Chapter 2
A Short Tour of the Predictive Modeling Process

Before diving in to the formal components of model building, we present a simple example that illustrates the broad concepts of model building. Specifically, the following example demonstrates the concepts of data "spending," building candidate models, and selecting the optimal model.

2.1 Case Study: Predicting Fuel Economy

The fueleconomy.gov web site, run by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy and the U.S. Environmental Protection Agency, lists different estimates of fuel economy for passenger cars and trucks. For each vehicle, various characteristics are recorded such as the engine displacement or number of cylinders. Along with these values, laboratory measurements are made for the city and highway miles per gallon (MPG) of the car.

In practice, we would build a model on as many vehicle characteristics as possible in order to find the most predictive model. However, this introductory illustration will focus high-level concepts of model building by using a single predictor, engine displacement (the volume inside the engine cylinders), and a single response, unadjusted highway MPG for 2010–2011 model year cars.

The first step in any model building process is to understand the data, which can most easily be done through a graph. Since we have just one predictor and one response, these data can be visualized with a scatter plot (Fig. 2.1). This figure shows the relationship between engine displacement and fuel economy. The "2010 model year" panel contains all the 2010 data while the other panel shows the data only for new 2011 vehicles. Clearly, 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.

M. Kuhn and K. Johnson, *Applied Predictive Modeling*,
DOI 10.1007/978-1-4614-6849-3_2,

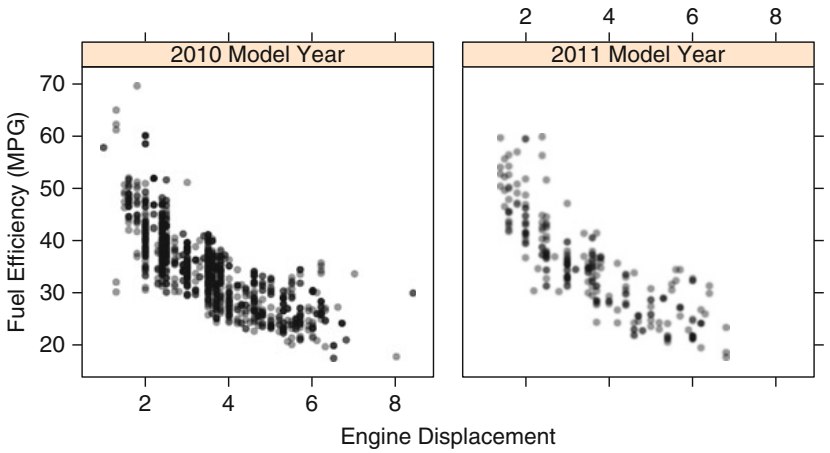

Fig. 2.1: The relationship between engine displacement and fuel efficiency of all 2010 model year vehicles and new 2011 car lines

If we had more than one predictor, we would need to further understand characteristics of the predictors and the relationships among the predictors. These characteristics may suggest important and necessary pre-processing steps that must be taken prior to building a model (Chap. 3).

After first understanding the data, the next step is to build and evaluate a model on the data. A standard approach is to take a random sample of the data for model building and use the rest to understand model performance. However, suppose we want to predict the MPG for a *new* car line. In this situation, models can be created using the 2010 data (containing 1,107 vehicles) and tested on the 245 new 2011 cars. The common terminology would be that the 2010 data are used as the model "training set" and the 2011 values are the "test" or "validation" set.

Now that we have defined the data used for model building and evaluation, we should decide how to measure performance of the model. For regression problems where we try to predict a numeric value, the residuals are important sources of information. Residuals are computed as the observed value minus the predicted value (i.e., $y_i - \widehat{y}_i$). When predicting numeric values, the root mean squared error (RMSE) is commonly used to evaluate models. Described in more detail in Chap. 7, RMSE is interpreted as how far, on average, the residuals are from zero.

At this point, the modeler will try various techniques to mathematically define the relationship between the predictor and outcome. To do this, the training set is used to estimate the various values needed by the model equations. The test set will be used only when a few strong candidate models have been finalized (repeatedly using the test set in the model build process negates its utility as a final arbitrator of the models).

