Predictive modelling process

A first tour

Michela Mulas

Today's goal

Today, we going to ...

- Introduce the predictive modeling process.
- Case study: Predict fuel economy.
- Detail the homework assignment.

Reading list

Max Kuhn and Kjell Johnson. *Applied Predictive Modeling*, Springer (2014)

Today's lecture is mainly based on Chapters 1 and 2 of the book.

Motivations

Information has become more readily available via the internet and media and our desire to use this information to help us make decisions has intensified.

Human brain can consciously and subconsciously assemble a vast amount of data but it cannot process the even greater amount of easily obtainable, relevant information for the problem at hand.

- Websites are used to filter billions of web pages to find the most appropriate information for our queries.
- These sites use tools that take our current information, sift through data looking for patterns that are relevant to our problem, and return answers.
- The process of developing these kinds of tools has evolved throughout a number of fields such as chemistry, computer science, physics, and statistics and has been called **machine learning**, **artificial intelligence**, **pattern recognition**, **data mining**, **predictive analytics**, and **knowledge discovery**.

Motivations

Each field approaches the problem using different perspectives and tool sets, the **ultimate objective is the same: to make an accurate prediction**.

Geisser[1] defines predictive modeling as the process by which a model is created or chosen to try to best predict the probability of an outcome.

Predictive modeling

The process of developing a mathematical tool or model that generates an accurate prediction.

Examples of artificial intelligence can be found everywhere[2]:

- The Google global machine uses AI to interpret cryptic human queries.
- Credit card companies use it to track fraud.
- Netflix uses it to recommend movies to subscribers.
- Financial systems use AI to handle billions of trades.

[1] Geisser S. (1993). *Predictive Inference: An Introduction*. Chapman and Hall
[2] Levy S. (2010). *The AI Revolution is On*. Wired.

Terminology

Predictive modeling is one of the many names that refers to the **process of uncovering relationships within data for predicting some desired outcome**.

Since many scientific domains have contributed to this field. So, there are synonyms:

- **Sample**, **data point**, **observation**, or **instance**: refer to a single, independent unit of data, such as a customer, patient, or compound.
  - **Sample** can also refer to a subset of data points (e.g. training set sample).
- **Training set** consists of the data used to develop models while the **test or validation sets** are used solely for evaluating the performance of a final set of candidate models.
- **Predictors**, **independent variables**, **attributes**, or **descriptors** are the data used as input for the prediction equation.
- **Outcome**, **dependent variable**, **target**, **class**, or **response** refer to the outcome event or quantity that is being predicted.

Terminology

Predictive modeling is one of the many names that refers to the **process of uncovering relationships within data for predicting some desired outcome**.

Since many scientific domains have contributed to this field. So, there are synonyms:

- **Continuous data** have natural, numeric scales. Blood pressure, the cost of an item, or the number of bathrooms are all continuous. In the last case, the counts cannot be a fractional number, but is still treated as continuous data.
- **Categorical data**, also known as **nominal**, **attribute**, or **discrete data**, take on specific values that have no scale. Credit status (“good” or “bad”) or color (“red”, “blue”,etc.) are examples of these data.
- **Model building**, **model training**, and **parameter estimation** all refer to the process of using data to determine values of model equations.

Limitations

After the breakdown of financial markets at 2008, Rodriguez[1] wrote:

“Predictive modeling, the process by which a model is created or chosen to try to best predict the probability of an outcome, has lost credibility as a forecasting tool”.

This is due to either the modeller’s expertise or knowledge, or the lack of resources.

The main common reasons that make a predictive model to fail are as below:

- Inadequate pre-processing of the data;
- Inadequate model validation;
- Unjustified extrapolation;
- Over-fitting the model to available data.

[1] Rodriguez M (2011). *The Failure of Predictive Modeling and Why We Follow the Herd.* Technical report, Concepcion, Martinez & Bellido.

Key ingredients

For an **effective predictive model** we need:

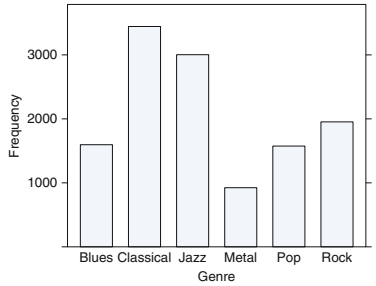
- Intuition and deep knowledge of the problem context:
  - Vital for driving decisions about model development.
- Relevant data:
  - The whole process begins with data.
- Versatile computational toolbox:
  - Including techniques for data pre-processing and visualization;
  - Suite of modeling tools for handling a range of possible scenarios.

