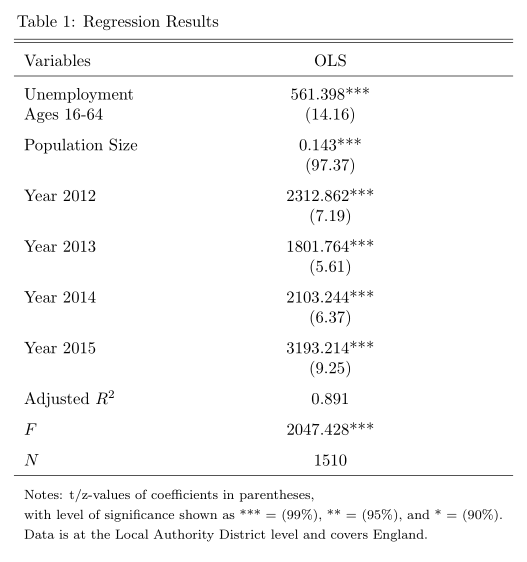
Supervised Findings

In order to directly test the hypothesis that higher levels of unemployment directly lead to higher levels of crime, we used ordinary least squares linear regression analysis using base R. We performed this analysis at the LAD level, and all variables were at the LAD level when this analysis was performed. Total number of crimes in an LAD was the dependent variable. The independent variables were the unemployment rate of people between the ages of 16 and 64, the population size of the LAD, and a dummy variable for each year in the study except 2011 (omitted to avoid collinearity).

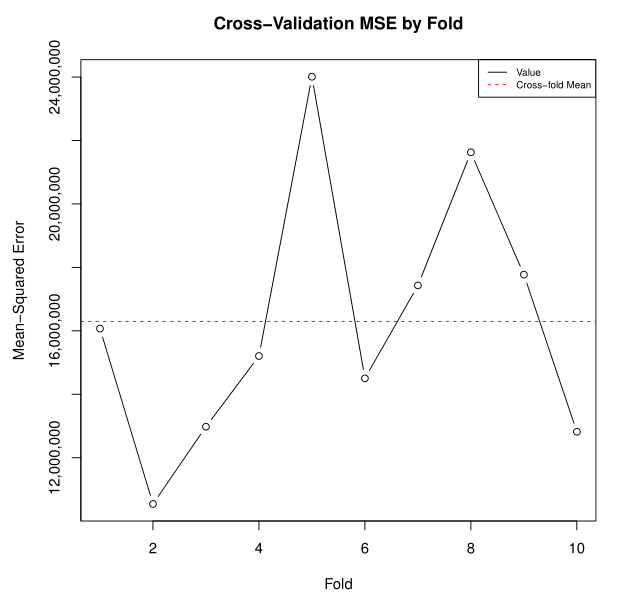
Of course while the unemployment variable was the focus of the analysis, we added other predictors to act as controls. Population size controlled for the large variation in population size (and therefore some of the variation in crime counts) in the LADs. Furthermore, we are not naïve enough to believe that the only predictor of total crime is going to be unemployment, though we do believe it makes methodological sense for it to be one of the strongest. As such, including a Year dummy variable for each year helps us to approximately control for various economic factors that were not readily available, such as yearly mean or median household income, or GDP. I say that it would approximately control for those things because in their absence, the dummy variable for each year is going to be given “credit” by the regression for variation that is actually caused by these missing macro-economic indicators changing between years. Recall that any coefficient of these year dummy variables should be interpreted as the additional change in the crime count for the given year compared to the effect of 2011.



