Predictive Modeling

Department of Mathematical Sciences
Qiuying Sha, Professor

Michigan Tech

Part I. General Strategies

- There are total 4 chapters in this part. The major goal is data pre-processing.

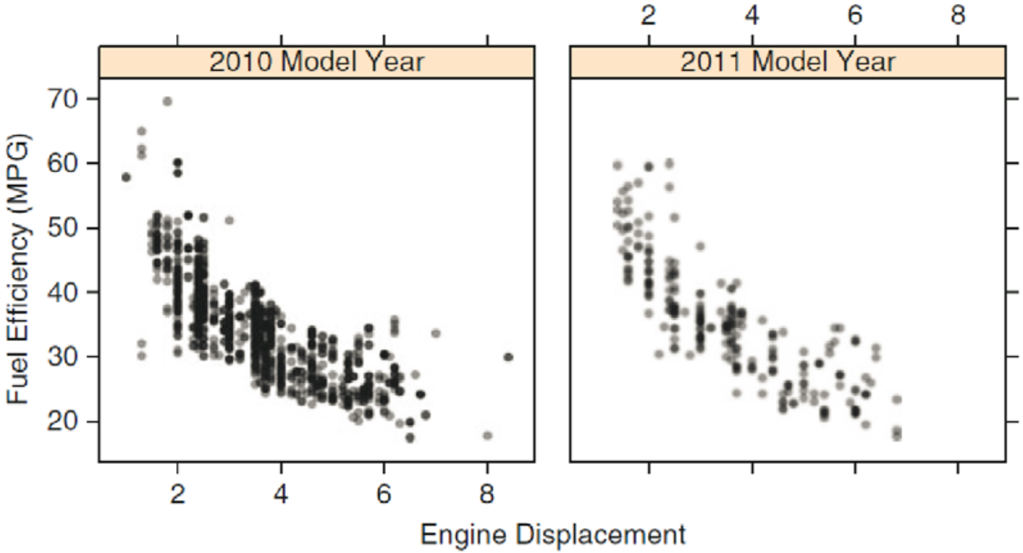
Chapter 2. A Short Tour of the Predictive Modeling Process

2.1 Case Study: Predicting Fuel Economy

Data:

- Response: unadjusted highway MPG for 2010–2011 model year cars.
- Predictor: engine displacement (the volume inside the engine cylinders)

First step: understand the data.



The relationship is somewhat linear but does exhibit some curvature towards the extreme ends of the displacement axis.

Second step: build and evaluate a model on the data.

- Decide training set and test set:
  - A standard approach is to take a random sample of the data for model building and use the rest to understand model performance.
  - We want to predict the MPG for a *new* car. In this situation, models can be created using the 2010 data (containing 1,107 vehicles) and tested on the 245 new 2011 cars.
- Decide how to measure performance of a model.
  - The root mean squared error (RMSE). RMSE is interpreted as how far, on average, the residuals are from zero.

MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$RMSE = \sqrt{MSE}$$

- Fit candidate models using training set.
  - Linear regression model:

