**GAME PLAN**

In the following plan, we describe (1) baseline models (2) null models (3) assessments of the predictive power (4) robustness tests for the predictive power (5) inference tests and (6) specification tests.

**BASELINE MODELS**

We’ve selected the following as our baseline models based on the performance of our experimental models.

For both MR and PS injuries, we will run both logit (binary outcome) models and negative binomial (count outcome) models. For each of these four models, we will implement a version utilizing violation variables with no lags, and violation variables with a lag of one (i.e. violations yesterday predict injuries today). We will run all of these models on all subparts, as well as part-level data. All of these models will be implemented at the year level. This results in 16 models, which we refer to as our baseline models, which assume the form:

***INJURIES*** *=* ***(COVARIATES OF INTEREST)*** *+ (MINE TIME) + (ONSITE INSPECTION HOURS) + (STATE DUMMIES) + (TIME DUMMIES) + EXPOSURE(HOURS) + CLUSTER(MINES)*

Bolded aspects of the models change across baseline models. These include:

**INJURIES:** Either MR or PS, either in binary or count form.

**COVARIATES OF INTEREST**: Either at the part of subpart level, either with no lag or with one lag. Always of the form counts of total violations.

**NULL MODELS**

For each of our baseline model specifications, we will also estimate the following models:

1. A strong null model, of the form above, except *excluding* all covariates of interest.
2. A weak null model, of the form above, but *only including* the exposure term and the clustering of standard errors.

However, only 8 null models will need to be performed in total (because no covariates of interest are included in either the strong or the weak specification).

In each of the 24 models described above (16 baseline, 8 null) we will train the model on ~70% of all observations in our dataset, and test the predictive power of models on the remaining ~30%. 69% of observations in our sample occur before 2012, and so **we include observations before 2012 in the training set, and all observations in and after 2012 in the testing (hold-out) set.** Later on, we will perform various tests to assess the appropriateness of our training and test set size and composition.

**ASSESSMENTS OF THE PREDICTIVE POWER OF OUR MODELS**

Having estimated all baseline and null models, we perform the following assessments of predictive power.

1. For all logit (binary outcome) models, we use the results of the models that were fit on the training sample (both baseline and null models) to generate predictions in the test (hold-out) set data. For each baseline and null model we then compute:
   1. The overall classification rate (# correctly predicted outcomes/total # observations)
   2. The false positive rate (# predicted positive outcomes/# of true positive outcomes)
   3. The false negative rate (# predicted negative outcomes/# of true negative outcomes)

We expect all baseline models to perform at least no worse than null models.

1. For all negative binomial (count outcome) models, we use the results of the models that were fit on the training sample (both baseline and null models) to generate predictions in the test (hold-out) set data. However, the use of count data makes the comparisons of predictions more difficult. Therefore, for each baseline and null model we then:
   1. Treat predictions and true count outcomes as binary outcomes (0’s remain 0’s, all else becomes 1’s)and then compute
      1. The overall classification rate
      2. The false positive rate
      3. The false negative rate
   2. Create various groups of real and predicted outcomes for comparison:
      1. We treats all 0’s as 0’s, all 1’s as 1’s, and all observations >1 as 2’s
         1. The overall classification rate
         2. The false positive rate
         3. The false negative rate
   3. Finally, we calculate the residuals (the difference between true and predicted outcome) for each baseline and null model.
      1. The sum of the absolute value of residuals
      2. The sum of all positive residuals
      3. The sum of all negative residuals
      4. The number of observations for which the residual is 0

We expect all baseline models to perform at least no worse than null models in terms of overall classification rate (although false positive/negative rate may not always be uniformly higher/lower, respectively), and all sums of residuals.

**ROBUSTNESS TESTS OF THE PREDICTIVE POWER OF OUR MODELS**

1. Comparing predictions to fake outcome data
   1. Randomly shuffle values of the outcome variable (i.e. sample randomly from the distribution without replacement) in the testing set and re-compute
      * 1. The overall classification rate
        2. The false positive rate
        3. The false negative rate
2. Testing the sensitivity of predictions to the size of the training/test set. In order to assess the appropriate of our selected cutoff of 2012, we re-estimate baseline and null models on various other cutoffs: 2010, 2011, 2013, 2014 and 2015.

**INFERENCE ROBUSTNESS TESTS**

1. We will apply **randomization inference technique** to confirm the validity of models.
   1. For models the yield statistically significant violation subparts, we will create new datasets, in which the corresponding subpart values have been randomized (selected random randomly from the column’s distribution, without replacement). We then re-estimate models.
   2. We will conduct this randomize then re-estimate procedure 500 times. After each iteration, we will record the estimated coefficient on our subpart(s) of interest.
   3. By repeating this procedure 500 times, we can obtain a distribution of coefficients on the variable(s) of interest that can be compared to the coefficient estimates from the true models. If this procedure yields a high number of coefficients (more than 0.05) with effect sizes as large as the true observed effects, it would cast doubt on the validity of our results.
2. Using **randomization to test for statistical artifacts**. Given the size of our data, there is concern that any observed effects may be due to statistical artifacts. For example, violations of part 75 of the Code of Federal Regulations Title 30 are by far the most common violations we observe in our dataset, and so it is possible that these covariates are coming as significant simply because they have higher variance. To test for this, we follow this procedure:
   1. Create a new dataset of pseudo-violations (for every violation subpart in our dataset we replace all observations with values randomly sampled from the distribution without replacement). Re-estimate models. Repeat this procedure 500 times.
   2. Obtain a distribution of p-values on all coefficients.
   3. We expect ~0.05 of these p-values to be significant by statistical artifact alone. However, a higher proportion of p-values less than 0.05 would suggest that the significant findings in our true models may be due to artifact.

**SPECIFICATION TESTS**

We estimate all baseline models with addition of two covariates: a **longwall** indicator and a **union** indicator. The inclusion of the union indicator requires us to drop all observations after 2013.

**FUTURE TESTS (ON THE BACKBURNER)**

If time allows, we will also experiment with creating training/testing sets based on the panel dimension of our data (as opposed to time). To do so, we would create a list of mineIDs that reflect ~70% of observations in our dataset and assign these to the training set, and the remaining observations to the test set, and then re-estimate models. In our to test the sensitivity of our results to the mine composition of our training and test set, we would need to conduct this procedure 500 times, each time re-sampling the mineIDs that would be selected for the training/testing sets.