1. Run a linear regression where you model the growth rate for the entire sample using a small macroeconomic model (reg\_small in the code). Report and discuss your results. In particular, report the estimates, the standard errors, the t-values, and the p-values. Discuss if the coefficients have the expected signs, or if any of them have a counterintuitive sign.
   1. TB3SMFFM: 0.519
   2. TB6SMFFM: -0.216
   3. UMCSENTx: 0
   4. UNRATE: -0.28
   5. PAYEMS: 1.481
   6. HOUST: 0.014
   7. PCEPI: -0.011
   8. The only variables that had any statistical significance were TB3SMFFM, PAYEMS, HOUST, and PCEPI
2. Run a linear regression where you model the growth rate for the entire sample using the entire data set available (reg\_large\_ols). You do not need to report this regression output because this table will be giant. Are the coefficients on PAYEMS, HOUST, and PCEPI significant in this very large regression? Are the signs and magnitudes similar to what you had in the small regression? What could be causing these differences in significance between the small and the large regression?
   1. None of the listed variables had any statistical significance which isn’t saying much because most of the variables had no significance while among the ones that did have any significance, only one had the highest level of significance (<0.0001)
3. Now we’ll generate the best tree for the whole data sample. Use rpart to estimate a tree that predicts the growth rate of industrial production using the entire data set. Note that we are using the exact same data we used for the regression model in Q2.
   1. What is the strongest predictor of high or low growth rates in industrial production (the first node of the tree)? You don’t have to interpret the meaning of this variable or look up the fred\_md legends, just report the variable in the first node.
      1. The strongest predictor is CMRMT SPLx
   2. What is the second node of the tree? You don’t have to interpret the meaning of this variable or look up the fred\_md legends, just report the variable in the first node.
      1. The second strongest is PAYEMS
4. Let’s estimate a simple AR(1) model for industrial production.
   1. Do the ACF and PACF indicate any evidence in favor of time dependence?
      1. The ACF has a large drop off at any lag while the PACF shows time dependance after the first lag
   2. What is the coefficient on the first lag from the AR(1) model?
      1. 0.186
   3. If you had to pick a model other than an AR(1) model, what ARMA models would be a reasonable starting point for the growth rate of industrial production.
5. Now we will check the predictive performance of our model. To assess this, I will try a test-train sample.
   1. Report the regression output for the small expert judgment regression for the training sample reg\_small1. Are they comparable to the regression output you got for the whole sample or do they change a lot if some observations are excluded because they are in the test sample?
      1. Most coefficients look the same, but now only PAYEMS is statistically significant
   2. Report the decision tree graphs for the training sample. Are the first two nodes the same as the decision tree for the whole sample or are they affected by the training/ test samples?
      1. The first two nodes are now USTRADE and CMRSTSPLx rather than CMRSTSPLx and PAYEMS
   3. Report the coefficients for the AR(1) model for the training sample. Are they similar to the results for the whole sample or do they change a lot if some observations are excluded because they are in the test sample?
      1. 0.241
   4. Does the forest give you any information about the link between the unemployment rate and the potential predictors?
6. Now let’s predict on the test sample. The code will also build a random forest model that predicts the growth rate of industrial production 200 trees (it picks 200 cross-validation models for you in the background, and averages the predictions across all 200 trees) and it will predict based on the AR(1) model.
   1. What is the RMSE for the small linear regression model, what is the RMSE for the giant regression that uses all o the data, what is the RMSE for the single tree, what is the RMSE for the forest, and what is the RMSE for the AR(1) model? Which one of these models did best on the test sample?
      1. Small OLS: 0.822
      2. Unrestricted OLS: 9.079
      3. Tree: 0.945
      4. Forest: 0.946
      5. AR1: 1.035
   2. If you had never taken an economics class and treated this as a data science problem, is it better to go with an AR(1) model or with a kitchen sink regression that includes all of the available data as explanatory variables?
      1. Based on the RMSE in the previous part the best model would be the small OLS that we first calculated which also was the model with the most significant variables. Between the AR1 and the kitchen sink approach I would choose the AR1
   3. Based on your answer to part a), do you think that expert judgment and our knowledge from other econ classes were helpful when we tried to build our predictive model here?
      1. Having real world knowledge of the data and its drivers will always be more important than coaxing regressions out of data. As mentioned in class: *its why economists have jobs*
7. Optional: the code also includes a Ridge and LASSO approach that uses the entire data set as potential right-hand-side variables, but it eliminates the variables that do not add predictive power based on the Ridge and LASSO criterion approaches.
   1. What is the RMSE for the LASSO approach? What is the RMSE for the ridge approach.
      1. LASSO: 1.083
      2. Ridge: 0.861
   2. If you had never taken an economics class and treated this as a data science problem, is it better to go with LASSO/ Ridge or with a kitchen sink regression that includes all of the available data as explanatory variables?
      1. The Ridge RMSE is comparable with the RMSE of the small regression and looks to be in that same echelon of predictive power
   3. Based on your answer to part a) in the previous section and to your answers in this part where we tried to use the entire data set more judiciously, do you think that expert judgment and our knowledge from other econ classes were helpful when we tried to build our predictive model here?
      1. Knowing the data did help with the selection of the actually important variables. Our knowledge worked like the LASSO/Ridge models in the sense that it eliminated variables that didn’t make sense to include