps

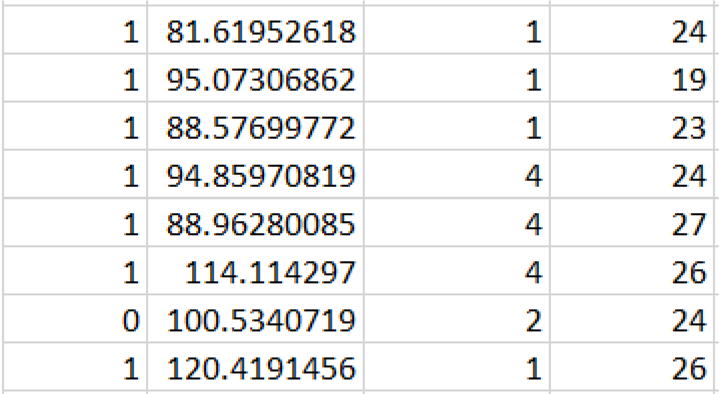
Training set:
<outcome, predictors>



| | | | | |
|---|---|---|---|---|
| 65 | 1 | 121.1175286 | 4 | 27 |
| 98 | 1 | 89.36188171 | 4 | 25 |
| 85 | 1 | 115.4431649 | 4 | 27 |
| 83 | 1 | 99.44963944 | 3 | 25 |
| 115 | 1 | 92.74571 | 4 | 27 |
| 98 | 0 | 107.9018378 | 1 | 18 |
| 69 | 1 | 138.8931061 | 4 | 20 |
| 106 | 1 | 125.1451195 | 3 | 23 |

1) Model is "fit" to the training data.

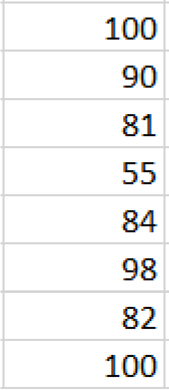
Testing set:
<predictors>, model outputs predictions



| | | | |
|---|---|---|---|
| 1 | 81.61952618 | 1 | 24 |
| 1 | 95.07306862 | 1 | 19 |
| 1 | 88.57699772 | 1 | 23 |
| 1 | 94.85970819 | 4 | 24 |
| 1 | 88.96280085 | 4 | 27 |
| 1 | 114.114297 | 4 | 26 |
| 0 | 100.5340719 | 2 | 24 |
| 1 | 120.4191456 | 1 | 26 |

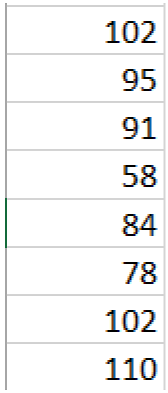
2) Model is given test data, outputs its predictions.

Predicted outcomes:



| |
|---|
| 100 |
| 90 |
| 81 |
| 55 |
| 84 |
| 98 |
| 82 |
| 100 |

"True" outcomes:

