0.1 The Chapter's Case Study: Predicting Fuel Economy

Contextual Notes From the Reading

- Website: fueleconomy.gov
- Dataset records vehicle characteristics like:
  - Engine Displacement
  - Number of cylinders
  - Laboratory Measurements of MPG
  - etc ...
- This example focuses on high-level design principles:
  - Uses a single predictor (engine displacement), and a single response (unadjusted highway MPG) for 2010 to 2011 model year cars.
- Steps to Solve the Case Study:
  1. Understand the Data. This case uses graphs to visualize displacement to efficiency.
  2. Build and Evaluate a Model. The standard approach is to take random samples; but in this case we can use the 2010 data as the training set, and the 2011 data as the validation set.
  3. Measure Performance. For regression problems that predict a numeric, the residual data contains important information.
     - Residuals are computed as the observed value minus the predicted value: $y_i - \hat{y}_i$.
     - The root mean square error (RMSE) is commonly used to evaluate models. This is evaluated as how far Residuals are from zero.
  4. Define Relationships Between Predictor and Outcome.
     - The training set is used to estimate values needed by the model.
     - The test set is used once a few strong candidate models have been finalized. If used too much beforehand, it negates its utility.

5. Resampling. Different subversions of the training data set are used to fit the model. In this model, 10-fold cross-validation was used to estimate the RMSE
6. Trying Other Models. In this case we try a quadratic model that contains a squared term. Quadratic models can perform poorly on the extremes of a data set.
   - Another model is the multivariate adaptive regression spline (MARS).
   - MARS fits separate linear regression lines for different ranges.
   - MARS has a, 'tuning parameter' which cannot be directly estimated from the data.
   - MARS allows up to five model terms.
7. The lowest RMSE value associated with four terms of the MARS model was chosen as the best predictor.

0.2 Themes

Data Splitting How we allocate data to tasks. In this example, we tested how the data extrapolates to a different population. If we wanted to predict from the same population, instead of a new population, a random sample would be more appropriate using interpolation.

Predictor Data This example used engine displacement as a predictor. We should use as many predictors as possible when building models. Look at how the case study was unable to predict fuel economy for engines with low displacement. More predictors could help us dial in these values with better accuracy.

Estimating Performance We used RMSE for quantitative assessment and resampling to help the user understand how techniques perform on the dataset. We used visualizations of a model to discover how it performed.

Evauluating Several Methods We evaluated three different models for our data. Without having substantive information about the problem, no single method could be said to perform better than another. It is always recommended to try multiple, widely variable techniques; then limit the choices based on the results.

Model Selection There are different types of model selection. In this case, we chose specific models, like linear regression, and also chose a type of a model, like the MARS parameters. In both cases, cross validation produced a quantitative assessment which allowed us to further filter out model choices.

Summary The process seems simple: Pick a model technique, plug in data, generate a prediction. Although this produces a predictive model, it most likely will not produce a reliable, and trustworthy model for new samples. We must understand both the data, and the objective in creating the model in the first place. We then must pre-process and select relevant data for our models.