Technical Appendix

Jiaxi Wu

1. Data management and overall workflow

For the CPI-U prediction, we estimate the performance of linear regression and CART. We follow the following workflow:

(1). Divide the data into a training sample and a hold-out sample.

(2). Run each prediction algorithm on training sample. For linear regression, just run the algorithm on the full training sample and store the result. For CART, we tune empirically on the training sample by cross-validation.

(3). After storing all the prediction functions, we run each algorithm on the hold-out sample and produce the statistics we are interested in such as R^2.

2. Clustering & stratification

If the data is clustered or stratified, we need to take the clusters into account when constructing the hold-out sample and in choosing fold for cross-validation by randomizig on the level of clusters. But in this assignment, all the indices are quantitative and not clustered.

3. Tuning & Cross Validation

For CART, we choose complexity parameters like depth by cross-validation within the training sample. We randomly divide the training data into 5 folds. For each fold, we fit the model for every tuning parameter value on all other folds and predict on the given fold. Then we obtain one prediction for every tuning parameter on each fold and average the prediction loss over the full training sample for each tuning parameter. Based on the loss, we choose the tuning parameter and then fit the model with the chosen tuning parameter on the full training sample. Finally evaluate the fitted function on the hold-out sample to test the performance of model.

When considering multiple tuning parameters, such as the depth of the tree and the minimal size of a non-terminal node, we choose these parameters jointly; that is, we run the above procedure over a grid of multi-dimensional parameter.

4. Feature Representations

The Consumer Price Index (CPI) measures the change in prices paid by consumers for goods and services. The CPIs are based on prices of food, clothing, shelter, fuels, transportation, doctors' and dentists' services, drugs, and other goods and services that people buy for day-to-day living. In calculating the index, price changes for the various items in each location are aggregated using weights, which represent their importance in the spending of the appropriate population group. All the indices have been normalized.

Certain variable transformations that would be redundant for an unregularized linear model and create collinearities can be valuable in penalized linear models or other machine-learning predictors: if meaningful coarser geographic subdivisions are included, the model may obtain better predictions with less complexity if reconstructing them from their components would incur a high complexity penalty. We therefore typically include coarser, redundant measures in addition to the base data when they are available.