Model Screening: Streamlining the Predictive Modeling Workflow

Discovery US, Oct 2021

Mia Stephens

JMP Principal Product Manager

STATISTICAL DISCOVERY FROM SAS

Abstract

Predictive modeling is all about finding the model, or combination of models, that most accurately predicts the outcome of interest. But, not all problems (and data) are created equal. For any given scenario, there several possible predictive models you can fit, and, no one type of model works best for all problems. In some cases, a regression model might be the top performer, in others it might be a tree-based model or a neural network.

In the search for the best performing model, you might fit all of the available models, one at a time, using cross-validation. Then, you might save the individual models to the data table, or to the Formula Depot, and then use Model Comparison to compare the performance of the models on the validation set to select the best one. Now, with the new Model Screening platform in JMP Pro 16, this workflow has been streamlined. In this talk, you'll learn how to use Model Screening to simultaneously fit, validate, compare, explore, select and then deploy the best performing predictive model.

Outline

- What is predictive modeling?
- Types of models can we build
- The predictive modeling workflow
- Using Model Screening to streamline
- Metrics for comparing model performance
- Examples

What is Predictive Modeling?

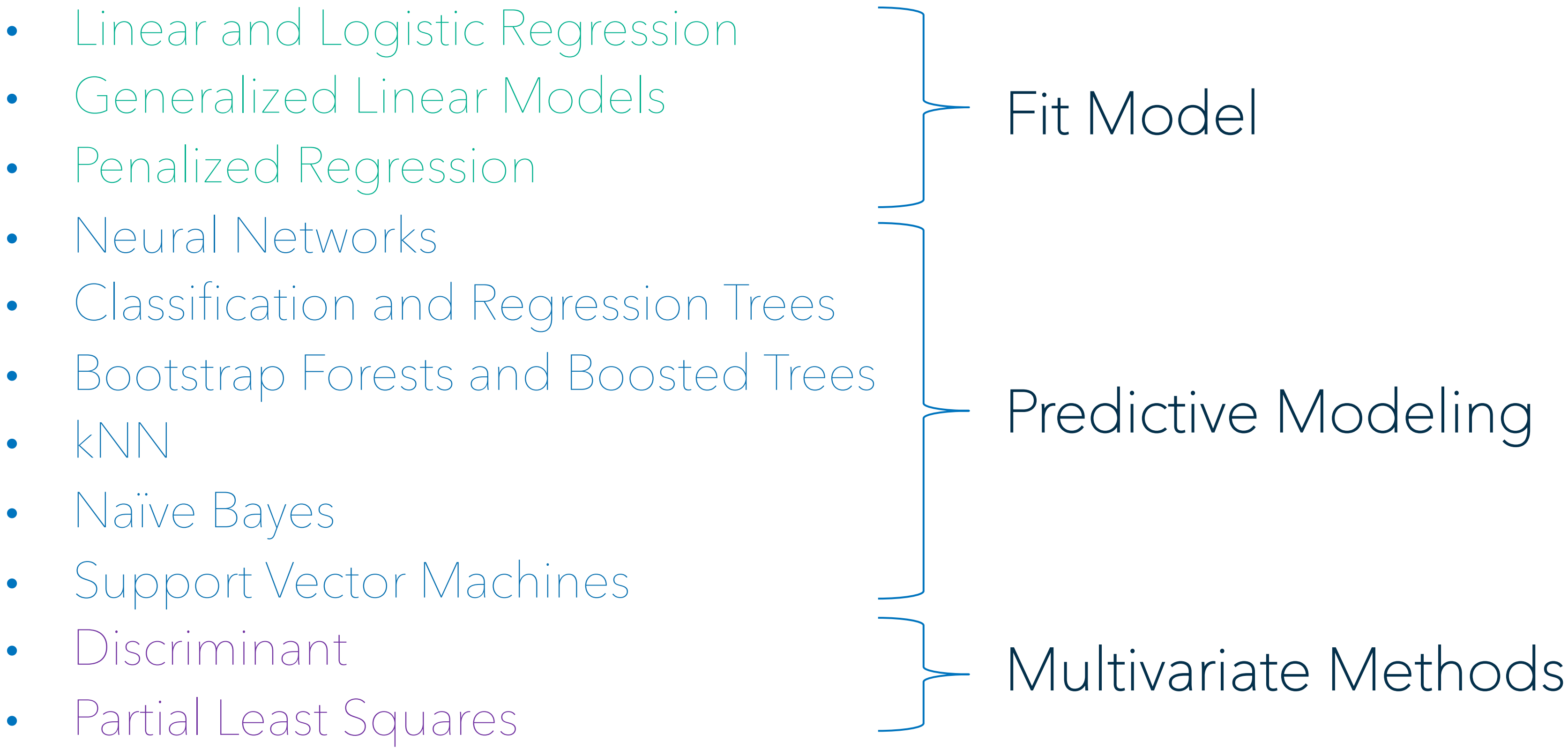
Explanatory Modeling

- Y = f(X's)
- Identify important variables.
  - Example: X1, X3, and X6 are potential causes of variation in Y.
- Understand how Y changes, on average, as a function of the X's.
  - Example: A 1 unit change in X is associated with a 5 unit change in Y.

