MODELING FRAMEWORK

Applied Analytics: Frameworks and Methods 1

Outline

- Machine Learning
- Prediction vs. Inference
- Model Accuracy
- Overfitting
- Splitting the Data
- The Model
- Inferential Statistics

Machine Learning

- The area of analytics has benefited from developments in many disciplines and in many cases has adopted their language
- Computer scientists are accustomed to programming rules for machines
- Machine learning is a family of techniques where these rules are determined from data and then can be applied to previously unseen situations.
- It has been argued, *machine learning* is really about *learning from data*
  - *See an interesting illustration in this clip from the movie Groundhog Day.*

Machine Learning

Draws from many disciplines

- Math and Statistics
  - *Draw inferences*
  - *Estimate models*
- Computer science
  - *Algorithms for enabling analytical techniques*
  - *Efficient, scalable computing*
- Application Domain: Finance, Geography, Genomics, Marketing, Physics,..

Machine Learning

- Predictors (also known as Inputs, Features, or Independent Variables)
  - *Denoted as X*
- Outcome (also known as Output, Response, or Dependent Variable)
  - *Denoted as Y*

- $Y = f(X) + \varepsilon$
- Machine Learning is a set of approaches for using data to determine the functional relationship (f) between predictor(s) (X) and outcome (Y)

Machine Learning

- Supervised Learning
  - *We have data on both predictors and outcome.*
  - *Also, known as labeled data*
  - *E.g., regression, trees*
- Unsupervised Learning
  - *There is data on a set of variables but no associated outcome or response variable.*
  - *E.g., cluster analysis, factor analysis, market basket analysis*
- This course will review only Supervised Learning methods

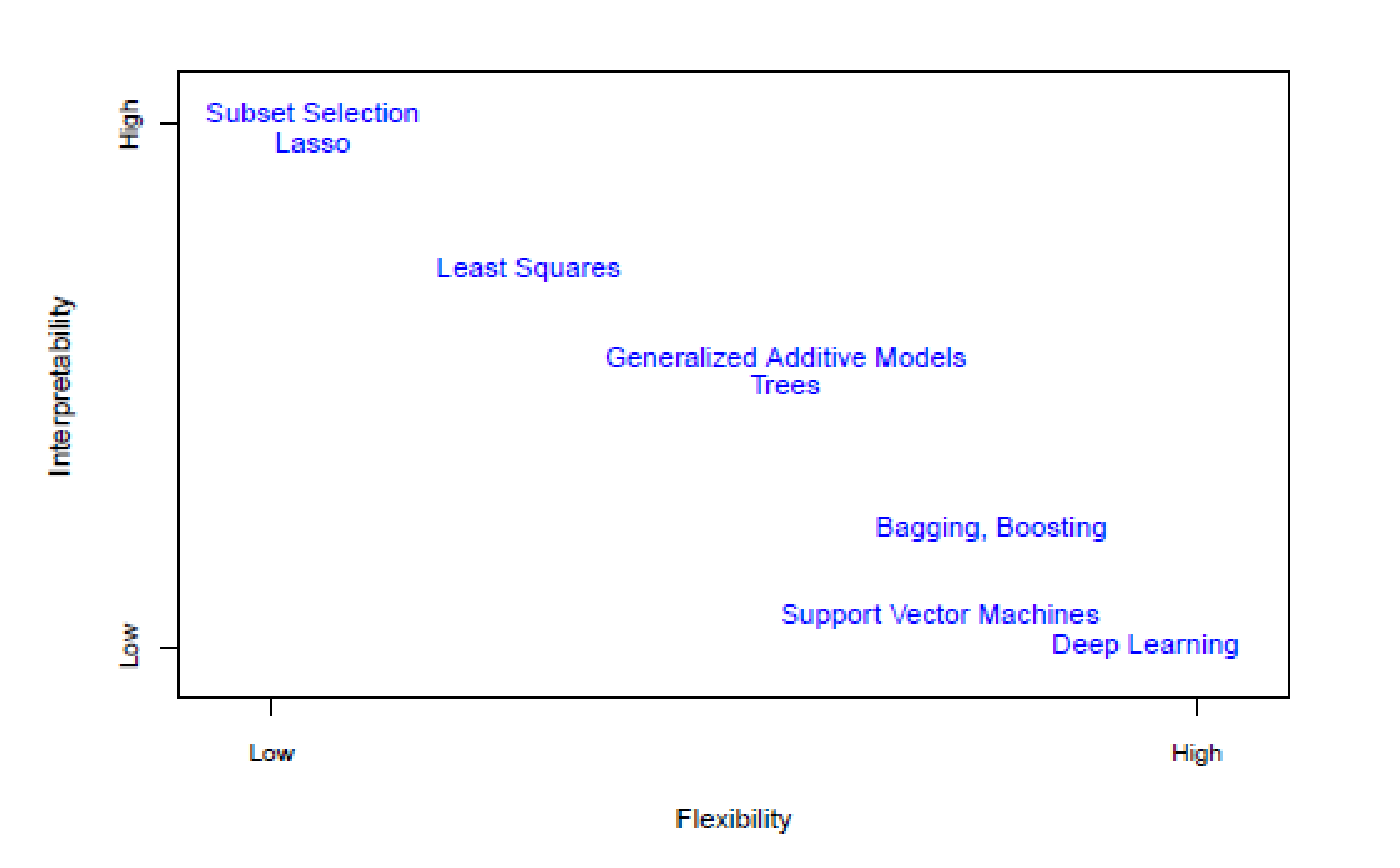
Supervised Learning

- Consider the model
  - *House_Sale_Price = f(Area, Age, Number_of_Bathrooms, Month_of_Listing)*
- Prediction
  - *Goal is to generate accurate predictions of House_Sale_Price*
  - *Prediction Error (ε) = Reducible Error + Irreducible error (Var(ε))*
  - *Techniques discussed in this class aim at estimating f with the aim of minimizing the reducible error*
- Inference
  - *Determine predictors associated with House_Sale_Price*
  - *Determine nature of relationship (e.g., valence, i.e., positive or negative; functional form such as linear or non-linear)*

Prediction vs. Inference

- Many problems are predominantly interested in only one of the two goals.
  - *New product development (Inference): Which product features influence sales and by how much?*
  - *Customer Targeting (Prediction): Using demographics and online behavior, predict which customers will click on the link in an email?*
  - *Of course, there are a few situations where both are of interest*
- Techniques that favor one don’t do so well at the other
  - *Models with the lowest prediction errors are generally hard to interpret*
  - *Flexible models are generally better for predictions while restrictive methods are better for explaining phenomena*

Prediction vs. Inference



Source: Introduction to Statistical Learning by James et al (2021)

Estimation Approaches

- Parametric methods
  - *Make an assumption of the functional form of relationship between predictors and outcome*
  - *Use training data to estimate parameters of equation*
  - *E.g., Linear regression*
- Non-parametric methods
  - *Does not make any assumption about the functional form of relationship*
  - *Can fit a wider range of shapes for f*
  - *But, needs a very large number of observations*
  - *E.g., splines*

Regression vs. Classification Problems

- Depends on nature of the outcome variable

- Regression problem: Outcome variable is numeric
  - *Least squares linear regression*
- Classification problem: Outcome variable is categorical
  - *Logistic regression*
- While some techniques can address only one i.e., regression or classification problems, others can address either. The latter include trees, forests, and boosting.