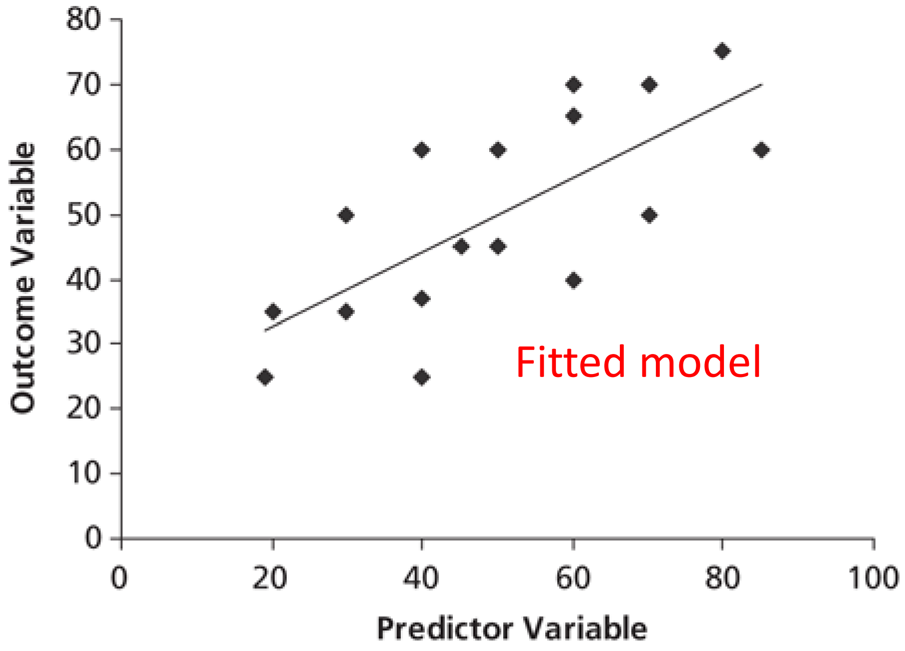


| |
|---|
| 102 |
| 95 |
| 91 |
| 58 |
| 84 |
| 78 |
| 102 |
| 110 |

3) Compare predicted with true outcomes to evaluate model performance.

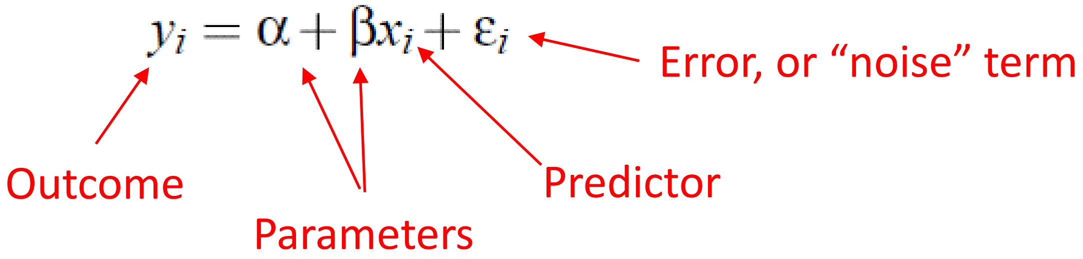
# Model Example

A model generalizes a theory from the observed data.



Fitted model

Assume a linear model can describe the data.
The general model formula (each “*i*” is a row of data):



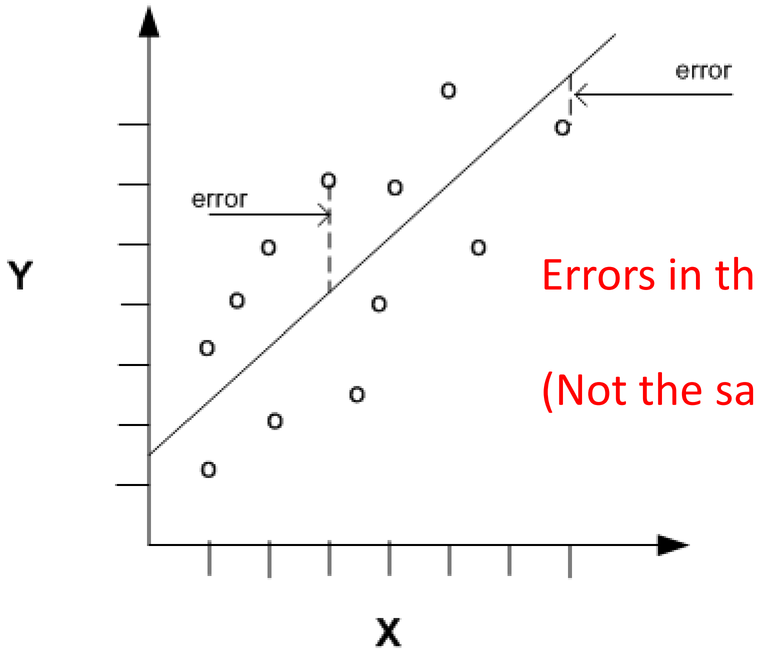
$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Error, or “noise” term