Predictive Modeling

- Accurately predict or classify future outcomes.
- Fit and compare many models.
- Advanced techniques, not easy to interpret.
- Overfitting can be a problem.
- Use validation for model comparison and to protect against over- and under-fitting.

Types of Predictive Models



- Linear and Logistic Regression
- Generalized Linear Models
- Penalized Regression
- Neural Networks
- Classification and Regression Trees
- Bootstrap Forests and Boosted Trees
- kNN
- Naïve Bayes
- Support Vector Machines
- Discriminant
- Partial Least Squares

Fit Model

Predictive Modeling

Multivariate Methods

*Not an exhaustive list

Types of Predictive Models

- Linear and Logistic Regression
- Generalized Linear Models
- Penalized Regression
- Neural Networks
- Classification and Regression Trees
- Bootstrap Forests and Boosted Trees
- kNN
- Naïve Bayes
- Support Vector Machines
- Discriminant
- Partial Least Squares

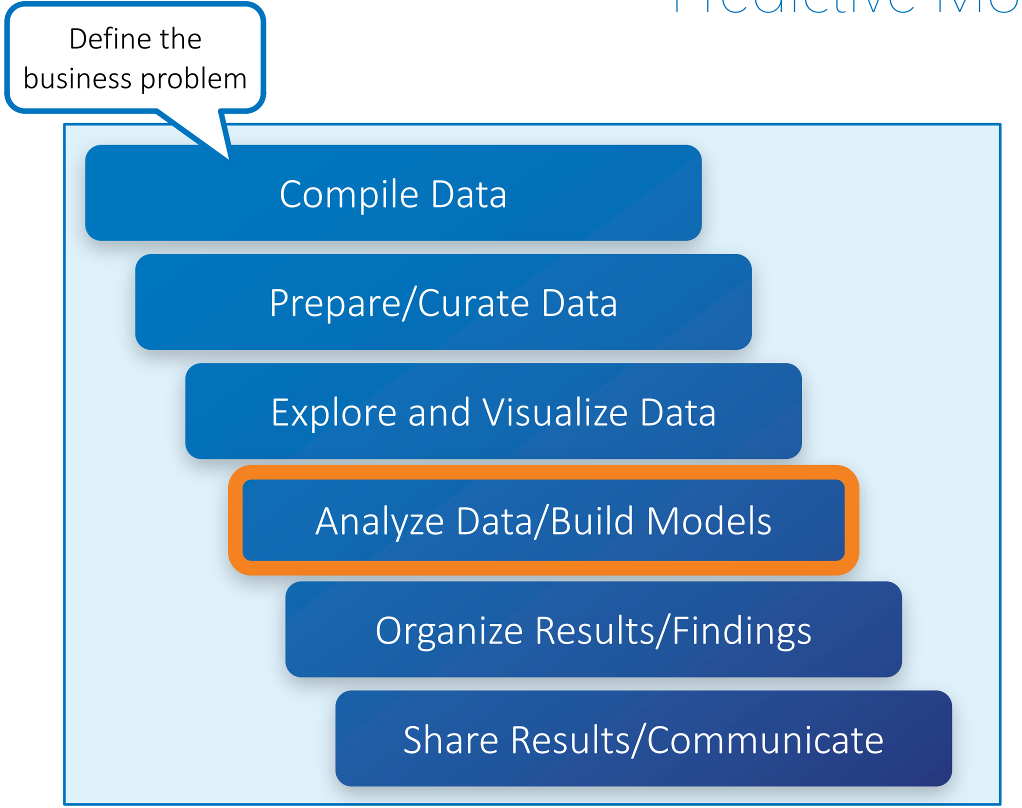
Why so many models?

No single type of model is always the best.