Regression vs. Classification Problems
Model performance metrics

Regression Problems

- Estimate Predictions

Classification Problems

- Decision Predictions
  - *Group 1 or group 2; High or Low*
  - *Often involves categorizing a probability outcome into class predictions*

Regression vs. Classification Problems

Regression Problems

Predictor1	Predictor2	Predictor3

Outcome
232.32
134.54
67.45
129.46
162.89

Classification Problems

Predictor1	Predictor2	Predictor3

Outcome
Not Buy
Buy
Buy
Buy
Not Buy

Regression Problems
Model Performance Metrics

- Measures of error
  - *Mean Squared Error (mse)*
  - *Root Mean Squared Error (rmse)*
  - *Mean Absolute Error (mae)*
  - *Mean Absolute Percentage Error (mape)*
- Measure of explained variance
  - R^2

Predictor1	Predictor2	Predictor3

Outcome
232.32
134.54
67.45
129.46
162.89

Classification Problems
Model Performance Metrics

- Class-probability based metrics
  - *Log-likelihood*
  - *Gini*
  - *Entropy*
- Accuracy-based metrics
  - *Accuracy,*
  - *Misclassification rate (= 1-accuracy),*
  - *Cohen's Kappa (adjusts for class imbalance)*
- Accuracy for specific classes
  - *Sensitivity and Specificity (distinguishes types of error for binary outcomes)*
- Area under the ROC curve (AUC)

Predictor1	Predictor2	Predictor3

Outcome
Not Buy
Buy
Buy
Buy
Not Buy

Model Accuracy

- Performance of a model is determined by comparing model predictions to true values.
- Performance can only be judged based on the data the researcher has, i.e., the data used to train the model.
- But, in most cases the researcher is interested in performance of the model in the real world, i.e., on data not used to train the model.

Model Accuracy: Simple vs. Complex

- As model complexity increases,
  - *models perform better on the sample used to train the model*
  - *but they also perform worse on datasets not used to train the model*
- The extent to which the model performs well on the data used to build it versus data not used to build it is called *Overfitting*.
- Overfitting is seen when in-sample performance far exceeds out-of-sample performance.
- This is the classic Bias-Variance tradeoff
- Let us review this issue.

OVERFITTING: BIAS VS. VARIANCE

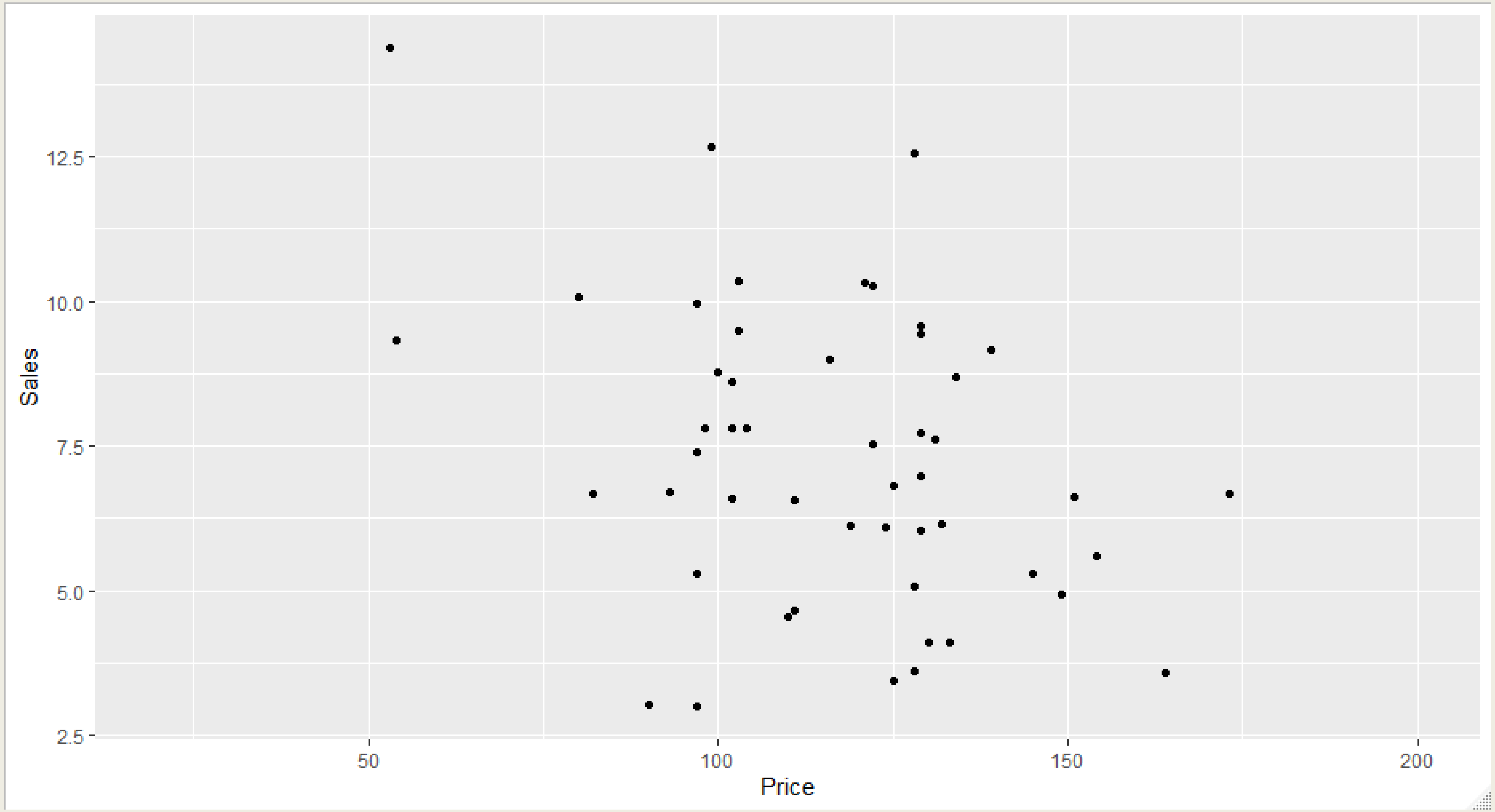
Illustrations from Life

- Car performance tuned on test tracks often falters on real roads.
- A student who practices hard for a standardized test sees his scores improving rapidly. The actual exam is a bit of a shocker as his score is significantly lower.