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Music genre

- This data set was published as a context data set: `http://tunedit.org/challenge/music-retrieval/genres`



- 12495 music samples with 191 characteristics.
- Response categories are not balanced.
- All predictors are continuous and many are highly correlated and
- The predictors spanned different scales of measurement.

- **Objective**: Develop a predictive model for classifying music into six categories.

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Grant applications

- This data set was published as a context data set: `http://www.kaggle.com`

  Historical database of 8707 University of Melbourne grant applications from 2009 and 2010 with 249 predictors.

  Grant status: “unsuccessful” or “successful” (with 46% successful).

  Australian grant success rates are less than 25%: the dataset is not representative of the Australian rates.

  Predictors include Sponsor ID, Grant Category, Grant Value Range, Research Field and Department.

  Data are continuous, count and categorical.

  Many predictor missing values (83%).

  The samples are not independent: the same grant writers occurred multiple times.

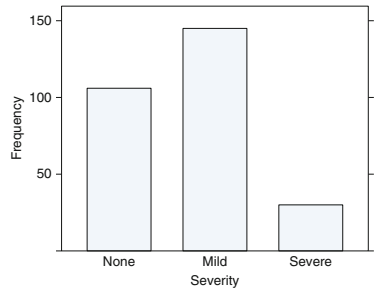
- **Objective**: Develop a predictive model for the probability of successful application.

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Hepatic injury

- Data set from the pharmaceutical industry.



- 281 unique compounds, 376 predictors measured or computed for each.
- Categorical response: “does not cause injury”, “mild injury”, “severe injury”.
- Highly unbalanced, common in pharmaceutical data.
- Measurements from 184 biological screens and 192 chemical feature predictors.

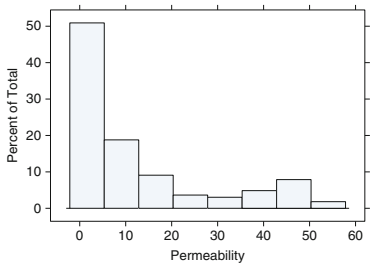
- **Objective**: Develop a model for predicting compounds’ probability of causing hepatic injury (i.e., liver damage).

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Permeability

- Data set from the pharmaceutical industry.



- 165 unique compounds.
- For each 1107 molecular fingerprints (binary sequence of numbers that represents the presence or absence of a specific molecular substructure).
- The response is highly skewed, the predictors are sparse (15.5% are present).
- Attempt to potentially reduce the need for the assay.

- **Objective**: Develop a model for predicting compounds’ permeability (the measure of a molecule’s ability to cross a membrane).

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Chemical Manufacturing Process

- Data set from a chemical process industry for producing pharmaceuticals.

  177 samples of biological material with 57 measured characteristics.

  12 of the biological starting material and 45 of the manufacturing process.

  Process variables included temperature, drying time, washing time, and concentrations of by-products at various steps.

  Some of the process measurements can be controlled, while others are observed.

  Predictors are continuous, count, categorical; some are correlated

  Some contain missing values.

  Samples are not independent (sets of samples come from the same batch of biological starting material).

- **Objective**: Develop a model to predict percent yield of the manufacturing process.

Example data sets and typical data scenarios

We can have very diverse problems as well as the characteristics of the collected data.

Fraudulent Financial Statements

- Data set from public data sources, such as U.S. Securities and Exchange Commission documents.

  Fanning and Cogger[1] sample non-fraudulent companies for important factors (e.g., company size and industry type).

  150 data points were used to train models and the 54 to evaluate them.

  The analysis started with an unidentified number of predictors derived from key areas, such as executive turnover rates, litigation, and debt structure.

  20 predictors were used in the models.

- **Objective**: Predict management fraud for publicly traded companies.

[1] Fanning K, Cogger K (1998). *Neural Network Detection of Management Fraud Using Published Financial Data*. Int. J of Intell Sys in Accounting, Finance & Management, 7(1), 21-41.