*Not an exhaustive list

Analytic Workflow

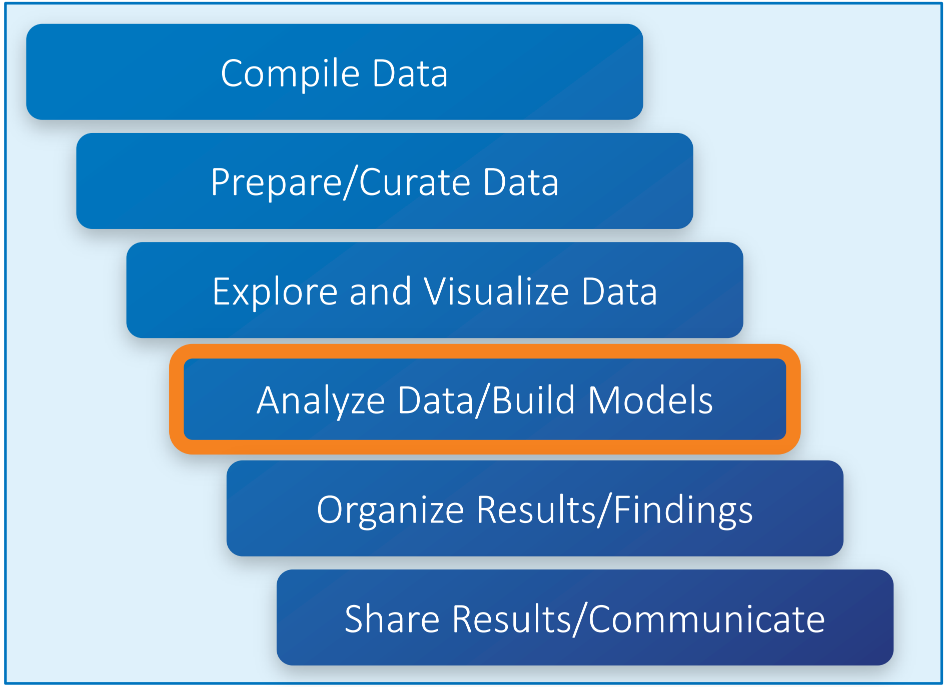
Predictive Modeling


Define the business problem
Compile Data
Prepare/Curate Data
Explore and Visualize Data
Analyze Data/Build Models
Organize Results/Findings
Share Results/Communicate

- Fit a model with validation
- Save prediction formula to data table (or publish to Formula Depot)
- Fit another model, repeat
- Use Model Comparison to evaluate the performance of each model on validation data
- Choose the best model (or set/combination of models)
- Deploy the model

Analytic Workflow

Predictive Modeling (with Model Screening)



Compile Data

Prepare/Curate Data

Explore and Visualize Data

Analyze Data/Build Models

Organize Results/Findings

Share Results/Communicate

- Fit the desired models in Model Screening
- Select the best model(s)
- Explore the model(s)
- Deploy the best model

Example 1: Diabetes

Scenario: Researchers want to predict rate of disease progression one year after baseline.

- Ten baseline variables, age, gender, body mass index, average blood pressure, and six blood serum measurements
- n = 442 diabetes patients.
- The response Y is a quantitative measure.
- The response Y Binary (High/Low)

Modeling goal: Predict patients most likely to have a high rate of disease progression, so corrective actions can be taken.

Efron, B., Hastie, T., Johnstone, J., and Tibshirani, R. (2004).

Comparing Predictive Models

How do you decide which model predicts the best?

- Compare measures of accuracy on the Validation set (or Test set).

For continuous responses:

- RMSE (or RASE) – lower is better
- AAE, MAD, MAE – lower is better
- RSquare – higher is better

For categorical responses:

- Misclassification (or error) and accuracy rates
- Precision (positive predictive value)
- AUC, Sensitivity (TP rate, recall), Specificity (TN rate)
- F1-Score and MCC – good with unbalanced data

Example 2: Credit Card Marketing

Scenario: Market research on acceptance of credit card offers.

Response:

- Offer Accepted (Only 5.5% of the offers are accepted)

Factors:

- Reward (Air Miles, Cash Back, Points)
- Mailer Type (Letter, Post Card)
- Financial information (Income Level, Credit Rating,…),

Modeling goal: Identify customers most likely to accept the credit card offer.

Summary

- Predictive Modeling: Accurately predict or classify future outcomes
- Fit and compare many models using validation

Model Screening streamlines the workflow:

- Fit models same time, one platform
- Select dominant model(s)
- Explore model details, and fit new
- Use Decision Threshold to explore cutoff for classification
- Deploy best model to the data table or Formula Depot

For More Information

Classification Metrics:

- https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Predictive Modeling, and Model Screening in JMP Pro:

- JMP User Community > Learn JMP
- STIPS, Module 7
- Model Screening Blog

JMP Early Adopter Program:

- See what's coming in JMP 17
- Provide early feedback

Thank you!