Carseats

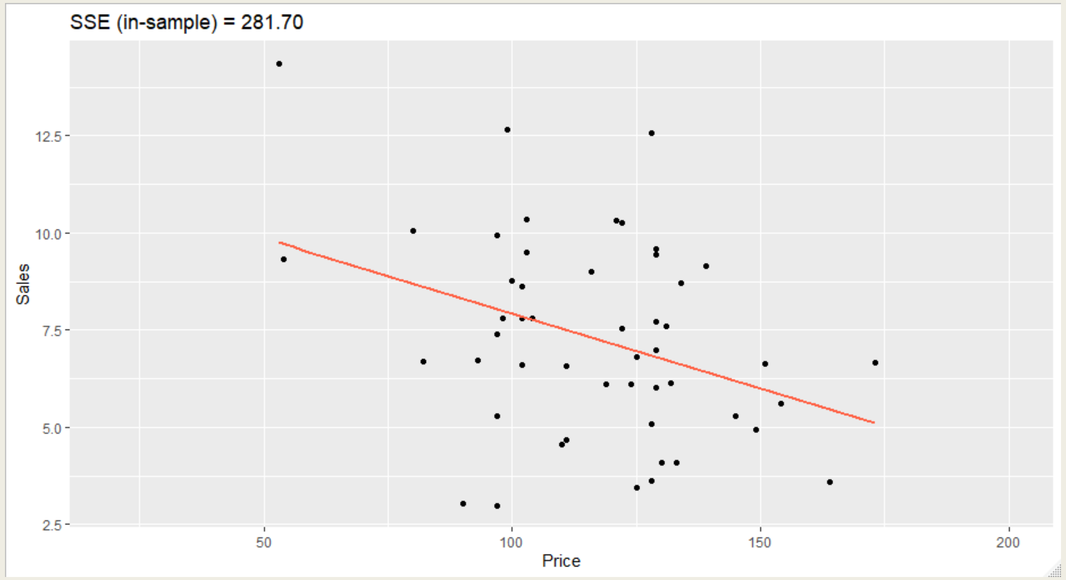
- Next few slides pictorially represent the Bias Variance tradeoff with data on Carseats.
- Sample (n=50) used to estimate model was randomly selected from Carseats data.
- The model was then evaluated on three other samples (n=50) drawn randomly from the Carseats data.

Predicting Carseat Sales with Price

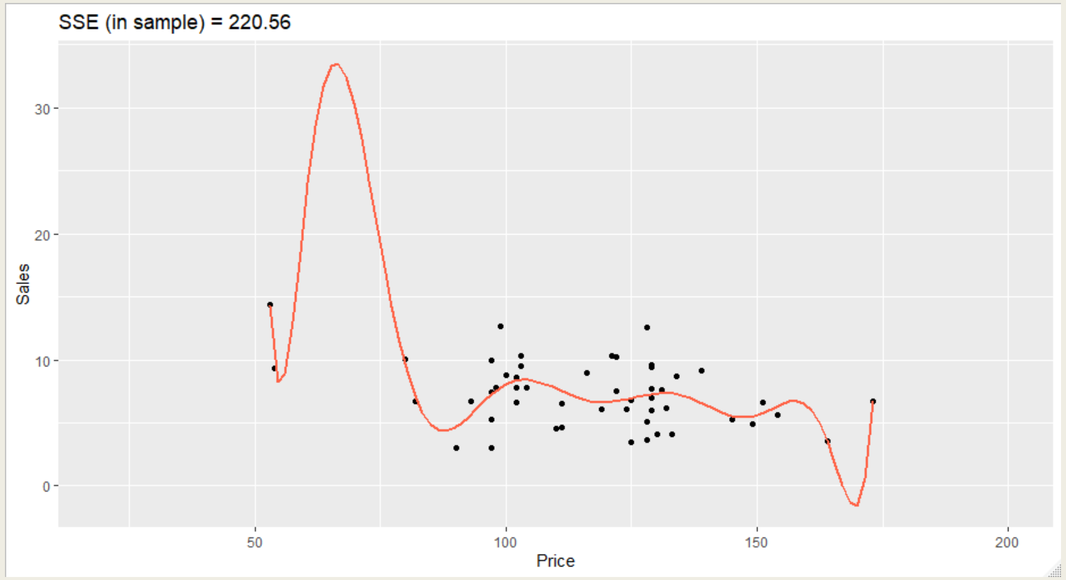


Which model looks better?

