STAT 302 - Chapter 4: Additional Topics in Regression - Part 2

Harsha Perera

Key Topics

- Cross-Validation
- Cross-Validation Correlation
- Shrinkage

Cross-Validation

- In chapters 1,2 and 3 we looked at fitting
  - Simple linear regression models
  - Multiple linear regression models

- We did validated the models based on
  - Checking model assumptions
  - R^2
  - SSE: $\sum(y - \hat{y})^2$
  - MSE

- But these methods only assess the accuracy of predicting data cases that were used to build and fit the model.

- Should we always expect the same level of accuracy to hold when using the model to predict completely new data cases ?
  No

Cross-Validation. . .

- Can we get some idea, in advance, for how the model might do when applied to new data ?
- One technique for doing this is Cross-Validation

Cross-Validation - Steps

1. Split the data set into two parts (simplest version)
   - training data (training sample)
   - test data (holdout sample)

2. Build and fit the model with training data

3. Predict the responses for the test data

4. Calculate the prediction accuracy by comparing the predicted values and actual values in the test data set

Cross-validation - Steps

- Prediction accuracy: test data vs. training data
  - Should be less accurate with test data
  - Should not be lot worse

- How should we split data into training and test sets?
  - Randomly assign data cases into training and test sets
  - 70% - 30% or 60% - 40% etc.

Example 4.3: House prices in NY

- There are 53 homes in this data set

- Let's split 53 cases into
  - training set: first 35 cases
  - test set: rest of the 18 cases

- For simplicity let look at the simple linear regression model between *Size* of the house and *Price* of the house

Cross-Validation Correlation

- Alternative way to assess the effectiveness of the training model for predicting responses in the testing sample.

- The correlation between actual (y) and predicted $\hat{y}$ in the test set is called Cross-Validation Correlation

- Usually, cross-validation correlation is smaller than the correlation between dependent/predictor variables in the training set

Shrinkage

- Difference between R^2 of the training sample and the square of cross-validation correlation
- In general, we should look for shrinkage $< 10\%$
- Shrinkage helps us to understand how well a model predicts the test set

Cross-Validation - Additional Notes

- 5-fold cross-validation
  - Split data into 5 samples
  - Consider one of the samples as the test set and rest as the training set
  - Repeat the process 5 times
  - Find the model with smallest MSE

- Leave one out cross-validation
  - An extreme version
  - One data point is considered as the test set and rest as training set
  - Repeat the process for n times
  - Find the model with smallest prediction error