STATISTICAL DISCOVERY
FROM SAS

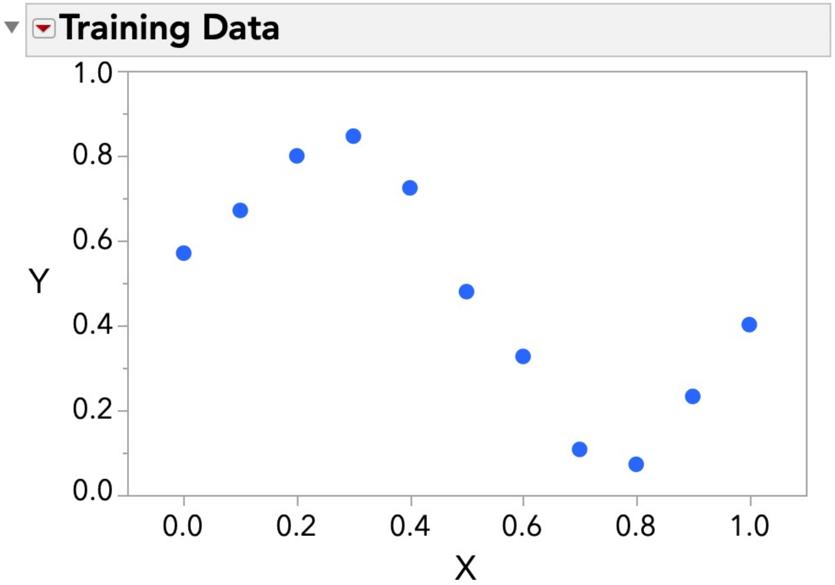
jmp.com

Why use Validation?

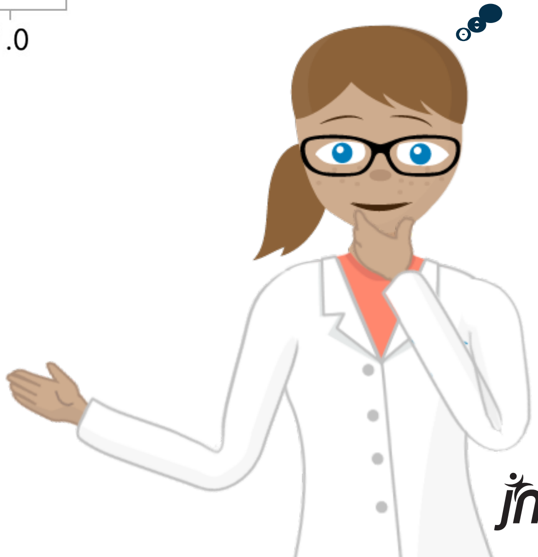
An illustration, borrowed from STIPS

(The Statistical Thinking for Industrial Problem Solving free online course)

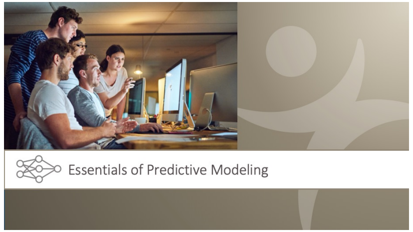
STATISTICAL DISCOVERY FROM SAS



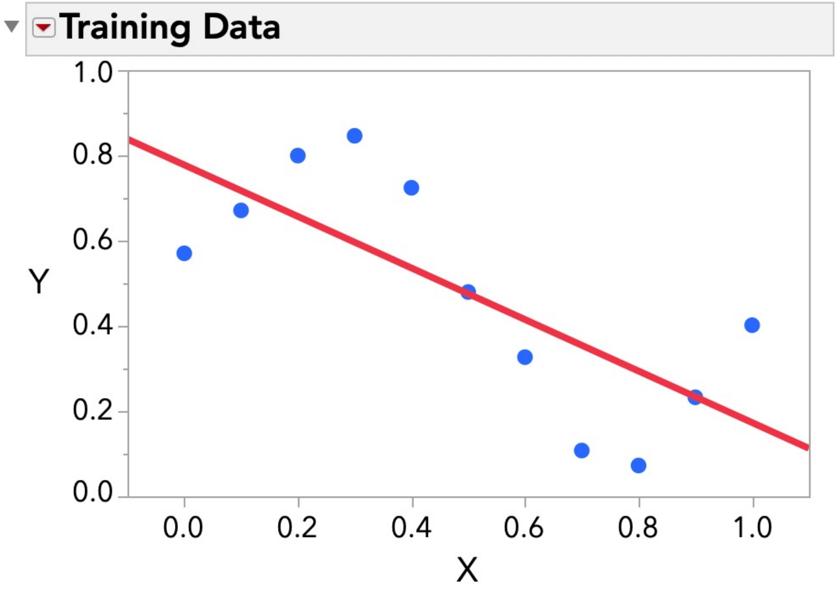
Can we predict Y from X?