Comparing the data

	Data set					
Data characteristic	Music genre	Grant applications	Hepatic injury	Fraud detection	Permeability	Chemical manufacturing
Dimensions						
# Samples	12,495	8,707	281	204	165	177
# Predictors	191	249	376	20	1,107	57
Response characteristics						
Categorical or continuous	Categorical	Categorical	Categorical	Categorical	Continuous	Continuous
Balanced/symmetric		×		×		×
Unbalanced/skewed	×		×	×		
Independent			×		×	
Predictor Characteristics						
Continuous	×	×	×	×		×
Count	×	×	×			×
Categorical		×	×	×	×	×
Correlated/associated	×	×	×	×	×	×
Different scales	×	×	×	×		×
Missing values		×				×
Sparse					×	

Case Study: Predicting Fuel Economy

Consider a simple example that **illustrates the broad concepts of model building**.

The data set of different estimates of fuel economy for passenger cars and trucks is provided by: `fueleconomy.gov`

From the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy and the U.S. Environmental Protection Agency.

Various characteristics are recorded for each vehicle such as engine displacement or number of cylinders.

Laboratory measurements are made for the city and highway miles per gallon (MPG) of the car.

Objective: Building a model to predict the MPG for a new car line by using

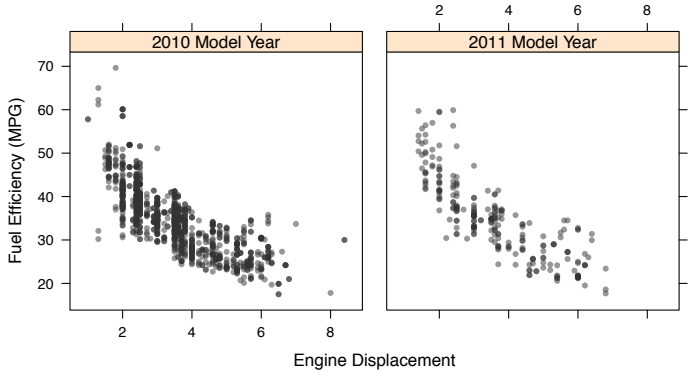
- A **single predictor**: Engine displacement (the volume inside the engine cylinders).
- A **single response**: Unadjusted highway MPG for 2010-2011 model year cars.

Case Study: Predicting Fuel Economy

Understand the data.

It can most easily be done through a graph.

- We have just one predictor and one response: We use a scatter plot.
- Left: Contains all the 2010 data.
- Right: Shows the data only for new 2011 vehicles.

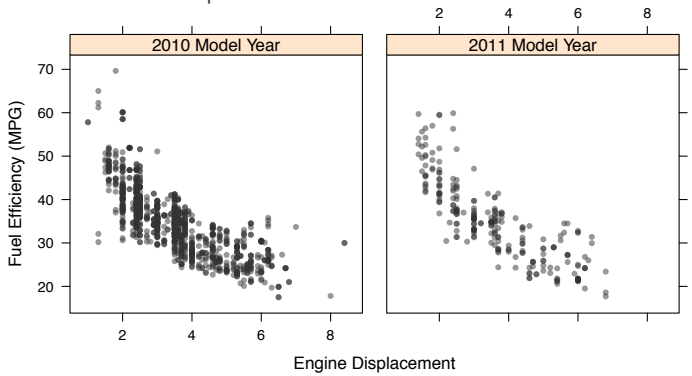


Case Study: Predicting Fuel Economy

Understand the data.

It can most easily be done through a graph.

- As engine displacement increases, the fuel efficiency drops regardless of year.
- The relationship is somewhat linear but does exhibit some curvature towards the extreme ends of the displacement axis.

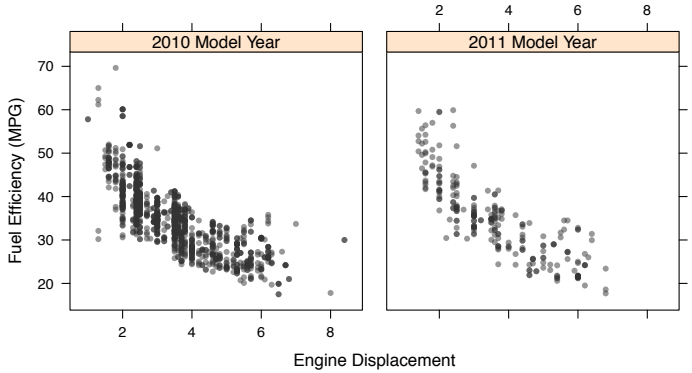


Case Study: Predicting Fuel Economy

Understand the data.