    efficiency = 50.6 – 4.5 × displacement

    RMSE = 4.6 MPG by 10-fold cross-validation

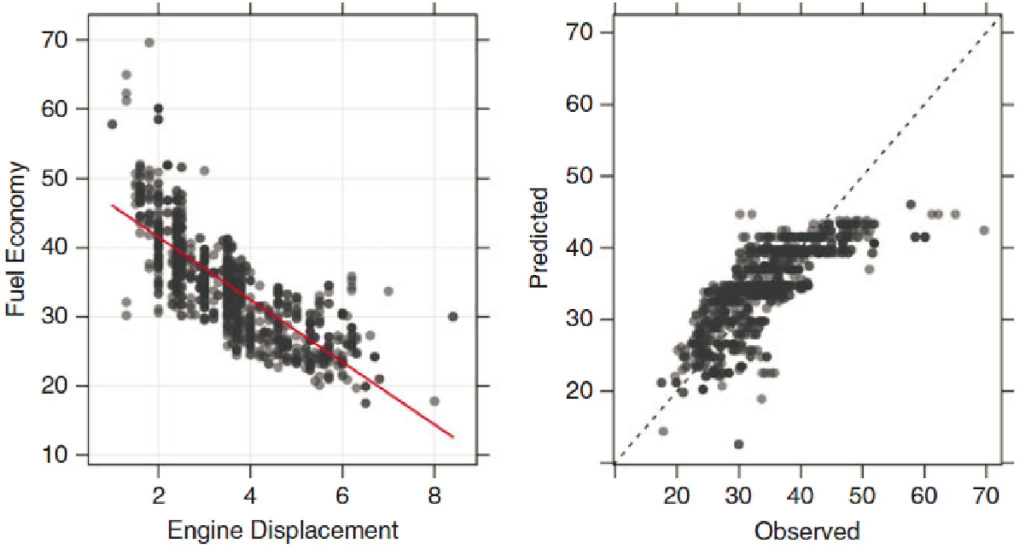


Fig. 2.2: Quality of fit diagnostics for the linear regression model. The training set data and its associated predictions are used to understand how well the model works

These plots shows: this model misses some of the patterns in the data, such as under-predicting fuel efficiency when the displacement is less than 2 L or above 6 L.

- Fit candidate models using training set.
  - Quadratic model

    efficiency = 63.2 − 11.9 × displacement + 0.94 × displacement^2

    RMSE = 4.2 MPG

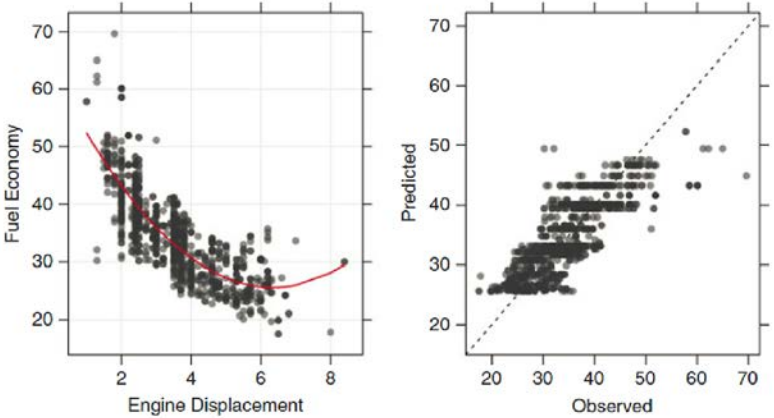


Fig. 2.3: Quality of fit diagnostics for the quadratic regression model (using the training set)

- The multivariate adaptive regression spline (MARS) model (Friedman 1991)

  MARS can fit separate linear regression lines for different ranges of engine displacement.

  There is a *tuning parameter* which cannot be directly estimated from the data. Tuning parameter is the # of segments used to model the data

  The MARS model has internal algorithms for making this determination, the user 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.

  RMSE = 4.2 MPG

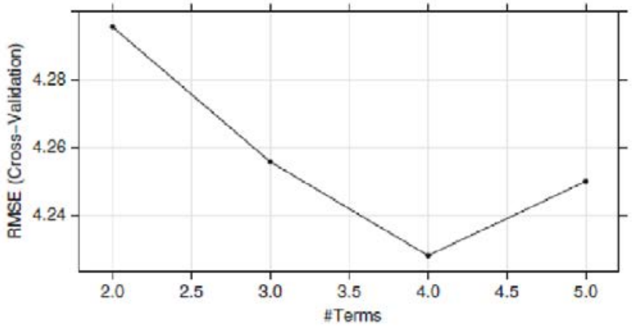


Fig. 2.4: The cross-validation profile for the MARS tuning parameter

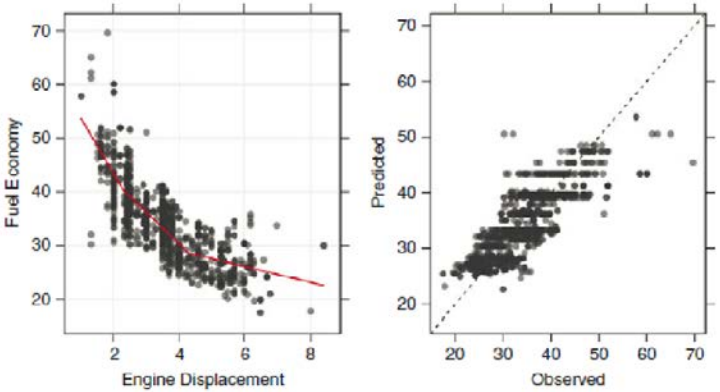


Fig. 2.5: Quality of fit diagnostics for the MARS model (using the training set). The MARS model creates several linear regression fits with change points at 2.3, 3.5, and 4.3 L

- Evaluate the performance of a few strong candidate models using test set.

