**NISOH GAME PLAN**

In this proposal, we describe:

1. The structure of the dataset that we will use for rigorous analyses;
2. The model specifications that we will evaluate with regard to predictive and inferential performance;
3. The null models against which we will compare our preferred models with regard to predictive performance;
4. The inference and prediction procedures;
5. The tests for the predictive performance of our preferred models;
6. The tests for the robustness of the predictive performance of our preferred models;
7. The tests for the reliability of the inferential performance of our preferred models; and
8. The procedures we will explore and implement if we have more time.

**(1) Dataset**

Our experimental analyses revealed that the dataset at the quarter-level was too large and sparse for frequentist modeling. Therefore, we collapse the dataset to the year-level by sums for years in which we have four quarters of data.

We split our dataset into a training and test set based on time. All observations representing mine-years before 2012 compose the training set. The remaining observations compose the test set. This split roughly corresponds to splitting the data 75%/25% (training/testing).

**(2) Preferred Model Specifications**

Each ***preferred model*** that we will evaluate in these analyses is specified in the following form:

Left-Hand-Side:

* **Injuries**
  + Can be maintenance and repair (MR) *or* pinning and striking (PS) injuries
  + Can be binary injuries *or* count of injuries; in the case of binary injuries, we specify a logit model, and in the case of count injuries, we specify a negative binomial model (our data is overdispersed)

Right-Hand-Side:

* **Variables of interest (i.e., counts of violations)**
  + Can be counts of violations at the part *or* the subpart level of the violation code
  + Can be counts of violations with no lag *or* with 1 lag (i.e., violations from the previous mine-year are used to predict the current mine-year)
* Onsite inspection hours
* State dummies
* Time dummies (i.e., fixed time effects)
* Mine time (i.e., number of active quarters the mine has been open since the beginning of our dataset)

For every model, we include an exposure variable (hours worked), and we cluster standard errors by mine.

Given the four ways in which the bolded aspects of these models are allowed to vary, we are left with **16** preferred models for our analyses.

Note: in our exploratory analyses, we investigated additional variations in specification. Below is a list of these variations with justifications for not including them in our preferred models:

* Use counts of significant and substantial violations (vs. count of violations) as variables of interest
  + We observed no meaningful differences when varying this dimension of specification
* Use penalty points (vs. count of violations) as variables of interest
  + Because the manner in which penalty points are assessed has changed over time, the use of penalty points as the variable of interest would require dropping all data before 2007, presenting meaningful sample size issues
* Use cumulative violations over the previous 4 time periods (vs. violations with no lag or 1 lag) as variables of interest
  + We observed no meaningful differences when varying this dimension of specification
* Use cumulative violations since the mine’s entry into our dataset (vs. violations with no lag or 1 lag) as variables of interest
  + We observed no meaningful differences when varying this dimension of specification

**(3) Null Model Specifications**

We propose the creation of strong and weak ***null models***, both of which we will compare against our preferred models regarding predictive performance.

***Strong null models*** are specified in the following form:

Left-Hand-Side:

* **Injuries**
  + Can be maintenance and repair (MR) *or* pinning and striking (PS) injuries
  + Can be binary injuries *or* count of injuries; in the case of binary injuries, we specify a logit model, and in the case of count injuries, we specify a negative binomial model (our data is overdispersed)

Right-Hand-Side:

* Onsite inspection hours
* State dummies
* Time dummies (i.e., fixed time effects)
* Mine time (i.e., number of active quarters the mine has been open since the beginning of our dataset)

For these models, we include an exposure variable (hours worked), and we cluster standard errors by mine.

***Weak null models*** are specified in the following form:

Left-Hand-Side:

* **Injuries**
  + Can be maintenance and repair (MR) *or* pinning and striking (PS) injuries
  + Can be binary injuries *or* count of injuries; in the case of binary injuries, we specify a logit model, and in the case of count injuries, we specify a negative binomial model (our data is overdispersed)

Right-Hand-Side: *None*

For these models, we include an exposure variable (hours worked), and we cluster standard errors by mine.

In all, we will run **8** null models – each of which will correspond (and be compared to) four of the preferred models with regard to predictive performance.

**(4) Training/Testing and Inference Procedure**

For the 24 models described above (16 preferred models, 4 strong null models, and 4 weak null models), we train the model on data in the training set. When probing a binary outcome, we specify a logit model; when probing a count outcome, we specify a negative binomial model (our data is overdispersed). We then use these trained models to predict the outcome of the test set – a procedure we refer to as ***prediction***. We also examine the discovered associations between the variables of interest (violations) and the outcome (injuries) in the training set alone – a procedure we refer to as ***inference***. For both procedures, we assess the performance of our models and propose numerous robustness checks.