What if we had more than one predictor?

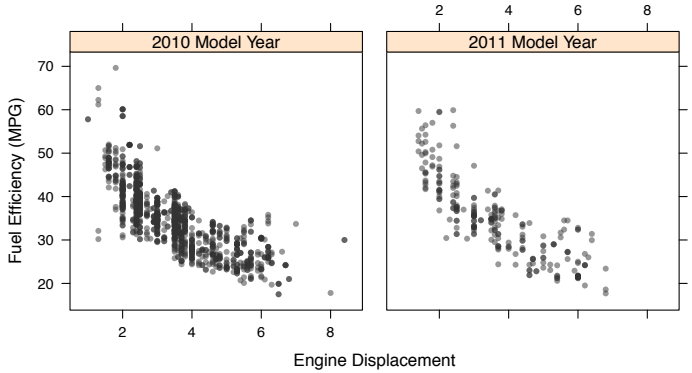
- Need to further understand characteristics of the predictors and their relationships.
- These characteristics may suggest important and necessary **pre-processing steps that must be taken prior to building a model**.



Case Study: Predicting Fuel Economy

Build and evaluate a model on the data.

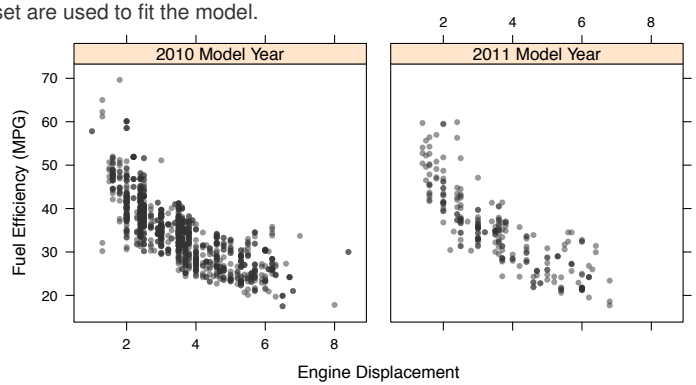
- **Standard approach**: Take a random sample of the data for model building and use the rest to understand model performance.
- Build the models using the 2010 data (1107 cars): **training set**
- Test the models using the new 2011 data (245 cars): **test** or **validation set**.



Case Study: Predicting Fuel Economy

Build and evaluate a model on the data.

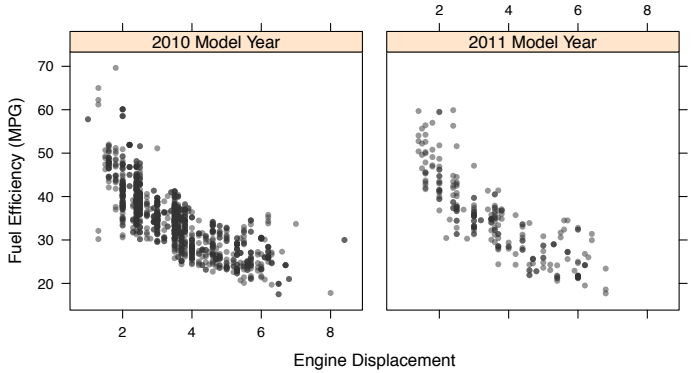
- We can **not simply evaluate model performance using the same data used to build the model**.
- Re-predict the training set data: potential to produce overly optimistic estimates.
- Alternative approach use **resampling**: different subversions of the training data set are used to fit the model.



Case Study: Predicting Fuel Economy

Measure the performance of the model.

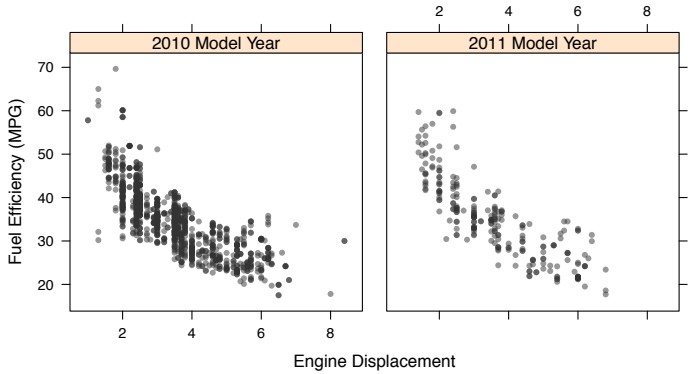
- For regression problems, the **residuals are important sources of information**.
- Computed as observed value minus the predicted value ($y_i - \hat{y}_i$).
- The **root mean squared error** is commonly used to evaluate models.
- RMSE is interpreted as how far, on average, the residuals are from zero.