Suppose a linear regression model was created where the predicted MPG is a basic slope and intercept model. Using the training data, we estimate the

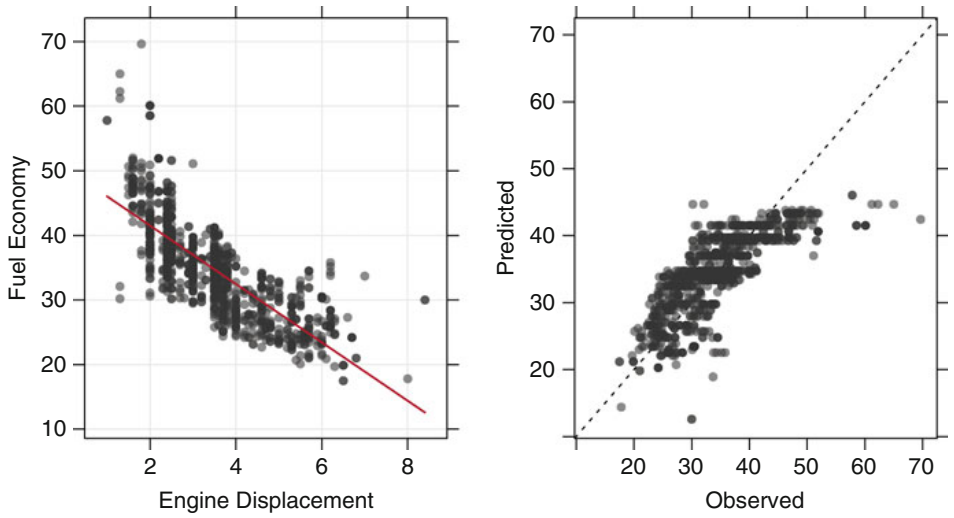

Fig. 2.2: Quality of fit diagnostics for the linear regression model. The training set data and its associated predictions are used to understand how well the model works

intercept to be 50.6 and the slope to be −4.5 MPG/liters using the method of least squares (Sect. 6.2). The model fit is shown in Fig. 2.2 for the training set data.[1] The left-hand panel shows the training set data with a linear model fit defined by the estimated slope and intercept. The right-hand panel plots the observed and predicted MPG. These plots demonstrate that 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.

When working with the training set, one must be careful not to simply evaluate model performance using the same data used to build the model. If we simply re-predict the training set data, there is the potential to produce overly optimistic estimates of how well the model works, especially if the model is highly adaptable. An alternative approach for quantifying how well the model operates is to use *resampling*, where different subversions of the training data set are used to fit the model. Resampling techniques are discussed in Chap. 4. For these data, we used a form of resampling called 10-fold cross-validation to estimate the model RMSE to be 4.6 MPG.

Looking at Fig. 2.2, it is conceivable that the problem might be solved by introducing some nonlinearity in the model. There are many ways to do this. The most basic approach is to supplement the previous linear regression model with additional complexity. Adding a squared term for engine displacement would mean estimating an additional slope parameter associated with the square of the predictor. In doing this, the model equation changes to

[1] One of our graduate professors once said "the only way to be comfortable with your data is to never look at it."

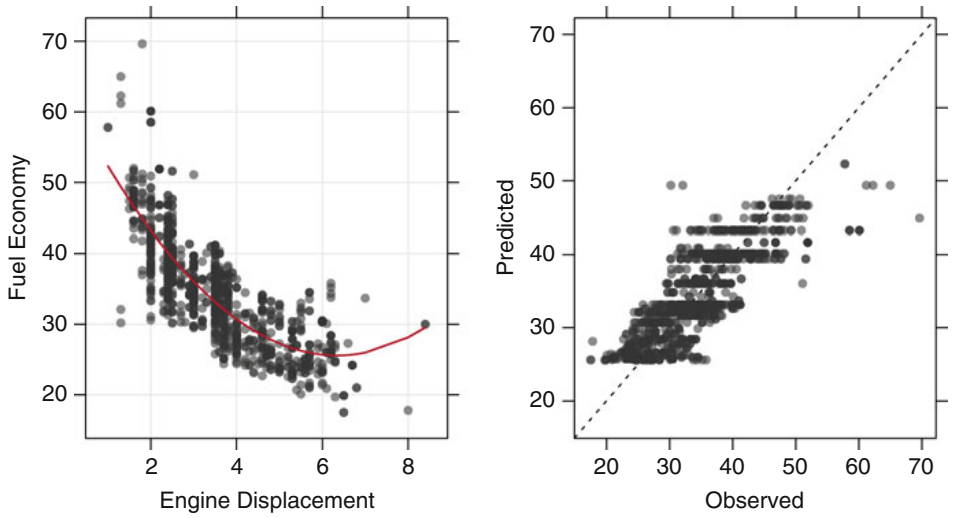

Fig. 2.3: Quality of fit diagnostics for the quadratic regression model (using the training set)

$$\text{efficiency} = 63.2 - 11.9 \times \text{displacement} + 0.94 \times \text{displacement}^2$$

