Problem Set I

Your Name

This problem set is due at 8:30 am on 9/12

Please upload both Rmd and pdf files on Sakai

In this problem set, we analyze a subsample of Professor Malesky's dataset that contains station-level air quality data of Beijing, China. Here is the variable description:

- `year`: year of the observation
- `month`: month of the observation
- `day`: day of the observation
- `code_station`: code of the monitor station
- `carbon`: carbon dioxide emission
- `nitro`: nitrogen record emission
- `lucky`: TO BE ADDED

1. Load the `CEOSAL2.dta`
2. Provide a summary of the dataset, you could use `summary`, `dim` etc. Also, answer the following questions:
    - what is the number of observations?
    - what is the time frame
    - what is the variable type of `year`,`code_station`, and `carbon` (categorical, numeric, ordinal)
3. Compute the following statistic manually (you are only allowed to use basic functions like `sum`, `length`, etc.). Then compare the results with the built-in functions (e.g. `mean()`, `cov()`, etc.). Do they have the same value?
    - mean of `carbon`
    - variance of `carbon`
    - standard deviation of `carbon`
    - covariance between `carbon` and `nitro`
    - correlation of `carbon` and `nitro`
4. What's the correlation of `carbon` and `nitro`? Does the correlation vary by year?
5. Use `plot` to show the following visualizations of `carbon`
    - histogram
    - pdf
    - empirical cdf
6. What's the potential research question(s) or hypotheses you would like e to explore using this data? (You can write one or two sentences stating your question of interest)