Case Study: Predicting Fuel Economy

Define the relationship between the predictor and outcome.

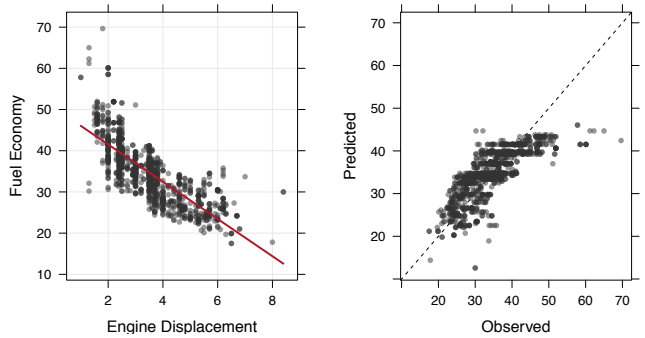
- The modeler will try various techniques to mathematically define this relationship.
- Training set used to estimate the various values needed by the model equations.
- Test set used only when a few strong candidate models have been finalized.



Case Study: Predicting Fuel Economy

First attempt: linear regression model.

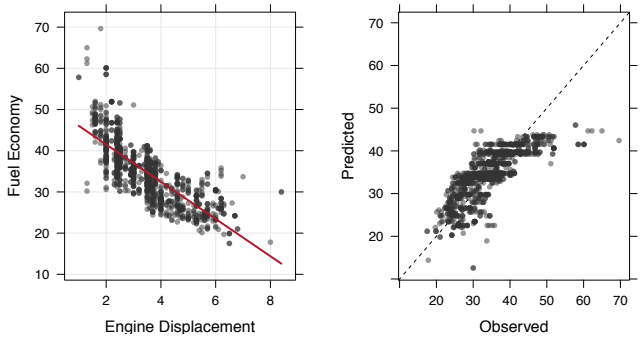
- The predicted MPG is a basic slope and intercept model.
- With the training data, we estimate the model using the **least squares method**.
- Left: A linear model fit defined by the estimated slope and intercept.
- Right: Observed and predicted MPG.



Case Study: Predicting Fuel Economy

First attempt: linear regression model.

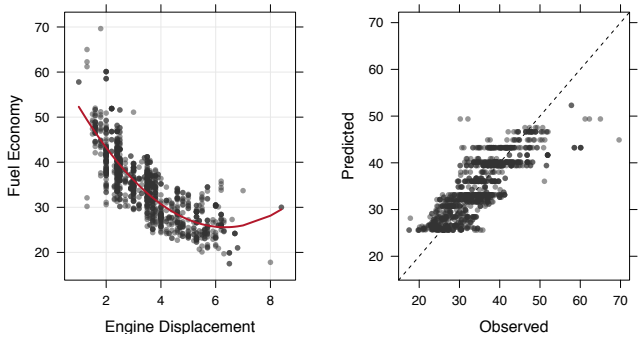
- The **model misses some of the patterns in the data**.
- Under-predicting fuel efficiency when the displacement is < than 2L or > 6L.
- We resample the data and estimate a RMSE=4.6 MPG.



Case Study: Predicting Fuel Economy

Improve the model: Introduce some nonlinearity.

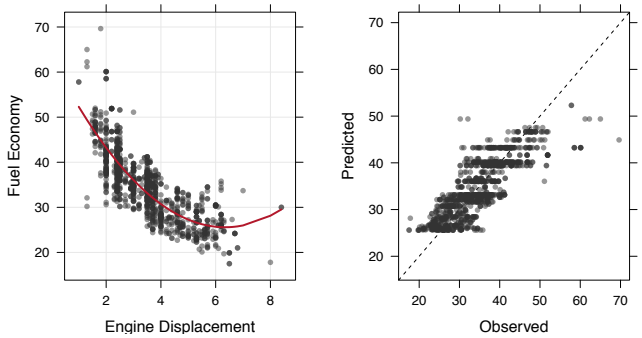
- The most basic approach is to **add complexity**.
- As for example, by adding a squared term:
  efficiency = $63.2 - 11.9 \times$ displacement + $0.94 \times$ displacement^2.
- This is a **quadratic model**: It includes a squared term.



