**Module 2: Building statistical models for air pollution and health**

In this module we will learn how to estimate the risk of mortality associated with short-term exposure to air pollution. We will provide the definition of "confounder" and we will estimate the mortality risk "adjusted for confounding" by temperature and season. Finally, we will estimate the association between air pollution and mortality by using linear regression models and log-linear regression models.

In order to remove the influences of potential confounders, we need to compare mortality when air pollution is higher to otherwise similar days where air pollution is lower. A key issue is how to define "similar". Two possibilities include to matching days based on the season or on temperature. For example, if we only look at days where the temperature is the same, then temperature cannot confound the relationship between air pollution and mortality. Similarly, if we only look at days that fall in the same season, then season cannot confound the relationship. We will explore these methods of handling potential confounders in this part.

Questions to consider

1. What variables/factors might confound the relationship between air pollution and mortality?
2. What happens to the $PM_{10}$-mortality relationship when we try to "adjust" for season and temperature? Is the relationship consistent across the various strata?
3. Are the estimates of the association between $PM_{10}$ and mortality the same in each city? Why might they be different?

**Part 1: Handling confounding factors**

Confounding by Season

1. Use the `lm()` function to fit a linear model of all-cause non-accidental mortality (`death`) versus $PM_{10}$ exposure at lag 1 (`l1pm10`). Do this for Chicago (`chic`), New York (`ny`), and Los Angeles (`la`).
2. For a given city's data frame, subset the data frame and create 4 separate data frames, one for each season/quarter of the year. To create a "winter" data frame for Chicago, you can do

   ```
   load("chic.rda")
   winter <- subset(chic, quarters(date) == "Q1")
   ```

   Data frames for the other 3 seasons/quarters can be constructed in a similar fashion.
3. Fit 4 separate linear models of non-accidental mortality versus lag 1 $PM_{10}$ using the 4 season-specific data frames.

Confounding by Temperature

1. Fit a simple linear model of non-accidental mortality and temperature (`tmpd`).
2. For each city, divide the temperature range into 3 categories: cold (`tmpd` $< 50$ degrees), warm ($50 \leq$ `tmpd` $< 80$), hot (`tmpd` $\geq 80$). Create 3 new data frames by subsetting the city's data frame into separate cold, warm, and hot data frames. (Note: You might want to think up some other definitions of cold, warm and hot.)
3. Fit 3 separate linear models of `death` $\sim$ `tmpd`, one for each temperature range.
4. Fit 3 separate linear models of `death` $\sim$ `l1pm10`, one for each temperature range.

Confounding by Season and Temperature

1. Divide the data into $3 \times 4 = 12$ different temperature-by-season data frames and fit a linear model of `death` $\sim$ `l1pm10` within each stratum.
2. Fill in the regression coefficient for `l1pm10` associated with each pairwise combination in the following table for each city (Chicago, New York, LA):

| | Winter | Spring | Summer | Fall |
|---|---|---|---|---|
| Cold | | | | |
| Warm | | | | |
| Hot | | | | |

**Part 2: Multiple linear regression**

In the previous section, some of the season $\times$ temperature categories have no data in them. This is a generaly problem when you start looking at a lot of variables at once. One way around this is the use multiple linear regression. With multiple linear regression we assume that the the variables have a linear relationship with the response variable.

Putting it all together

1. Use the `lm()` function to fit a linear model of non-accidental mortality and lag 1 $PM_{10}$ (`death` $\sim$ `l1pm10`) for Chicago, New York, and Los Angeles.
2. Create 2 new categorical variables (factors), one named `season` corresponding to the 4 seasons/quarters and one named `temp` corresponding to 3 temperature ranges (cold, warm, hot).
3. See what happens to the coefficient for `l1pm10` when you add `season` and `temp` to the model.
4. Try adding other variables to the model and see how the log relative risk for `l1pm10` changes.