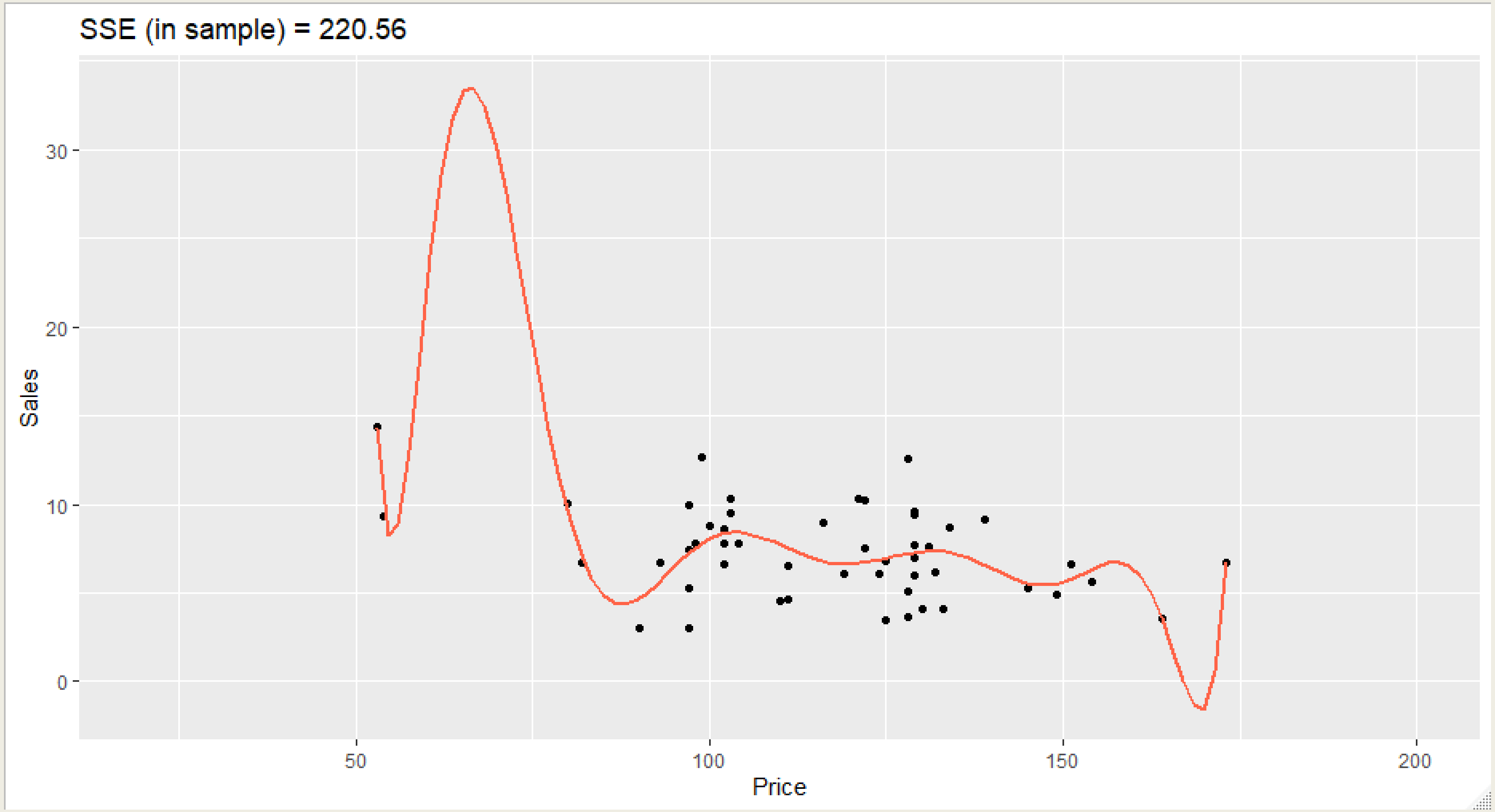
Simple



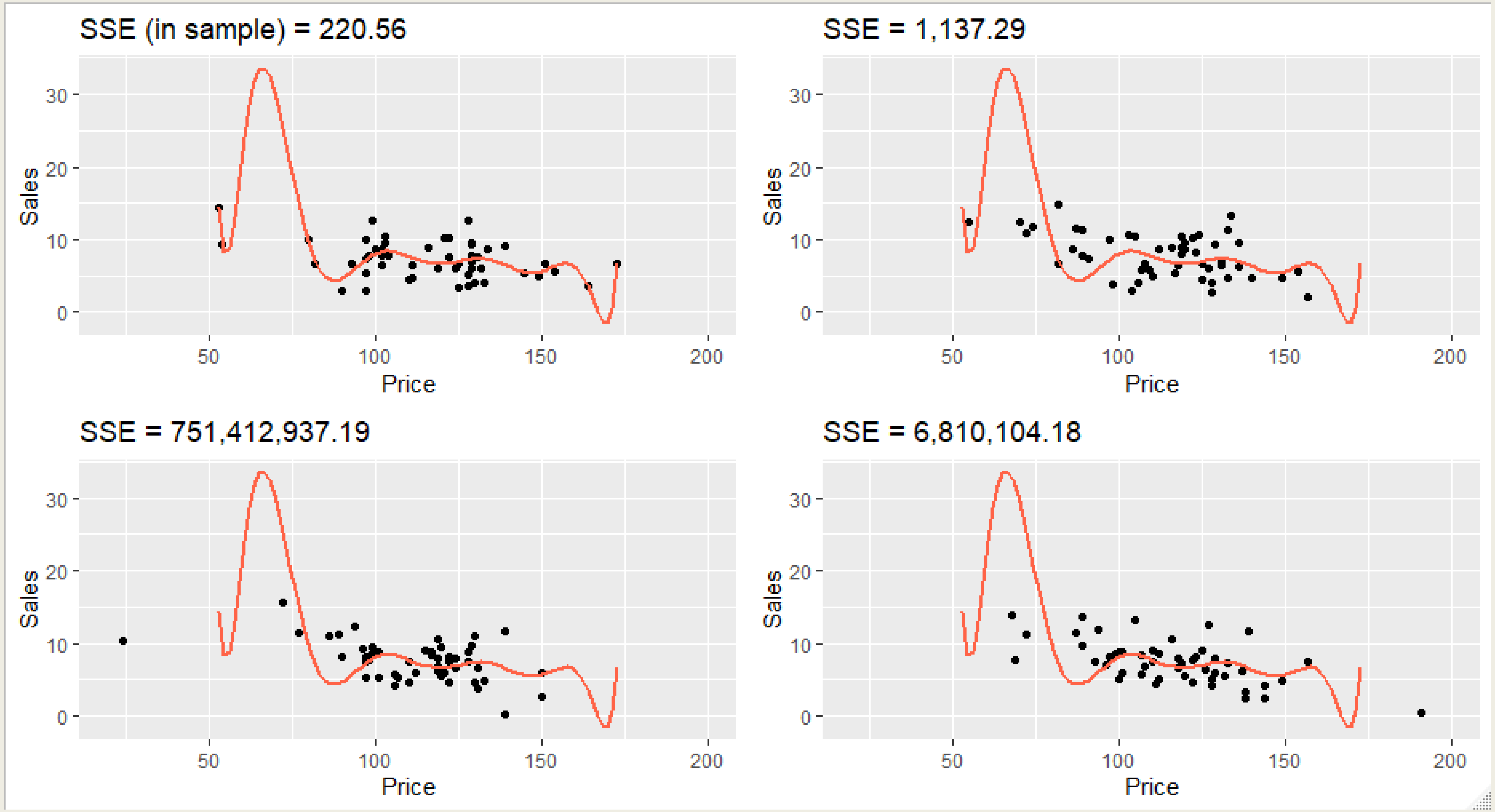
Complex



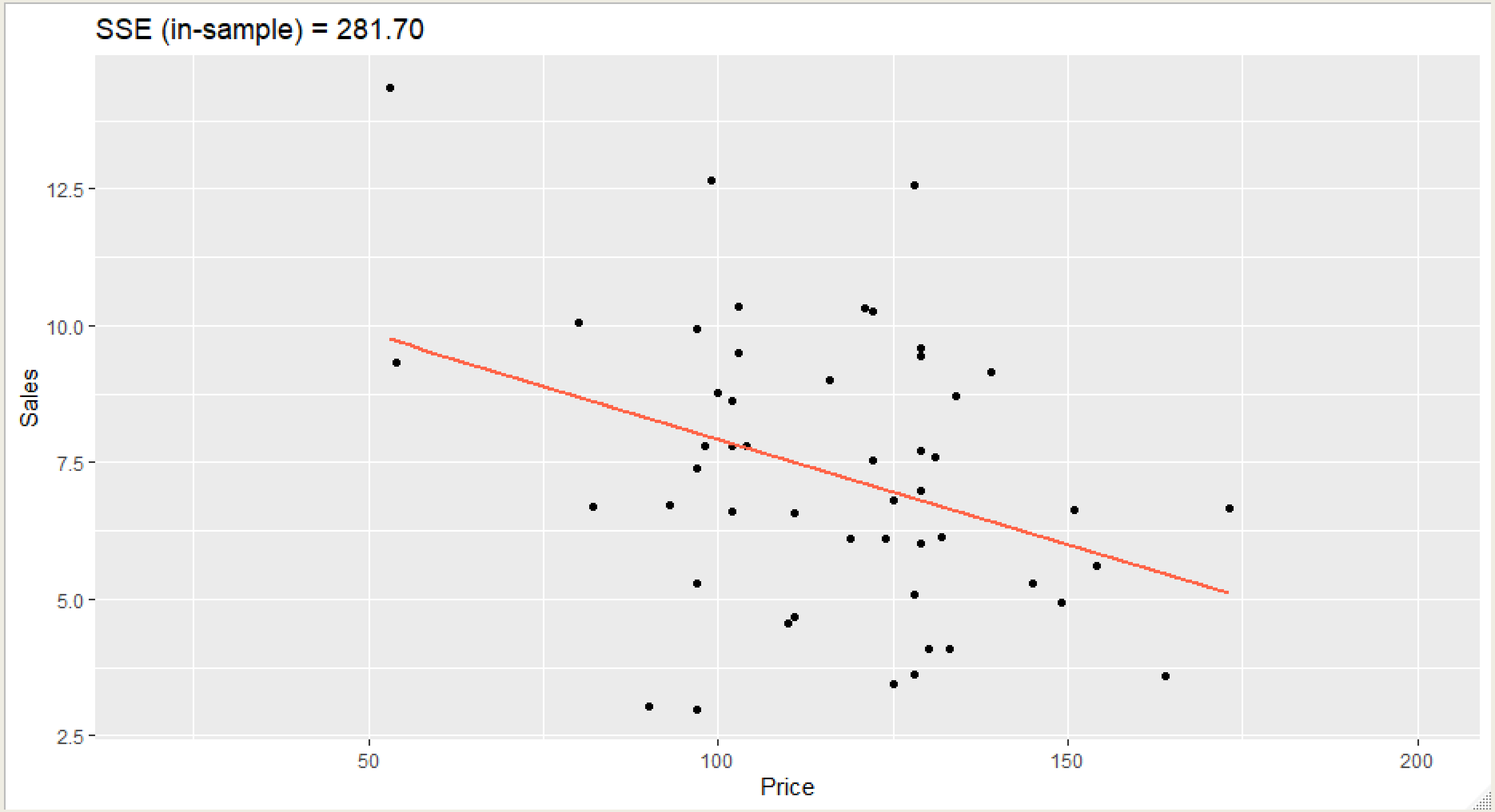