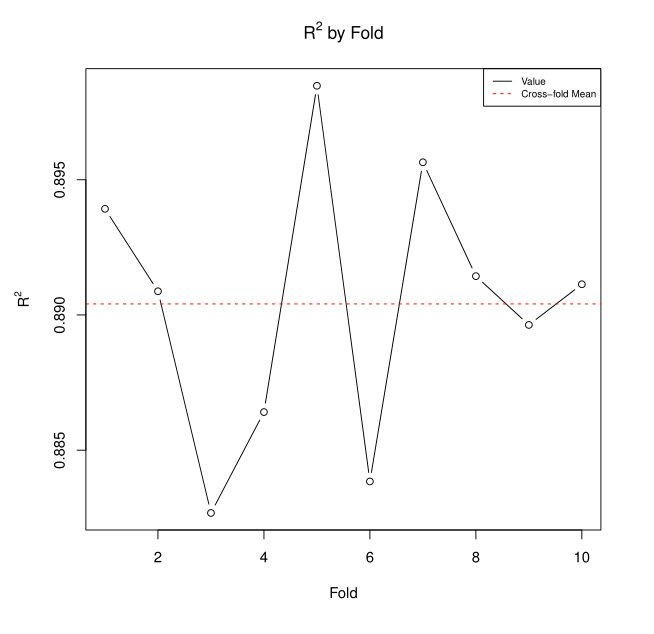
Consider Table 1. As you can see, all of the predictors were statistically significant at a 99% confidence level, and the adjusted R2 value was 0.891, indicating that these 6 predictors can, together, explain approximately 89.1% of the variation in total crimes across local authority districts. It is clear that the model has explanatory power as indicated by the F-statistic, which is far greater than 1. There were 1510 records included in the regression. That is one record per year per local authority district between 2011 and 2015. There were about 150 records that had to be removed prior to running the regression due to missing data.

In order to perform validation on our analysis, we used 10-fold cross-validation. For each fold, we tracked the mean-squared error (MSE), adjusted R2 value, and the value of the coefficients. Through the validation and analysis of our results we found that the model was fairly stable, but that there may be ways that the model could be improved in the future.

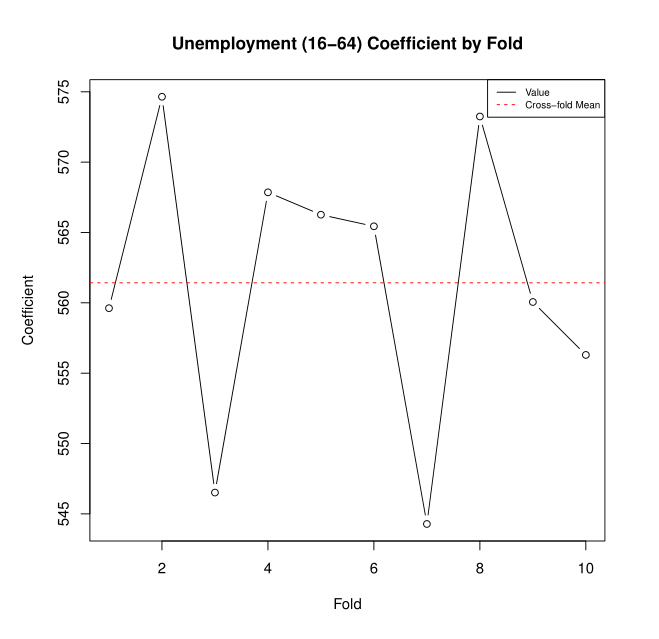
In the graph below, you can see our mean-squared error, which is pretty high, averaging just above 16 million. This indicates that, if we were to use this model to perform prediction, the prediction of crimes for any given LAD would be under or over-predicted by about 4,000 crimes. For larger LADs, that isn’t so erroneous. However, the mean value of total crime in our data is about 14,280 crimes, and the median is just under 10,000 crimes. As a result, we would recommend against using this model for prediction.



