**Code Review: PM2.5–ALS Project (with Yanelli)**

This is a code review document for the PM2.5–ALS Project.

I’ve included all comments in (more or less) sequential order based on the README file that you provided. There are four sections (Packages, Data Prep, Models, and Figures), and each are divided into subsections as necessary based on the individual files.

Comments are generally provided in bullet point format. Whenever possible, line numbers relevant to my comments are also referenced.

**Packages**

“packages.R” file only, which calls all required packages. No additional comments.

**Data Prep**

*p\_00\_patient\_data\_prepping*

L95: sorry that you had to do this by hand…I assume there wasn’t a more effective way to filter this?

*p\_01\_combining\_poll\_hosp\_data*

General comment: did you clean/match the air pollution data, or was this also provided directly by Johnni? I guess I’m just wondering how the air pollution was initially matched to each individual in the first place.

*p\_02\_windows\_exposure*

L55: slightly confused about how the nesting works: ask Yanelli. What is ck?

General comment: missingness % is small, and patterns appear to be random; does not seem to be much of an issue.

Section 5 (Data selection): I’m having a bit of trouble understanding how the exposure is matched here…how do you end up with so many rows of data after this is matched? (Ask Yanelli to walk through one of these)

Note to self: basically (cases + control) \* number of months ~ total number of rows

*p\_03\_averaging\_exposures*

Note to self: exposure data is monthly

General comment: by the end of this, you essentially end up with just (cases + controls) # of rows again, correct? But this differs slightly for each exposure window since there’s slightly different bits of missingness for each.

*p\_04\_lagged\_exposures*

L117-119: can you please clarify how the lagged exposures are assigned?

General comment: if I understand this correctly, you create the lag variables and exposure corresponding to each lag variable in separate datasets, then you piece this together (“rowbind”) all into one data frame, with an indicator for the lag year?

*p\_05\_add\_parish\_ses*

No comments.

**Models**

*functions\_yanelli*

Some potential questions about extracting the lag/cumulative lag estimates, but will wait until I see the model code.

*m\_00\_load\_datasets*

No comments.

*m\_01\_pm2.5\_fix\_exposures*

No comments.

*m\_02\_pm2.5\_exposures\_nonlinearities*

General question: how come you went straight for the natural spline to test non-linearities? Did you use penalized splines first? All the splines also came out to be non-significant, correct? AIC results make sense.

*m\_03\_pm2.5\_lag\_analysis*

L61: Okay! This makes sense now :) (converting lag back to wide format)

L82: why did you create the matrix by hand? I thought crossbasis generated the matrix for you when you input this directly…actually, is this because your data is a cohort? This works for a time series (and makes sense), but since your data is at the individual level, I guess it wouldn’t work…

L109: did you try any other df’s beyond 3 and 4? My experience in the past is that 3 always ends up being the lowest AIC, but overly may overly constrain the model to not depict any non-linearities. Though your results for 4 df does not appear to be much different from that of 3 df…so it is probably okay.

L307: these results are interesting…because the numbers are so borderline significant (especially for lower lags), to me it’s really hard to distinguish whether the difference in results is a meaningful difference or a statistical fragment…would love to talk about this.

**Figures**

*f\_01\_tables*

The package used to make tables look great!

No comments otherwise.

*f\_02\_figures*

No comments.