This is referred to as a *quadratic model* since it includes a squared term; the model fit is shown in Fig. 2.3. Unquestionably, the addition of the quadratic term improves the model fit. The RMSE is now estimated to be 4.2 MPG using cross-validation. One issue with quadratic models is that they can perform poorly on the extremes of the predictor. In Fig. 2.3, there may be a hint of this for the vehicles with very high displacement values. The model appears to be bending upwards unrealistically. Predicting new vehicles with large displacement values may produce significantly inaccurate results.

Chapters 6–8 discuss many other techniques for creating sophisticated relationships between the predictors and outcome. One such approach is the multivariate adaptive regression spline (MARS) model (Friedman 1991). When used 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, this technique 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. While 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 and used for prediction.

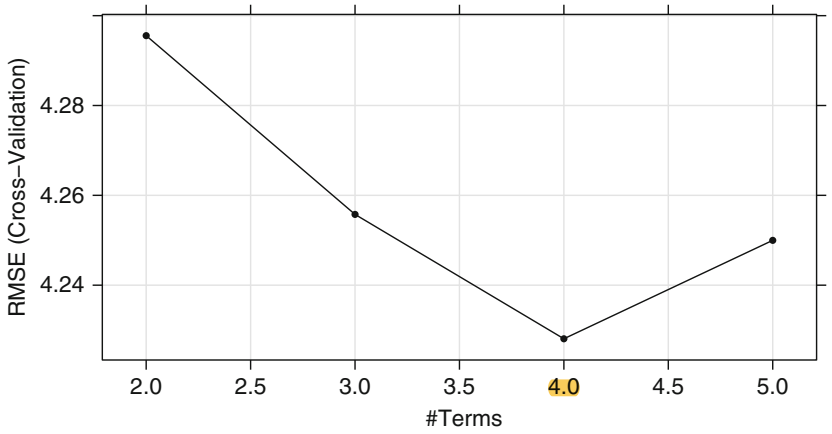

Fig. 2.4: The cross-validation profile for the MARS tuning parameter

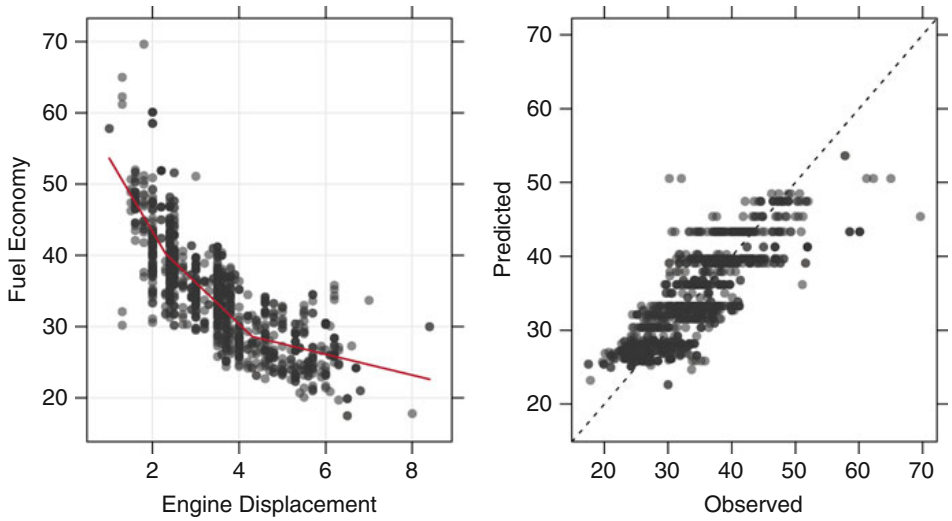

Fig. 2.5: Quality of fit diagnostics for the MARS model (using the training set). The MARS model creates several linear regression fits with change points at 2.3, 3.5, and 4.3 L

For a single predictor, MARS can allow for up to five model terms (similar to the previous slopes and intercepts). Using cross-validation, we evaluated four candidate values for this tuning parameter to create the resampling profile which is shown in Fig. 2.4. 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. The RMSE associated with the optimal model was 4.2 MPG. After fitting the final MARS model with four terms, the training set fit is shown in Fig. 2.5 where several linear segments were predicted.