**(5) Assessing Predictive Performance**

After training all preferred and null models and using the results to generate predictions in the test set, we perform the following assessments of predictive performance:

For all ***logit*** models (probing a binary outcome), compute:

* The overall classification rate (# of correctly predicted outcomes / # of true outcomes)
* The false positive rate (# of falsely predicted positive outcomes / # of true positive outcomes)
* The false negative rate (# of falsely predicted negative outcomes / # of true negative outcomes)

Then calculate differences between these measures for the preferred models and corresponding weak and strong null models. Record whether and by how much the preferred models outperform each of the null models.

For all ***negative binomial*** models (probing a count outcome):

* Convert predictions and true outcomes to be binary (if x == 0, y = 0; else y = 1), then compute:
  + The overall classification rate (# of correctly predicted outcomes / # of true outcomes)
  + The false positive rate (# of falsely predicted positive outcomes / # of true positive outcomes)
  + The false negative rate (# of falsely predicted negative outcomes / # of true negative outcomes)
* Group predictions and true outcomes (if x == 0, y = 0; if x == 1, y = 1; else y = 2), then compute:
  + The overall classification rate (# of correctly predicted outcomes / # of true outcomes)
  + The false positive rate (# of falsely predicted positive outcomes / # of true positive outcomes)
  + The false negative rate (# of falsely predicted negative outcomes / # of true negative outcomes)
* Calculate “residuals” (i.e.., true outcome – predicted outcome), then compute:
  + Sum of the absolute value of residuals
  + Sum of positive residuals
  + Sum of absolute value of negative residuals
  + Number of residuals == 0

Then calculate differences between these measures for the preferred models and the corresponding weak and strong null models. Record whether and by how much the preferred models outperform each of the null models.

**(6) Robustness Checks on Predictive Performance**

We endeavor to answer the following questions:

*Null hypothesis*: Predicted injuries are independent of violations

*Test*:

1. Randomly shuffle values of the outcome variables in the test set (i.e., sample randomly without replacement from the distribution)
2. Predict outcomes using trained models (preferred and strong null only – our preferred model should always perform better than the weak null)
3. Re-assess predictive performance of preferred models, following section (5)
4. Repeat 500 times

*Interpret*: Under the null hypothesis, the preferred models should not outperform the null models as often or as much with randomized data as with the true data

*Null hypothesis*: The predictive performance of the preferred models is dependent/sensitive to the size of the training/test sets

*Test*:

1. Re-estimate preferred and null models using training/testing cutoff years: 2010, 2011, 2013, 2014
2. Re-assess predictive performance of preferred models, following section (5)

*Interpret*: Under the null hypothesis, the preferred models should not outperform the null models as often or as much with differently sized training/test data as with the chosen cutoff

**(6) Robustness Checks on Inferential Performance**

We endeavor to answer the following questions:

*Null hypothesis*: Injuries are independent of violations

*Test*: Randomization inference, as in Ho & Donahue

1. For models the yield statistically significant variables of interest (part or subpart level violations), for each variable of interest, create a new dataset in which the variable of interest is randomized (i.e., sampled randomly without replacement from the distribution)
2. Re-estimate preferred and null models using randomized data; record the estimated coefficient on the variable of interest
3. Repeat 500 times
4. Obtain a distribution of coefficients and calculate p-value to test the null hypothesis (see Ho & Donahue)

*Interpret*: Under the null hypothesis, the randomized data should yield a high number of coefficients with magnitude as large as the observed effect in our preferred model

*Null hypothesis*: Significant violations are statistical artifacts of the sparse data

(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 variables are coming as significant simply because they have higher variance.)

*Test*:

1. For models the yield statistically significant variables of interest (part or subpart level violations), for each variable of interest, create a new dataset in which the variable of interest is randomized (i.e., sampled randomly without replacement from the distribution)
2. Re-estimate preferred and null models using randomized data; record the estimated p-value on the variable of interest
3. Repeat 500 times
4. Obtain a distribution of p-values

*Interpret*: Under the null hypothesis, the randomized data should yield a high number (> 5%) of significant p-values for the variable of interest

**(7) Future Tests**

If time allows, we will also experiment with:

* Specification tests: We estimate all preferred and null models with the addition of two covariates: a longwall indicator and a union indicator. The inclusion of the longwall indicator requires us to drop all observations representing mine-quarters after 2015, and the union indicator requires us to drop all observations representing mine-quarters after 2013.
* Creating training/testing sets based on the panel dimension of our data (as opposed to based on time), and replicate all analyses: It would be important to test the sensitivity of these results to the size and mine composition of the training and test set.