Complex Model (in-sample)



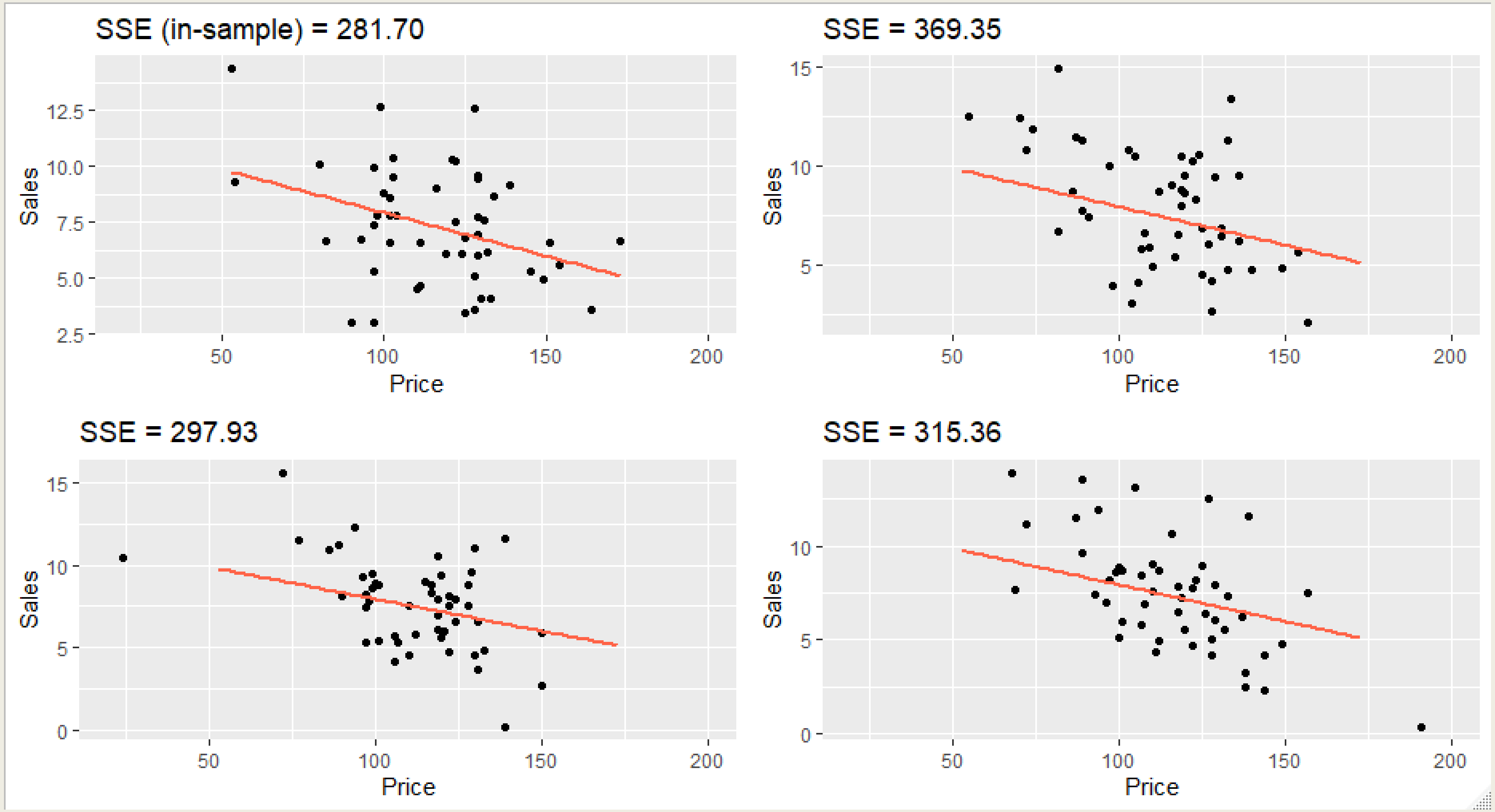
Complex Model (out of sample)