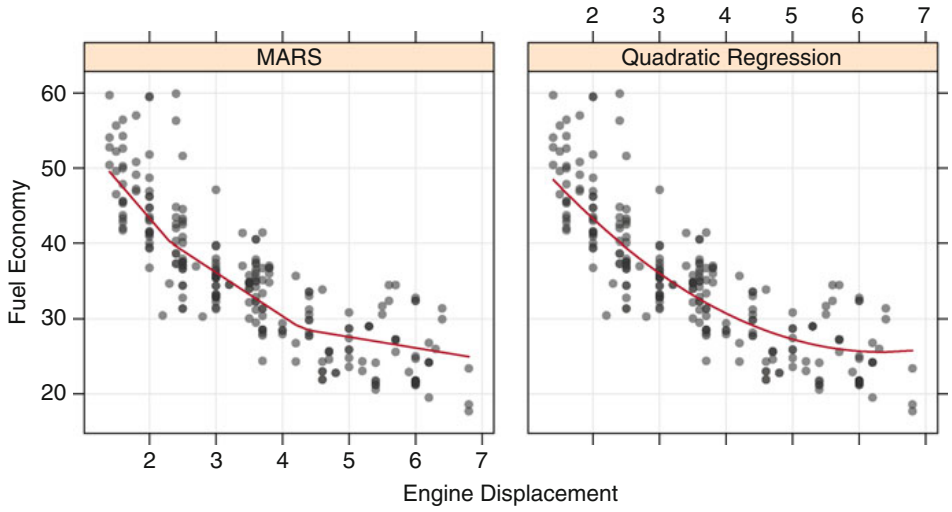

Fig. 2.6: The test set data and with model fits for two models

Based on these three models, the quadratic regression and MARS models were evaluated on the test set. Figure 2.6 shows these results. Both models fit very similarly. The test set RMSE values for the quadratic model was 4.72 MPG and the MARS model was 4.69 MPG. Based on this, either model would be appropriate for the prediction of new car lines.

2.2 Themes

There are several aspects of the model building process that are worth discussing further, especially for those who are new to predictive modeling.

Data Splitting

Although discussed in the next chapter, how we allocate data to certain tasks (e.g., model building, evaluating performance) is an important aspect of modeling. For this example, the primary interest is 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 should reflect how the model will be applied.

How much data should be allocated to the training and test sets? It generally 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. In this case, a sole reliance on resampling techniques (i.e., no test set) might be more effective. Large data sets reduce the criticality of these decisions.

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. 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.

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. This common task is discussed in Chap. 19.

Estimating Performance

Before using the test set, two techniques were employed to determine the effectiveness of the model. First, quantitative assessments of statistics (i.e., the RMSE) using resampling help the user understand how each technique would perform on new data. The other tool was to create simple visualizations of a model, such as 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.

Evaluating Several Models

For these data, three different models were evaluated. It is our experience that some modeling practitioners have a favorite model that is relied on indiscriminately. The "No Free Lunch" Theorem (Wolpert 1996) argues that,

without having substantive information about the modeling problem, there is no single model that will always do better than any other model. Because of this, a strong case can be made to try a wide variety of techniques, then determine which model to focus on. In our 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.

Model Selection

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

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. One goal of this book is to help the user gain intuition regarding the strengths and weakness of different models to make informed decisions.

2.3 Summary

At face value, model building appears straightforward: pick a modeling technique, plug in data, and generate a prediction. While this approach will generate a predictive model, it will most likely *not* generate a reliable, trustworthy model for predicting new samples. To get this type of model, we must first understand the data *and* the objective of the modeling. Upon understanding the data and objectives, we then pre-process and split the data. Only after these steps do we finally proceed to building, evaluating, and selecting models.