Project Notes/Log

Gift and Kunal

Questions for cct

* Model selection for logistic regressions – k-fold cross validation?
* How to interpret negative natural log model estimates?
  + When I take the log, I add it to the whole real line
  + When they are exponentiated, it becomes a value between 0 and 1, which then describes the the %13 shrink.
  + Scale my contiuous variables: (x-mu)/sd
  + Exp(intercept) – positive number, the [As] at the mean sampling distance
  + Exp(coefficient) – positive number, proportional change in As per 1 km distance from smelter, starting at the average distance
  + Predict functions should do it okay, but should work – plz test and report back
  + If I am not going to predict something, I do not need machine learning, so I don’t need bootstrapping or k-fold
* How to handle overlap in independent variables, lmm

To Dos 11/14/23

* Email gift hds spreadsheet - kunal
* Email stats class for more info on sparsepca - kunal
* Email monica for meeting – gift
* HDS discrepancies – kunal
* Logistic regressions by community – kunal
* Natural log transformed cfactor models by community - gift

To Dos 11/9/23

* Cfactor model test and summaries with natural log - gift
* Log transform distribution comparison for cfactors - kunal
* PCA – kunal
* Logistic regressions – kunal
* HDS Index – maintainence - kunal
  + 0 and 1
  + If missing, assume 0 for risk averse public health analysis
    - Q67, Q71, Q76, Q77, Q79,

To Dos 11/2/23

* PCA – Kunal
  + community
* PLI Summary – Gift
* Cfactor Summary and Model – Gift
  + community
* Exceedance logistic regression – Kunal
  + By community? Fine if not
* exceedance summary – Kunal
* Concentration summaries – Gift
* HDS index – Kunal and Gift
* Cfactor modeling w/ HDS – Kunal and Gift

Meeting with Stats Class

* How should we handle missing/unanswered categorical data from the home description survey? When combining this data with contaminant data, our sample size drastically diminishes because not all samples have home description data associated with them.
* We are interested in looking at 13 analytes, 10+ variables, and many interactions. What is the best way to pick which analytes and variables are most influential?
* How do we handle collinear variables in our models?
* PCA
  + Looks at covariance with each other, sometimes we can find info on sources
    - Source signals would get mixed up
    - Do separate MFAs by community by season? But potentially have an overfitting issue
  + Sparse PCA algorithms? For ex. Lasso regression
    - Adds a penalty to the optimization of modeling – this helps overfitting and can favor solutions where components are zeroed out (variable selection) – most important variables in terms of error
      * ## Sparse PCA
      * 1. N.B.Erichson,P.Zheng,K.Manohar,S.Brunton,J.N.Kutz,A.Y.Aravkin."Sparse
      * Principal Component Analysis via Variable Projection." Submitted to IEEE
      * Journal of Selected Topics on Signal Processing(2018).
      * (available at <E2><80><98>arXiv https://arxiv.org/abs/ 1804.00341).
      * Tuning parameters – shrinkage + robustness – chosen by cross validation, prediction score
    - Definitely important to look at outliers, scale the data, play around with the sparseness of the solutions (how many zero coefficients)
* Different data ranges
  + Use logs or natural logs
  + Model the correlation matrix (use scale = true or something like that, scale all the data, using a z-score) – still need to look at outliers
* Non-detects
  + Around 70% non-detect, probably don’t use the data
* HDS
  + Missingness mechanisms
    - If data is missing for some reason that doesn’t inform the answer, as opposed to not answering for a specific reason (missing vs informed missing)
    - Sometimes code missing as a variable – if missing is significant….
    - Use covariates and do imputation of the missing values so we aren’t forced to using complete cases, are there associations between missing data and data we do have?
  + Making an index? A composite predictor? If they’re all going in the same direction, making a maintainence score – missing values???
* General notes, do all the modeling by community

Next steps

* What is the PCA goal?
* Research questions
* Give summary stats for contaminant data and HDS
* Metadata

Research Questions

* Is it safe for gardening?
* Is infrastructure or point source contributing more? If so, how? Why? What can we do about it?
  + Where do we put our energy?
* Not too interested in differences between contaminants.
* How interested are we in comparing mining and urban
* We are trying to make a generalizable conclusion, studying individual community dynamics is not as useful

Priority action Items

* Continue testing pli models (KP)
* How to handle missing data (KP)
* Cf models (Gift)
* Play around with multiple metals in a model, starting with two or three (KP and GC)
* Figure out how to represent, understand, communicate complex interactions (GC)
* Add pH and EC into the models
* Figure out how to look at multiple rwhi questions simultaneously – maybe just compare p-values
* Look at dimensional reduction methods – PCA to figure out which metals to focus on? (KP)

Regular action items

* MFA with hds info?
* Add pH and EC into all analysis

Double check the following

* Al summary stats, some values might actually be lead stats
  + Why was period:season taken out of the model? It was a better fit, but then period was visualized later?
* Do the control samples fall into PH sampling scheme? Do we exclude control samples that were collected after our project ended?

List of assumptions, data notes

Control Modeling options

* Put all the metals together?
  + Would the different metal scales fuck up the model interactions?
    - Metal:season and metal:community and metal:period and metal:period:season?
* Model PLI -> model cfactor comparing metals, with signif variables from pli -> individual concentration models with highest contributors -> HDS models
  + 1) PLI – narrow variables down, generalized pollution analysis
  + 2) Cfactor – assess contribution only
  + 3) Concentration – highest contributor analyte trends
  + 4) Concentration + HDS - highest contributor analyte trends
    - Do we need pli models for hds?
    - Can we add hds questions together? How do we handle missing hds data?
* IF we can analyze multiple HDS ?s in one PLI model, repeat above analysis
  + Model PLI -> model cfactor comparing metals, with signif variables from pli -> individual concentration models with highest contributors -> HDS models
    - 1) PLI + HDS – narrow variables down, generalized pollution analysis
    - 2) Cfactor + HDS – assess contribution only
    - 3) Concentration + HDS – highest contributor analyte trends
* How do we graph complex interactions?

HDS modeling

* How do we rank different hds questions?
  + Can we compare p-values to each other?
* Do we look at multiple metals at once for one hds question at a time?
* Add distance to point source to hds models? Which is more important?
* How do we get creative with HDS? Can we create an index?
  + For ex every home before 1960 is a minus 1, first flush is a plus 1, etc etc
  + Normalize for number of responses
  + But some questions are really hard – 1s and 0s and -1s could be dependent on analyte
  + Only variables that people realistically have control over?
  + Conservative, risk averse approach
    - Wherever there is a NA or missing value, we assume they are not following the best practice

Pull back a little bit

* Take top 5 average cfactor analytes by community, by season
  + Ranked differently, but perhaps we see trends? Cu in all the mining communities for example
  + And then go to infrastructure
* We look at PLI split by community or land use

Questions

* How to handle missing categorical data from home description survey? In order to attach to contaminant data.
  + Do we create an index of decision making? How? How to normalize? Assume NAs are 0s? Bootstrapping and imputation?
  + How do we assess home age and roof type then?
* Advice on the best way to pick metals based on contamination?
* How should we approach collinearity in the modeling?
* Do we model each community individually? Combine analytes?
* Do we need to use cfactor or pli for individual analyte analysis? Or concentrations?
* Do we need enrichment factors to help us understand pollution?