22 June 2021

**Summer Outline for OSHA dataset analysis**

**Question of interest:**

1. Determine how serious the violation type was based on the penalty fees applied to the facilities?
2. Does the involvement of union status increase or decreases the case duration to resolve?
3. What is the relationship between key toxicity variables and owner type? Is the relationship strong or weak?
4. How long does it take for a union to resolve the issue?

**Suggested methods:**

1. Apply logistic regression model for a categorical and numeric variable.

The variables to include are viol\_type (categ), cur\_penalty (num), init\_penalty (num), and nr\_exposed (categ)

1. Probably descriptive stats like table and a graph to show.

The variables to include are union\_status (categ), nr\_instances (categ), insp\_type (categ), and insp\_scope (categ)

1. Apply logistic regression model for a categorical and numeric variable.

The variables to include are Owner\_type(categ), viol\_type (categ), cur\_penalty (num), init\_penalty (num), insp\_type(categ), insp\_scope(categ), and nr\_exposed (categ)

1. Apply graphs:

The variables to include are union\_status (categ), open\_date (time), close\_date (time), nr\_instances (categ), insp\_type (categ), and insp\_scope (categ)

I like that you list the variables! I think one thing that would help me is to list which is the dependent variable and which are the independents. From what I read I see a few proposed regression models.

Linear regressions:

DV: current penalty (or initial penalty)

IVs: ??? Up to you. What might predict the penalty amount? Gravity should but a lot is missing. Union\_status, owner\_type, nr\_exposed (this is a total we could make an average nr exposed if you prefer), nr\_violation, inspection\_type, viol\_type, and time open - could all be good!

DV: average number of days the inspection was open

IVs: ??? Up to you. What might predict longer duration? You mentioned union status. Owner\_type, nr\_exposed, nr\_violation, inspection\_type, inspection\_scope might make sense but could be collinear with nr\_violations, viol\_type, gravity (but a lot missing), initial penalty (this would be super interesting bc perhaps greater initial penalty means the case is open longer because they might try to contest having to pay that much).

Logistic:

DV: Viol\_type (categorical) - the categories are other than serious, serious, and repeat. So in the “long” data it would be easy to do logistic regression because you might think about adding “serious” and “repeat” together into one variable. That way you have 0 - other than serious and 1 - serious or repeat. Because repeat violations are, in a way, serious, because they had an issue once and then they had the same issue again later so they didn’t really solve the problem!

However, the “wide” data are a little tricky. In the wide data instead of other than serious or serious/repeat we have total number of times they had an other than serious violation, total number of times they had a repeat violation, and total number of times they had a serious violation. So we have three separate variables.

So you have a few options: 1) You could do another linear regression predicting total other than serious violations. 2) You do another linear regression model predicting total serious (or repeat) violations. 3) You could re-code the data and create a binary variable 0 - if they never had any serious violations and 1 if they ever had a serious violation. I don’t think this would be too hard to do and would allow you to meet your goal to do a logistic regression.

IVS: ??? This is up to you. What might predict ever having a serious (or not serious) violation? My suggestions might be Union\_status, owner\_type, average number of people in the facility, nr\_instances or number of times inspected/cited. I’m not sure the other variables make much sense.