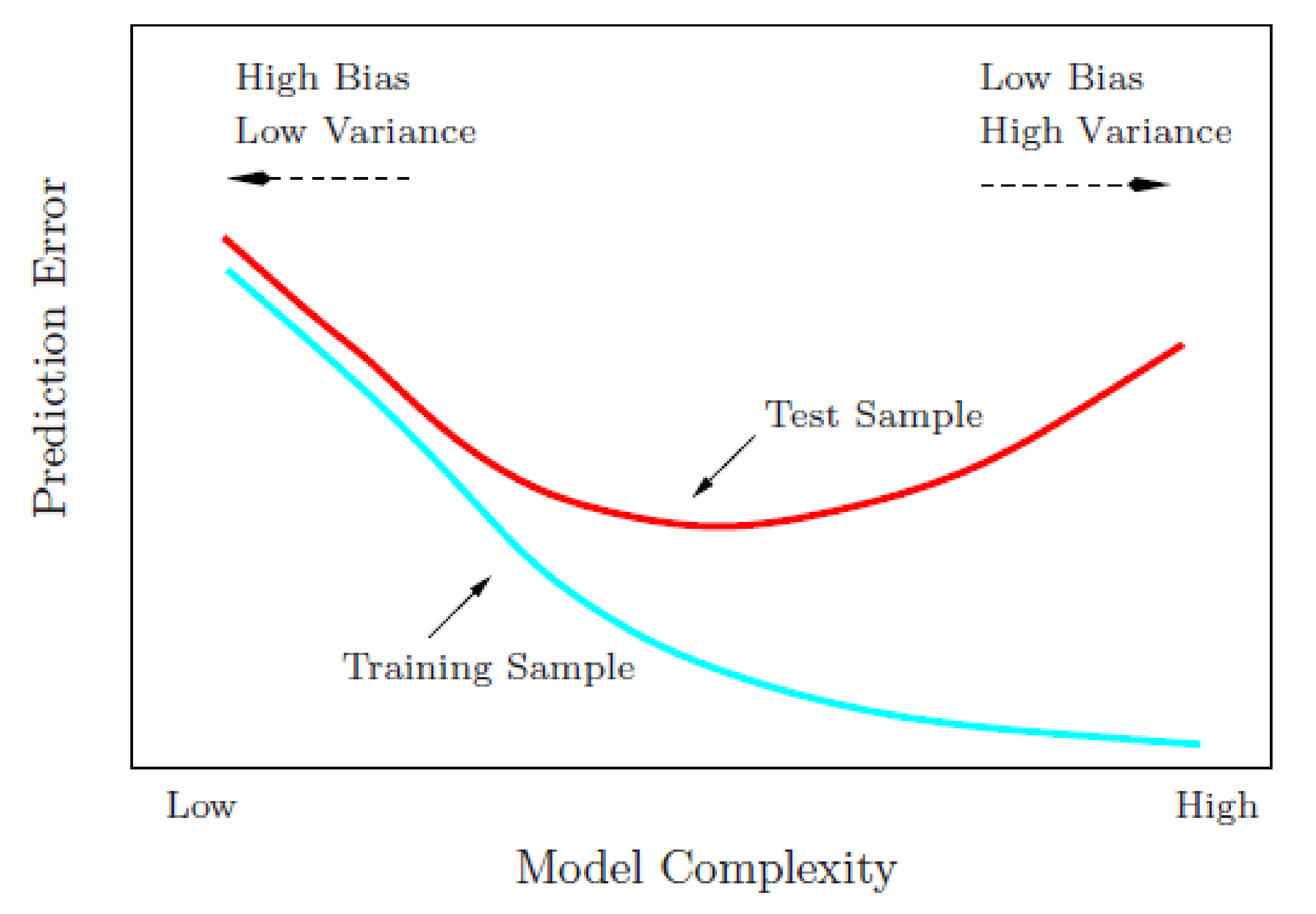
Simple Model (in-sample)



Simple Model (out of sample)