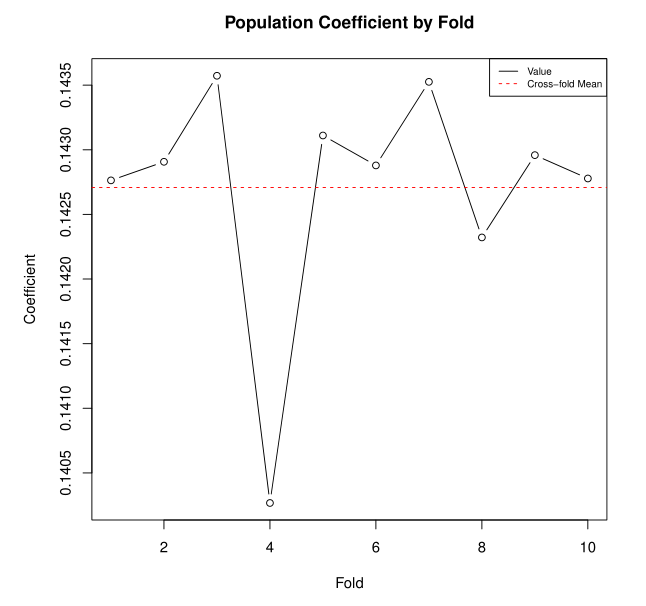
Given the graph below of the adjusted R2, you might think that such a high error rate of the model is unexpected. The adjusted R2 graph indicates that on average (across folds) we explain approximately 89% of the variation in the data. However, there is a way that this could be occurring, and that is due to omitted variable bias, which will be a theme as we continue to look at the cross-validation results and analyze this model. It is what we tried to control for with the year dummy variables.



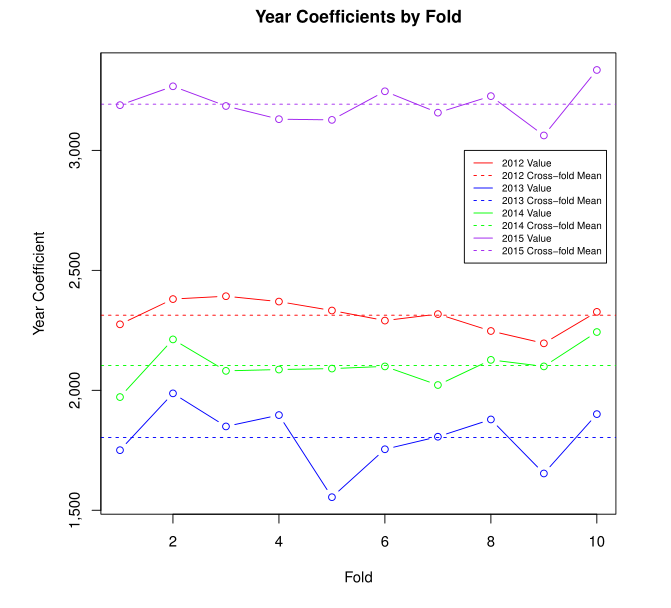
The next three graphs show the stability of the coefficient estimates across folds. Note that although the graphs look relatively variable, the scale on the y-axis is very small, so in reality we are looking at very little variation.



Note that our average coefficient for Unemployment was approximately 562, indicating that for each additional 1% of economically active 16 to 64 year olds that are unemployed, there will be, on average, 562 more crimes.



The coefficient for Population indicates that, ceteris paribus, we would expect another crime for every 7 people in the local authority district. Recall that “crimes” are loosely defined here, as “anti-social behavior” is the most common crime in our data.



Finally, we can take a look at the variance of the coefficients across years and within year across folds. Again, these values are very stable. Recall that these have to be considered in relation to 2011. So if 2011 is considered the baseline, then we could interpret the 2012 year coefficient on the first fold to mean that, on average, in 2012 there are roughly 2,300 more crimes than there were in 2011.

Unfortunately, with such high magnitudes and a non-trivial variance, since we have no methodological reason to believe that one year should be so different from the others, we can be sure that the analysis suffers from omitted variable bias. That is to say, the year variable is clearly acting as a proxy for other variables that should have been included in the regression, were they available. Given the large magnitude of these coefficients and the high variance among them, we would need to be even more careful about predicting values outside of the time range of our data than we are with predictions without our time range.