Case Study: Predicting Fuel Economy

Improve the model: Introduce some nonlinearity.

- The quadratic term improves the model fit (RMSE = 4.2 MPG).
- Drawback: it can **perform poorly on the extremes of the predictor**.
- The model appears to be bending upwards unrealistically. Predicting new vehicles with large displacement values may produce significantly inaccurate results.



Case Study: Predicting Fuel Economy

Improve the model: Multivariate Adaptive Regression Spline[1].

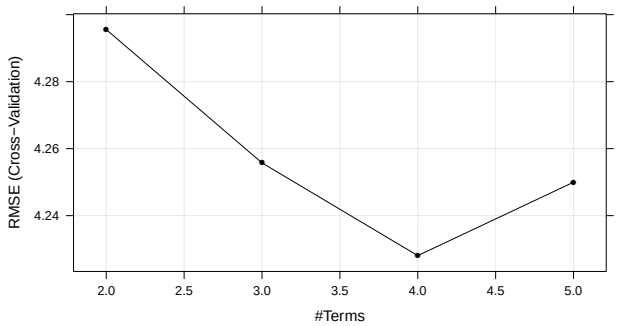
- With a single predictor, MARS can fit separate linear regression lines for different ranges of engine displacement.
- The slopes and intercepts are estimated for this model, as well as the number and size of the separate regions for the linear models.
- Unlike the linear regression models, MARS has a tuning parameter which cannot be directly estimated from the data.
- There is no analytical equation that can be used to determine how many segments should be used to model the data.
- We can try different values and use resampling to determine the appropriate value.
- Once the value is found, a final MARS model would be fit using all the training set data and used for prediction.

[1] Friedman J. (1991). *Multivariate Adaptive Regression Splines.* The Annals of Statistics, 19(1), 1-141.

Case Study: Predicting Fuel Economy

Improve the model: Multivariate Adaptive Regression Spline.

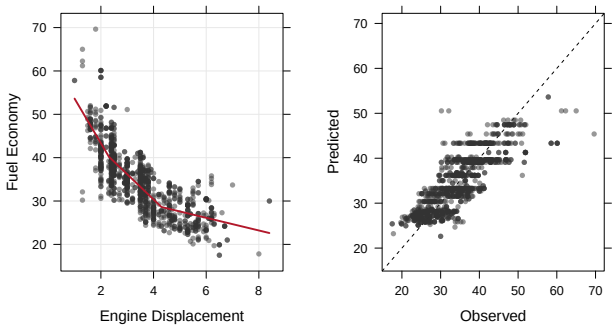
- For a single predictor, MARS can allow for up to five model terms (similar to the previous slopes and intercepts).
- The lowest RMSE value is associated with four terms, although the scale of change in the RMSE values indicates that there is some insensitivity to this tuning parameter.



Case Study: Predicting Fuel Economy

Improve the model: Multivariate Adaptive Regression Spline.

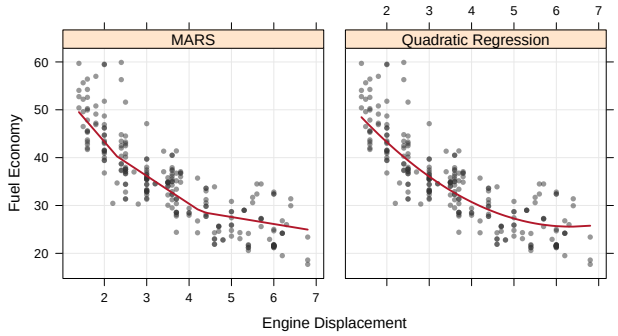
- After fitting the final MARS model with four terms, the training set fit is shown below where several linear segments were predicted.



Case Study: Predicting Fuel Economy

Compare the models on the test set

- Both models fit very similarly.
- For the **test set**: $\text{RMSE}_{\text{quadratic}} = 4.72$ MPG and the $\text{RMES}_{\text{MARS}} = 4.69$ MPG.
- Either model would be appropriate for the prediction of new car lines.