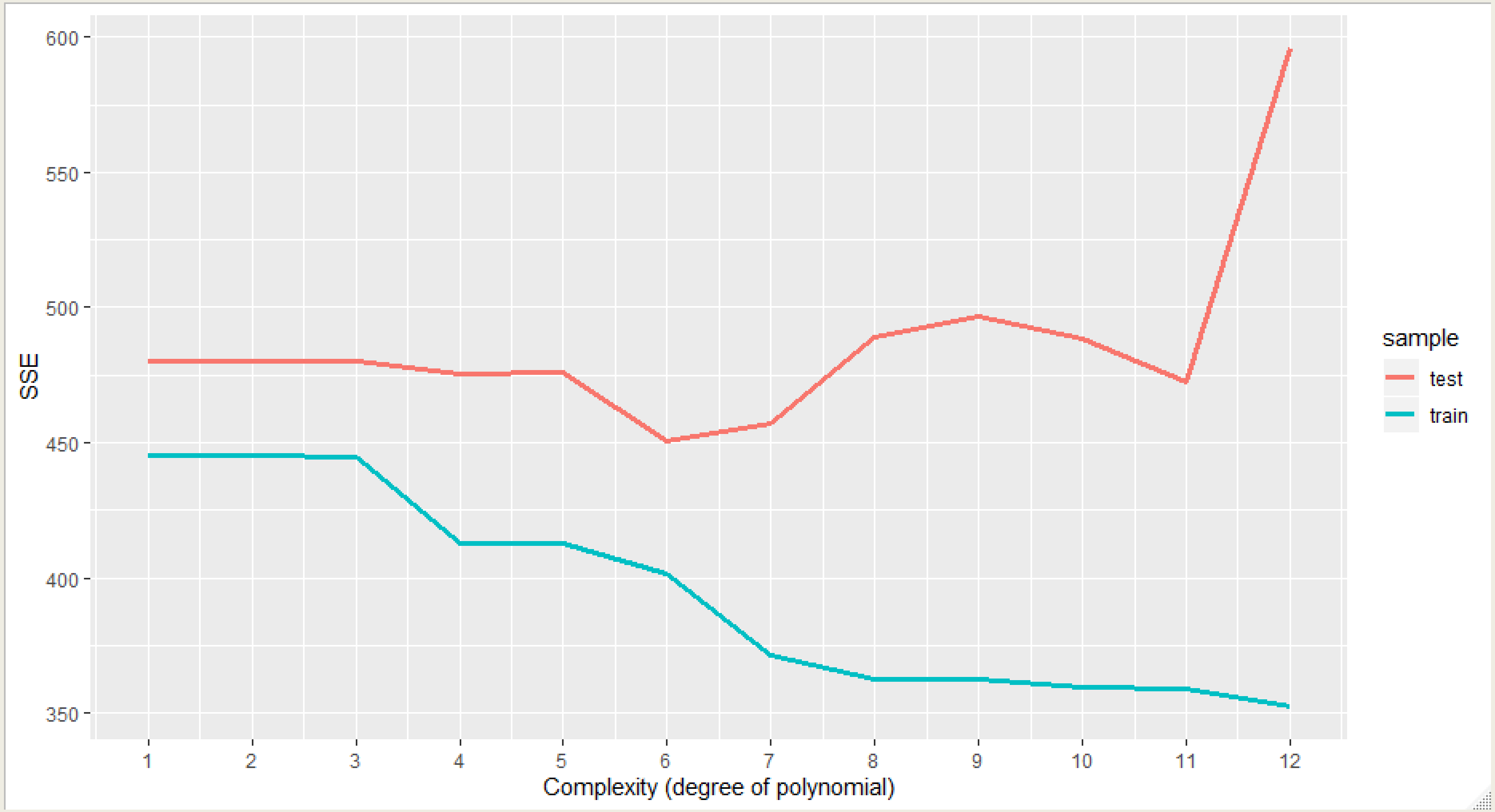


Training vs. Test Set



Training vs. Test Set

Prediction Accuracy vs Complexity: Sales = f(Price^d), where d is degree

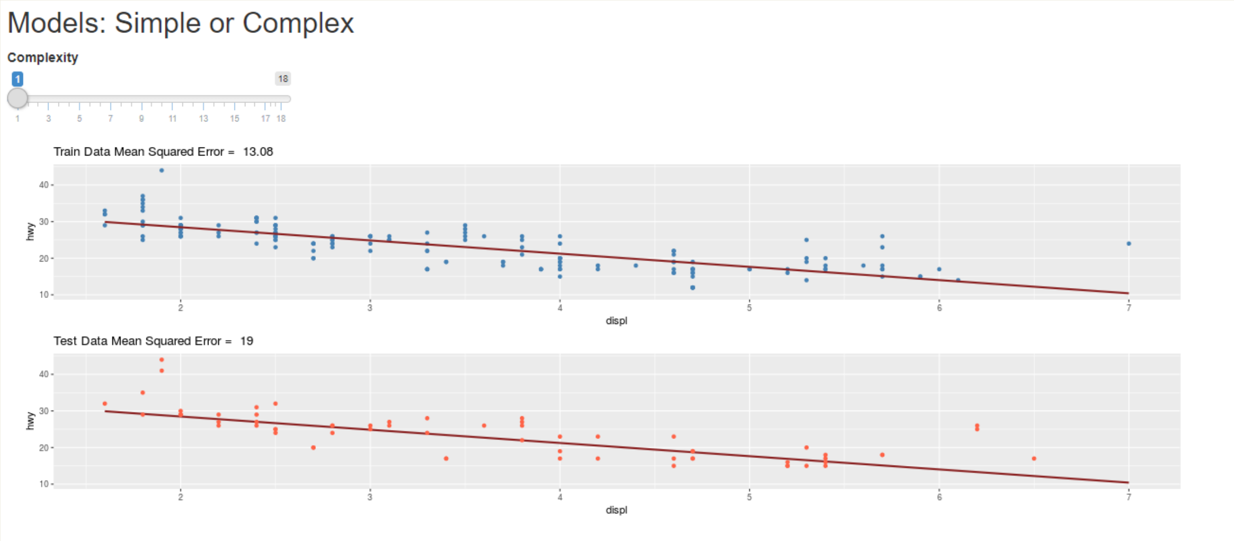


Training vs. Test Set

- Researcher is generally interested in developing a model that performs well out-of-sample.
- In practice, we only have training data, therefore not possible to assess performance out-of-sample.
- Also, as noted in foregoing illustration, in-sample performance is a poor proxy for out-of-sample performance.

Training vs. Test Set

- Here is an interactive chart to examine the effects of complexity on train and test set performance.
- Complexity is reflected by the degree of a polynomial regression model
- Model uses mpg data (from library(ggplot2) to predict hwy gas mileage using displ for different degrees of displ.