Borrowed from STIPS, Module 7



Essentials of Predictive Modeling



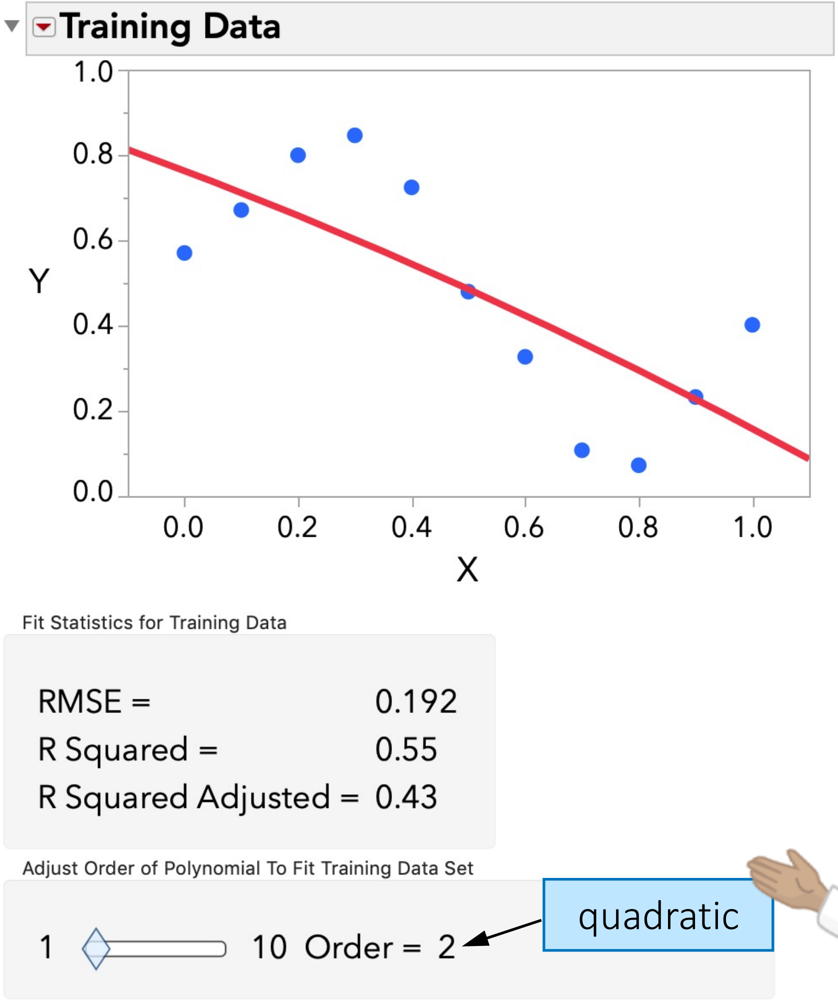
Fit Statistics for Training Data

RMSE =	0.183
R Squared =	0.55
R Squared Adjusted =	0.5

Adjust Order of Polynomial To Fit Training Data Set

1 ◇ 10 Order = 1 ← linear

Simple linear model doesn't make sense.