The quadratic regression and MARS models were evaluated on the test set.

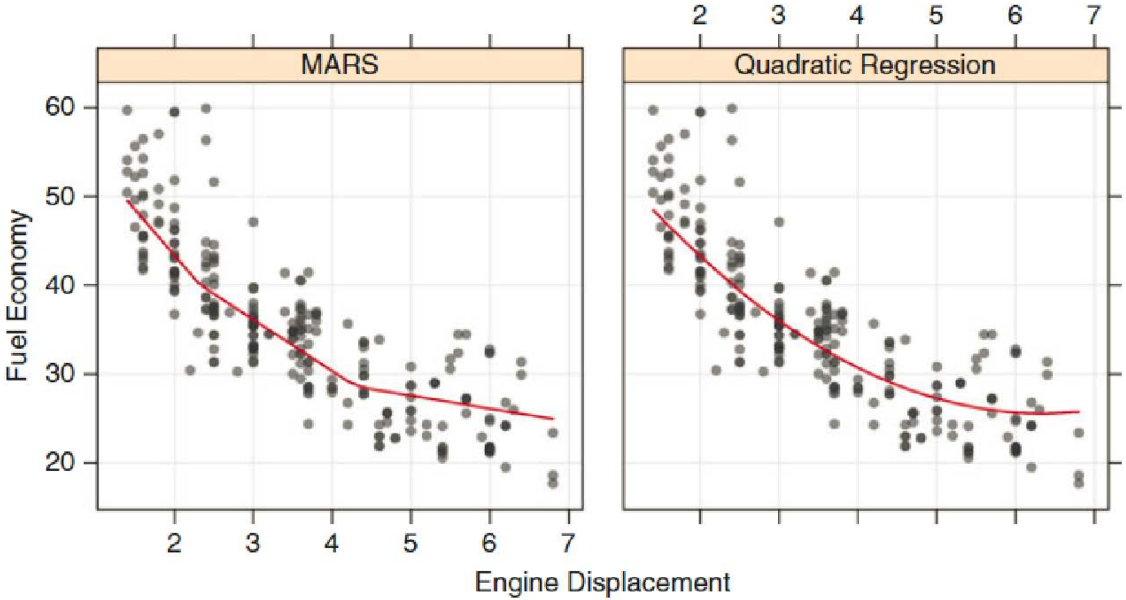


Fig. 2.6: The test set data and with model fits for two models

Conclusion:

- Fit candidate models using training set.

2.2 Themes

There are several aspects of the model building process that we will discuss in this section.

- ***Data Splitting***

  Allocate data to test set and training set to build models and evaluate performance of a model.

  How much data should be allocated to the training and test sets? It depends on the situation.

- ***Predictor Data***

  This example has revolved around one of many predictors: the engine displacement.

  The original data contain many other factors, such as the number of cylinders, the type of transmission, and the manufacturer.

  Using more predictors, it is likely that the RMSE for the new model cars can be driven down further.

  An aspect of modeling that was not discussed here was feature selection: the process of determining the minimum set of relevant predictors needed by the model.

- ***Estimating Performance***

  Before using the test set, two techniques were employed to determine the effectiveness of the model.

  - Quantitative assessments of statistics (i.e., the RMSE) using resampling help the user understand how each technique would perform on new data.
  - Visualizations of a model, such as plotting the observed and predicted values.

- ***Evaluating Several Models***

  Without having substantive information about the modeling problem, there is no single model that will always do better than any other model.

  Try a wide variety of techniques, then determine which model to focus on.

- ***Model Selection***

  This example demonstrated two types of model selection.

  - ➢ First, we chose some models over others: the linear regression model did not fit well and was dropped. In this case, we chose *between models*.
  - ➢ There was also a second type of model selection in MARS. For MARS, the tuning parameter was chosen using cross validation. This was also model selection where we decided on the *type of MARS model* to use. In this case, we did the selection *within* different MARS models.

2.3 Summary

To get a reliable, trustworthy model for predicting new samples, we must

- Understand the data and the objective of the modeling.
- Pre-process and split the data.
- Building, evaluating, and selecting models.