Train and Test Samples

- One solution is to split the sample into two parts: train and test.
  - *Other solutions such as cross-validation will be discussed later.*
- Estimate the model on train set and evaluate using the test set.
- Performance of model on test set can be used as an indication of out-of-sample performance.
- Note:
  - *train sample is also referred to as estimation sample*
  - *test sample is also known as validation or holdout sample*

Train and Test Samples
Factors to Consider

- Size of train and test sample
  - *If data is sufficiently large, a 50:50 split may be done*
  - *Generally, train sample is larger than test sample, with the split being 60:40 or 70:30. These are heuristics not rules.*
- Method of split
  - *Non-random approaches: Only used in very specific situations. E.g. time-series data.*
  - *Random approaches*
    - Simple random sampling: Designed to make train and test sample as similar as possible. In R, sample()
    - Stratified sampling: Applies random sampling within subgroups. In R, caTools::sample.split(), caret::createDataPartition()
      - *On outcome: Random sampling while ensuring the distribution or proportion of outcome is the same across samples*
      - *On predictors: Same idea as above but for specific predictors such as gender or location*

THE MODEL

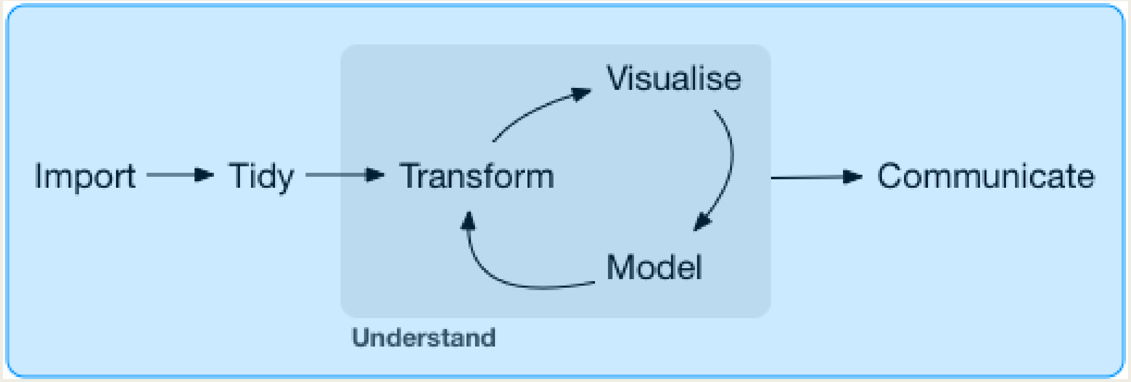
The “Best” Model

- *The No Free Lunch Theorem shows that under certain assumptions*
  - *No single predictive model can be declared to be the best*
- While certain models work with certain data characteristics (e.g., missing values), they may fail with different data characteristics
- Rather than seeking a silver bullet, analysts, should examine the problem or data at hand, before deciding on the models to use.

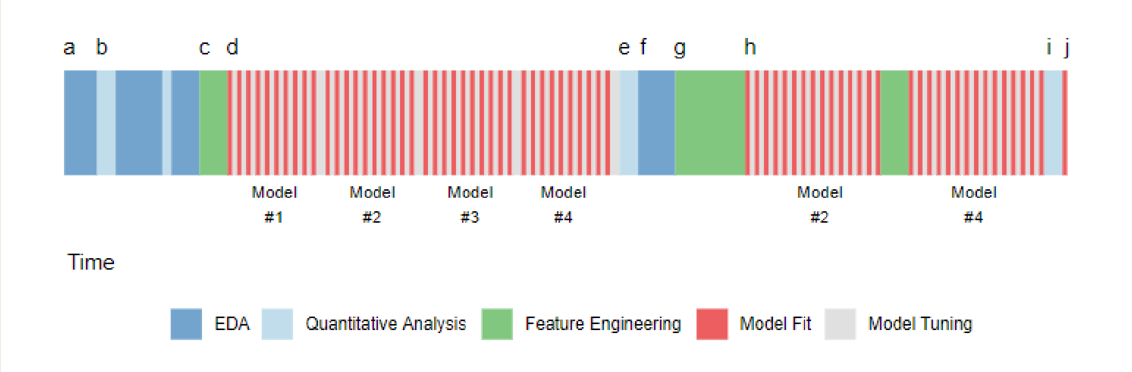
Road to the Best Model

- Modeling process is iterative, not linear

- Predictive analysis is much more than just fitting a single model to tidy data



Source: R for Data Science



Source: Kuhn and Johnson (2019)

INFERENTIAL STATISTICS

Inferential Statistics

- Population
  - *Collection of all units for the study*
- Sample
  - *Subset of the population*
- Sample is used to draw inferences about the population
- Most studies are based on a sample

Process of inferential statistics

- Generate a hypothesis about the population, null hypothesis (H_0) and an alternative hypothesis (H_1) such that the two cover the Universe of possibilities
- Select a statistical technique to generate a test statistic. Test statistic often follows a well known distribution such as t, F, or χ^2.
- Choose a level of significance (e.g., $\alpha = 0.01$) to reflect tolerance for Type I error, i.e., rejecting H_0 when in fact it is true.
- Gather data and calculate value of test statistic
- Determine the probability (p-value) of obtaining the test statistic assuming null hypothesis is true.
- If $p<\alpha$, reject H_0

Illustration

Consider the Linear Model: Sales = b_0 + b_1*AdSpend

- Hypotheses, being tested (although not always explicitly stated)
  - *H_0: $b_1 = 0$*
  - *H_1: $b_1 \neq 0$*
  - *If coefficient of AdSpend (b_1) in the population is 0, one would conclude AdSpend does not drive Sales*
- Test statistic: t value for coefficient of AdSpend
- Level of Significance (α) = 0.01
  - *Values used tend to be 0.1, 0.05, 0.01, 0.001 but whatever the threshold, it should be set before looking at the data*
- Gather data and calculate value of test statistic
- Translate t value into p-value. Let's say p = 0.002. This means if b_1 is 0 then there is only a 0.2% chance of obtaining the sample data.
- Since the chance (p=0.002) is below our threshold (α = 0.01), one would reject the null hypothesis and conclude that the coefficient of AdSpend is not zero. In other words, AdSpend influences Sales.

In Practice

- Desirable results are generally in H_1, so analysts generally seek to reject H_0 in favor of H_1.
- p-value does not reflect strength of effect
- p-value is sensitive to sample size. With large samples, even very small effects are statistically significant
- Statistical significance does not imply practical significance.
- On the other hand, before one can examine practical significance, it is imperative that the results are statistically significant.

Conclusion

- In this module, we reviewed
  - *machine Learning*
  - *goals of prediction vs. inference*
  - *assessing model accuracy*
  - *problem of overfitting*
  - *splitting the data to estimate test error*
  - *the iterative modeling process*
  - *inferential statistics to determine significance of results*