Fit Statistics for Training Data

RMSE = 0.192

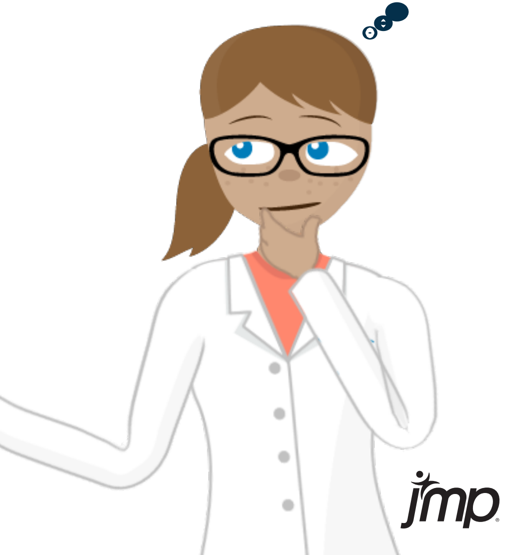
R Squared = 0.55

R Squared Adjusted = 0.43

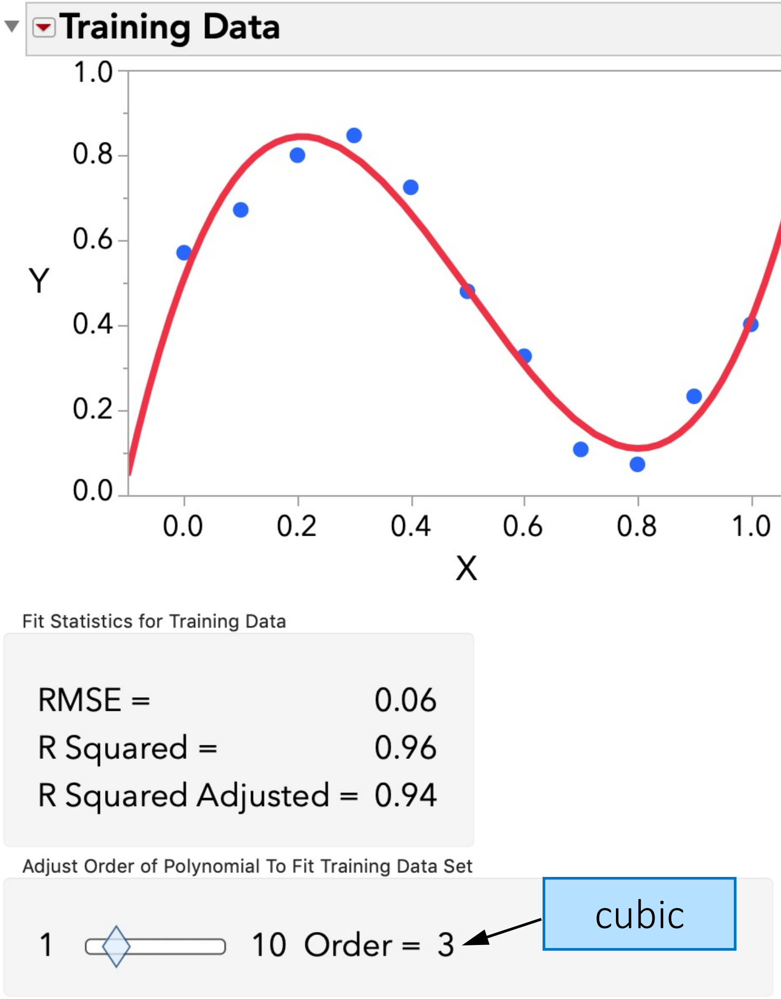
Adjust Order of Polynomial To Fit Training Data Set

1 10 Order = 2

quadratic



Quadratic model doesn't describe the relationship.

