**NIOSH Data Analysis Strategy**

***“Statistical significance”***

“Which, if any, equipment-related violations are statistically significant predictors of subsequent equipment-related injuries?”

Performing frequentist hypothesis testing will fulfil this requirement.

Bayesian analysis with Bayes factors will not produce p-values or traditionally-understood measures of statistical significance. We can measure confidence in our results, but it will not technically mean “statistical significance.” Is this kind of analysis still worth pursuing?

For general background on the interpretation of Bayesian measures of significance (as well as for a discussion of the similarities/differences between these and frequentist approaches), see, e.g.:

<https://www.jstor.org/stable/pdf/2669779.pdf>

<http://truebra.in/?p=690>

<http://bayesfactor.blogspot.com/2015/01/on-verbal-categories-for-interpretation.html>

***Our intuition(s)***

1. We suspect that what may be predictive is not only certain/groups of violations, but also patterns of certain/groups of violations over time. To test this hypothesis, we would have to use lagged variables.

However, if we lagged all of the potential predictors (4 quarters back), we lose many observations (approx. 20%); we (potentially) reach a place with more predictors than observations; and we have to deal with more severe collinearity issues.

To read more about the questions lagged variables answer, as well as the danger of this procedure, see, e.g.:

<http://www.mathworks.com/help/econ/examples/time-series-regression-viii-lagged-variables-and-estimator-bias.html>

1. We suspect that interaction terms between parts and sub-parts may also be predictive.

However, including interaction terms increases the number of predictors in our model, which means that we (potentially) reach a place with more predictors than observations and we have to deal with more severe collinearity issues.

***Complicating features of the data***

* Sparse
* High-dimensional
* Panel dependence (potentially 2 layers)
* Time trends (trend, seasonal, fixed effects)
* Nested/collinear predictors

***Investigated solutions (Frequentist)***

* Sparse
  + **Data imputation** - Not theoretically sound for our purposes
  + **First differences** - Not theoretically sound for our purposes
* High-dimensional
  + **PCA** - Produces un-interpretable results for our purposes
  + **Low variance filter** - Ineffective because of sparsity
  + **High correlation filter** - Only catches linear relationships; also hard to choose cutoff because of high collinearity of predictors
  + **Variable selection + inference** - Possible, but we lose data for both tasks; the procedure may be sensitive to outlier mines; the procedure is still constrained by collinearity issues; and robustness checks are not straight-forward
    - **Random Forest** - Possible tool for variable selection, but still subject to pitfalls of variable selection + inference method (stated above)
* Panel dependence (potentially 2 layers)
  + **Fixed effects** - Possible, but includes more predictors in the model and still requires work to define the rest of the model (for which we are somewhat lost)
  + **Hierarchical models** - Useful/necessary for dealing with multiple levels of panel dependence, but not applicable otherwise (decide whether we care about higher-order dependencies than mines like operators); also cannot use lagged variables
* Time trends (trend, seasonal, fixed effects)
  + **Time series regression** - Must be combined with other methods
  + **Time trend, seasonality, and fixed effects** - We should include these effects; this includes more predictors in the model and may result in further collinearity issues
* Nested/collinear predictors
  + **Nested model comparison with F-test** - Possible, but very computationally expensive and non-traditional (also we still need to define the models, and we are still somewhat lost here)

***Potential contacts***

* Statistics
  + Montanari – high dimensional statistics
  + Joseph P. Romano – econometrics, multiple hypothesis testing, large sample theory, time series analysis
  + David O. Siegmund – probability theory, nonlinear regression
  + Wing Hung Wong – multivariate analysis, statistical inference, machine learning
  + John Duchi – machine learning, algorithms
  + Persi Diaconis – machine learning, computing
  + Susan Holmes – computer intensive methods, phylogenic trees, multivariate analysis
  + Tze Leung Lai – empirical Bayesian modeling, quantitative finance, time series
* Economics
  + Timothy Bresnahan – econometrics
  + Han Hong – econometrics
  + Thomas E. MaCurdy – econometrics