Outcome

Parameters

Predictor

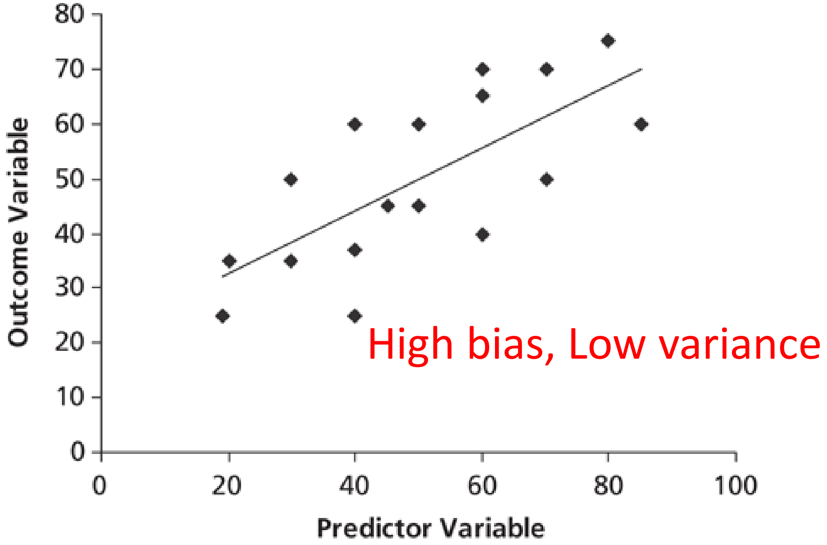


Errors in the fit.

(Not the same as errors in prediction)

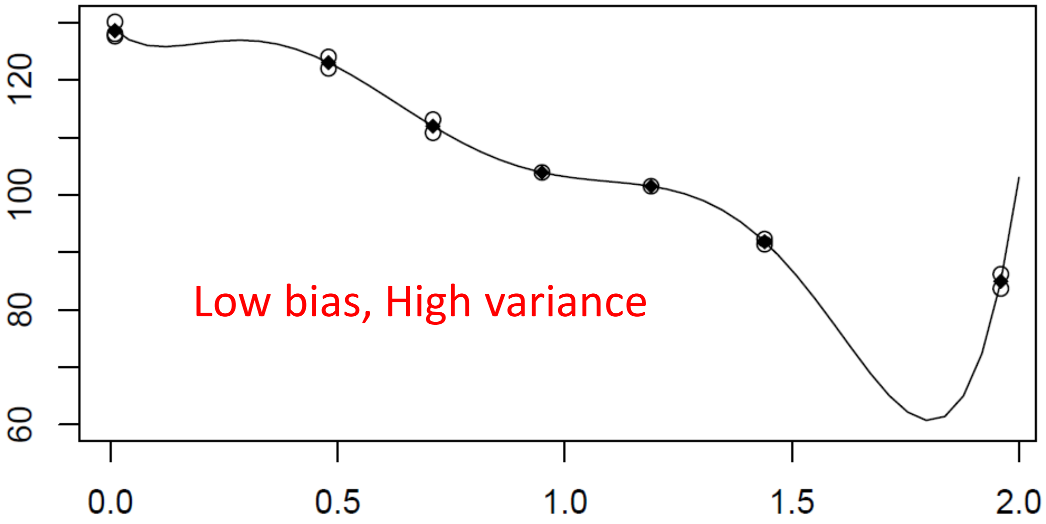
# Model Fit- bias vs. variance

Modeling is a trade-off between fitting the training set well and generalizing enough to predict new data well.



High bias, Low variance

Error in the fit, but hopefully it generalized well. That is determined when we test its predictions.



Low bias, High variance

A perfect fit, but no generalization: This model has ***overfit*** these data and Is most likely a poor predictor.

# Modeling

Modeling is not an exact science- we choose models based on previous experience and by trial and error:

pick the type of model, fit to data, evaluate fit, repeat

There are many modeling techniques, and we will be learning and practicing some of them.

Always be thinking about what biases and assumptions you are dealing with. What are the “threats to validity” inherent in your model and the data you are using?