Case Study: Predicting Fuel Economy

Considerations on the model building process: Data Splitting

How allocate data to model building and evaluating performance?

- The primary interest was to predict the fuel economy of *new* vehicles, which is not the same population as the data used to build the model.
- This means that, to some degree, we are testing how well the model **extrapolates** to a different population.
- If we were interested in predicting from the same population of vehicles (i.e., **interpolation**), taking a simple random sample of the data would be more appropriate.
- How the training and test sets are determined reflects how the model will be applied.

Case Study: Predicting Fuel Economy

Considerations on the model building process: Data Splitting

How much data should be allocated to the training and test sets?

- Generally **it depends on the situation**.
- If the pool of data is small, the data splitting decisions can be critical.

  A small test would have limited utility as a judge of performance.

  A sole reliance on resampling techniques might be more effective.

- Large data sets reduce the criticality of these decisions.

Case Study: Predicting Fuel Economy

Considerations on the model building process: Predictor data.

The fuel economy example has revolved around one of many predictors: the engine displacement.

- The original data contain many other factors: number of cylinders, the type of transmission, and the manufacturer.
- An earnest attempt to predict the fuel economy would examine as many predictors as possible to improve performance.
- Using more predictors, it is likely that the RMSE for the new model cars can be driven down further.
- Some investigation into the data can also help. For example, none of the models were effective at predicting fuel economy when the engine displacement was small. Inclusion of predictors that target these types of vehicles would help improve performance.

Feature selection, the process of determining the minimum set of relevant predictors needed by the model, is a common way to approach the problem.

Case Study: Predicting Fuel Economy

Considerations on the model building process

Considerations on the model building process: Estimating performance.

We used two techniques to determine the effectiveness of the model.

1. **Quantitative assessments** of statistics (i.e., the RMSE) using resampling help the user understand how each technique would perform on new data.
2. **Simple visualizations** (e.g., plotting the observed and predicted values) to discover areas of the data where the model does particularly good or bad.

This type of qualitative information is critical for improving models and is lost when the model is gauged only on summary statistics.

Case Study: Predicting Fuel Economy

Considerations on the model building process: Evaluating several models.

Three different models were evaluated.

- The **No Free Lunch Theorem**[1] argues that, without having substantive information about the modeling problem, there is no single model that will always do better than any other model.
- A strong case can be made to try a wide variety of techniques, then determine which model to focus on.

In the fuel economy example, a simple plot of the data shows that there is a nonlinear relationship between the outcome and the predictor.

Given this knowledge, we might exclude linear models from consideration, but there is still a wide variety of techniques to evaluate.

One might say that “model X is always the best performing model” but, for these data, a simple quadratic model is extremely competitive.

[1] Wolpert D (1996). “The Lack of a priori Distinctions Between Learning Algorithms”. Neural Computation, 8(7), 1341–1390 (`http://www.no-free-lunch.org`)

Case Study: Predicting Fuel Economy

Considerations on the model building process: Model selection.

At some point in the process, a specific model must be chosen.

This example demonstrated two types of model selection:

- **Between models**: The linear regression model did not fit well and was dropped.
- **Within models**: For MARS, the tuning parameter was chosen using cross-validation.

In either case, we relied on **cross-validation** and the test set to produce quantitative assessments of the models to help us make the choice.

Because we focused on a single predictor, which will not often be the case, we also made visualizations of the model fit to help inform us.

At the end of the process, the MARS and quadratic models appear to give equivalent performance. However, knowing that the quadratic model might not do well for vehicles with very large displacements, our intuition might tell us to favor the MARS model.

Few details about the homework

The homework will represent 40% of your final grade:

- **HW1**: Data analysis (September 7 – 25, 2017)[1]
- **HW2**: Linear methods (October 9 – 23, 2017)[1]
- **HW3**: Nonlinear methods (November 6 – 20, 2017)[1]
- **HW4**: Evolutional computation or Fuzzy system (December 1 – 12, 2017)[1,2]

The **grading and evaluation** of the homework is based on the:

- Level of difficulty (in case you choose another case-study).
- Comments and discussion on the results.
- Quality of reporting.

Extra assignment for the post-grad students ...

[1] **Dates might be subject to change... it will depend on you!!**
[2] Not compulsory. Extra points on final HW grade.