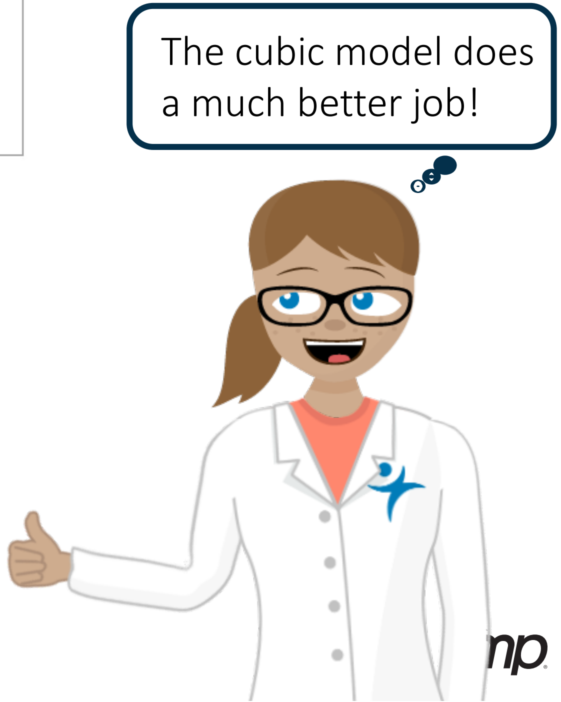


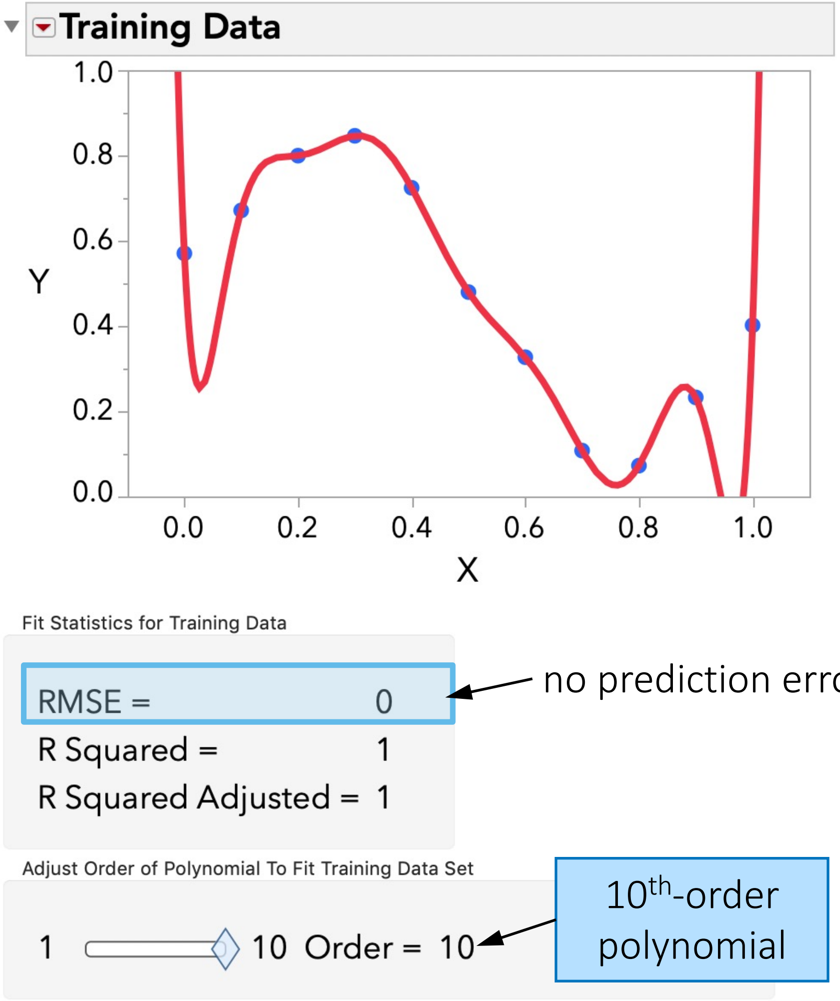
Fit Statistics for Training Data

RMSE =	0.06
R Squared =	0.96
R Squared Adjusted =	0.94

Adjust Order of Polynomial To Fit Training Data Set

1 10 Order = 3





Fit Statistics for Training Data

RMSE = 0

R Squared = 1

R Squared Adjusted = 1

Adjust Order of Polynomial To Fit Training Data Set

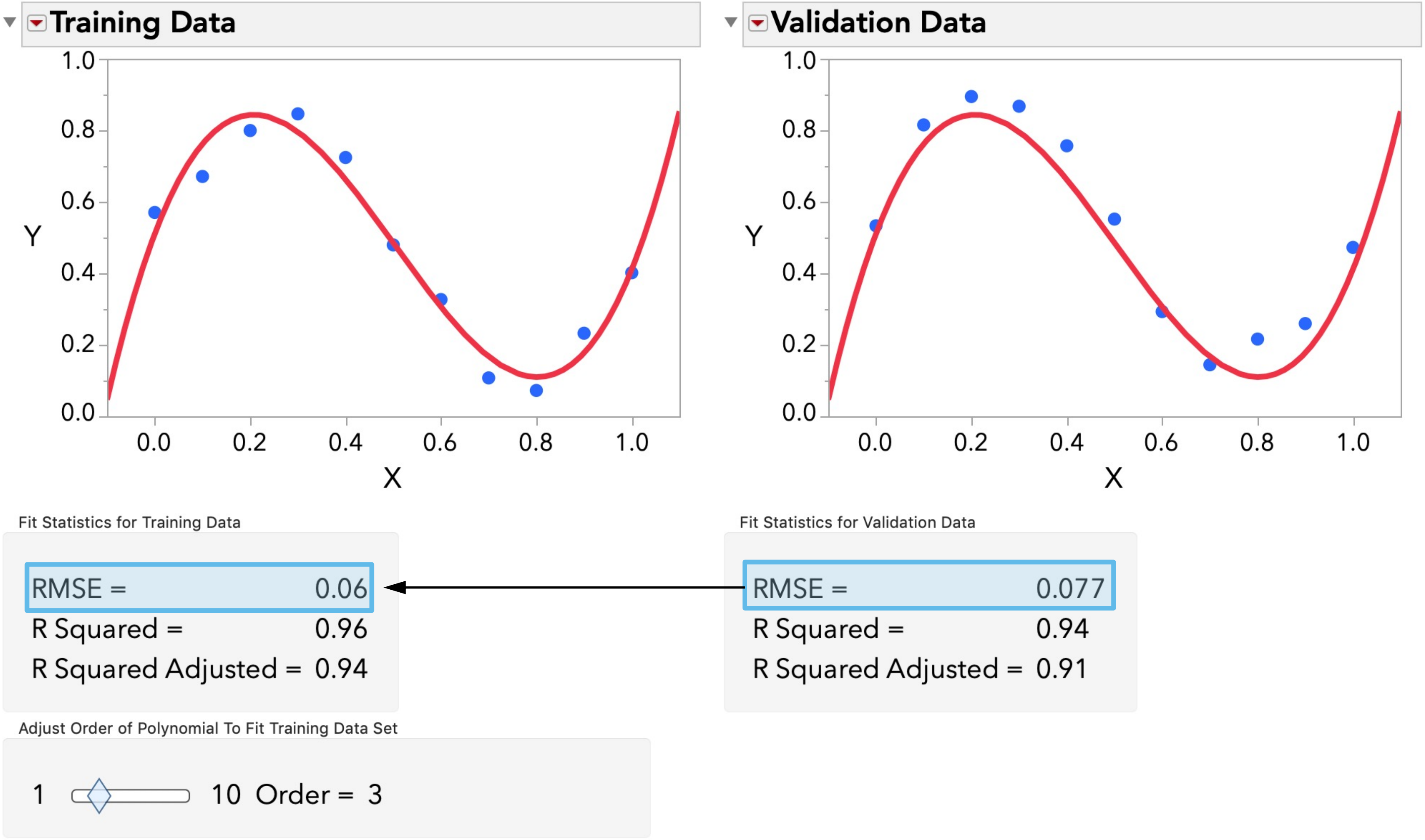
1 10 Order = 10

no prediction error

10th-order polynomial

