Module 3: Pooling air pollution health risks across locations and understanding spatial heterogeneity

In this module you will learn how to pool health risks estimates across locations to produce a national average estimate of the health effects of air pollution. You will also learn how to estimate the spatial heterogeneity of the county-specific health risks estimates, and how to properly rank the estimates.

Data:

1. Maximum likelihood (MLE) county-specific estimates of the relative rates of hospitalization for all the outcomes, their corresponding statistical variance, their regional indicators (7 regions and the east and west US) for the 204 counties included in Dominici et al JAMA 2006.*(Use the lag that leads to a statistically significant effect of the national average estimate, see Table 1 of the JAMA paper).* Add few county-specific characteristics, SES (average income or percentage people in poverty) and average O_3, NO_2).

Questions to consider (Part I)

1. Is there evidence that short-term variations in $PM_{2.5}$ are associated with hospitalization risk for cerebrovascular disease on average across the nation?

2. Is this evidence the same across geographical regions and between the East and West U.S.? Is there evidence of heterogeneity across counties and geographical regions of the health effects of air pollution?

3. Are counties with the largest health risks for $PM_{2.5}$ also the ones with largest average O_3 and NO_2? Can we identify county-specific characteristics that might explain why in some counties have larger health risks than others? County-specific characteristics can be found in the `countyinfo.rda` file.

4. Are the relative risks estimates similar across cities? Is it reasonable to pool these estimates across geographical locations to provide a national estimate of risks? If yes why? If no, why?

5. Should we estimate health relative risks in Los Angeles by using data from Los Angeles only or should we also use data available in the other counties?

6. What are the three cleanest counties in the US (that is, that have the smallest health risks associated with air pollution)? Should we move there? What are the three dirties counties in the US? If you live in one of them, should you move out?

Part I: estimating a national average and heterogeneity

Tasks (Part I)

1. Load the `MCAPS.rda` dataset. Plot the county-specific estimates of the relative rates for cerebrovascular disease and their 95% confidence intervals using the function `plot.estimates`.
2. Try plotting the same data but ranking the estimates from the largest to the smallest. You can use the `sort = TRUE` option to `plot.estimates` for this.
3. Apply the function `pooling` to estimate the national average and the heterogeneity parameter by using the methods of moment. Interpret these two estimates.
4. Apply the function `lm` to estimate an average effect for each of the seven geographical regions
5. Appy the function `pooling.tlnise` to estimate the national average and the heterogeneity parameter by using Bayesian computation methods. Compare these estimates with the ones obtained in 3.
6. Do the same as 5 but separately to the counties located in the East and West U.S.
7. Apply the function `pooling.tlnise` to all the outcomes and produce national average estimates and their 95% confidence intervals. Do you obtain the same results are the ones summarized in the Table 1 of the paper Dominici et al JAMA 2006?

Part II: estimating county-specific risks by borrowing strength across locations and explaining heterogeneity

Tasks (Part II)

1. Apply the function `pooling.tlnise` to the county-specific estimates (MLE) for cerebrovascular disease and obtain the Bayesian county-specific estimates (BE) and their 95% posterior intervals.
2. Apply the function `plot.estimates` to plot side by side the MLE and the Bayesian estimates with their 95% uncertainty intervals versus county. You can use the `shrink = TRUE` option to plot the Bayesian estimates.
3. Rank the BE and their posterior intervals from the largest to the smallest estimate. Trying to plot the ranked estimates and their 95% posterior intervals.
4. Plot the MLE county-specific estimates versus the county-specific characteristics. County-specific characteristics can be found in the `countyinfo.rda` file.
5. Fit a weighted linear regression model having as dependent variable the county-specific MLE estimates and as independent variable each of the county-specific covariate. Weight the observation by 1 divided by the statistical variance of the MLEs.