If a cubic model is good, a higher order polynomial must be better!

But do I need a model with this much complexity?



Fit Statistics for Training Data

RMSE = 0.06

R Squared = 0.96

R Squared Adjusted = 0.94

Fit Statistics for Validation Data

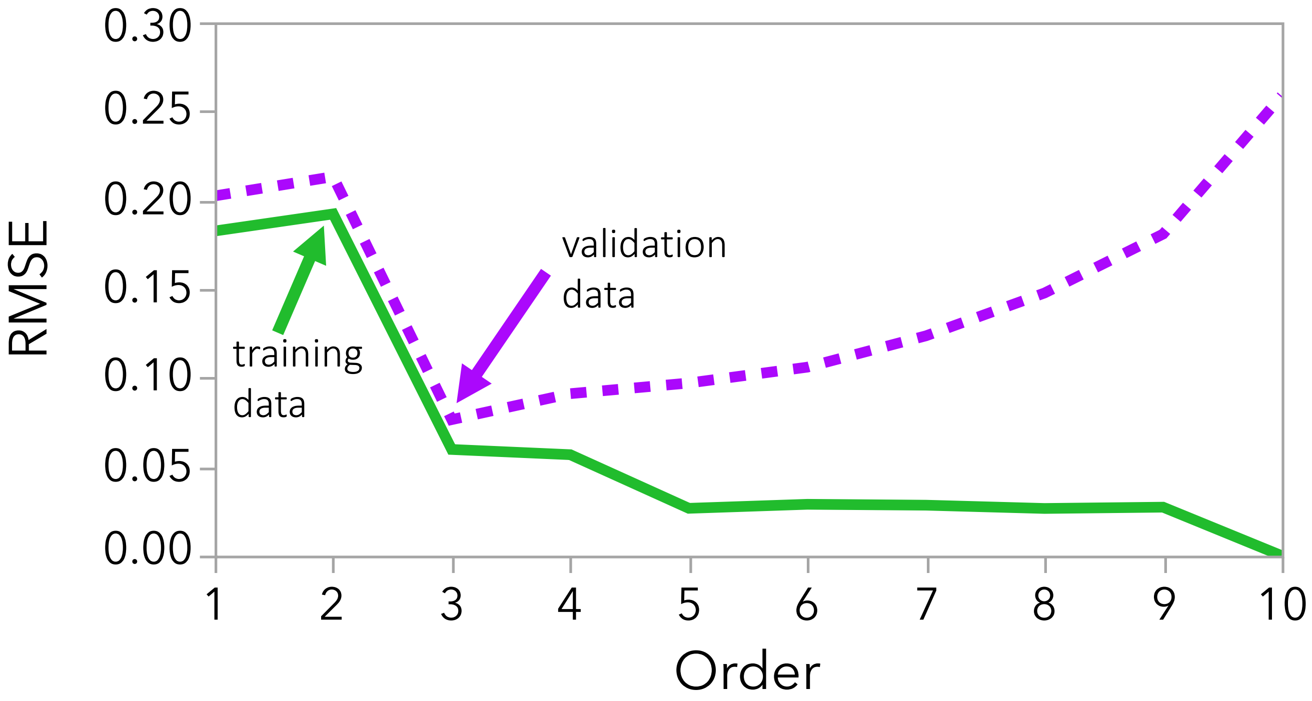
RMSE = 0.077

R Squared = 0.94

R Squared Adjusted = 0.91

Adjust Order of Polynomial To Fit Training Data Set

1 10 Order = 3


0.30
0.25
0.20
0.15
0.10
0.05
0.00
RMSE
1
2
3
4
5
6
7
8
9
10
Order
training data
validation data