

# STAT331:Final Project

Due: April 14, 2021 at 5pmEST on Crowdmark

## General Instructions

- Due: April 14 at 5pm.
- Each group consists of 3–4 students (see below). Students who have not enrolled in a group by Wednesday March 17th will be randomly assigned to a group.
- Each project consists of a typed report between 7-10 pages (12 point font with standard 1-inch margins and single-spaced) including figures and tables, but excluding a mandatory Appendix containing (but not limited to) all R code.
- Reports must be created with R Markdown and submitted online via LEARN and/or
- Late Penalty: 10% per day. Projects turned in after April 18 will not be graded.
- Your project grade will be worth 40% of your final grade

## Group Enrolment

- Log in to LEARN and join a Group: At the top of the screen, click Connect > Groups.
- Agree on a Group number between 1–70 with your other team members, and select a group.
- The names of all collaborators must be written on your report.

## Project Details

The goal of this project is to analyze the `pollutants.csv` data and write a report on your analysis. The specific goals of your analysis are up to you to decide.

## Dataset

The dataset (posted on LEARN) contains a sample of  $n = 864$  adults included in a study investigating the relationship between exposure to persistent organic pollutants and telomere length—a marker of cellular aging that may be related to certain cancers. The data include the following variables:

- Outcome:
  - `length`: mean leukocyte telomere length
- Exposures (organic pollutants)
  - `POP_PCB1--11`: concentration of 11 different PCBs (Polychlorinated biphenyls; a class of persistent organic pollutants), in pg/g

- POP\_dioxin1--3: concentration of 3 different dioxins (a class of persistent organic pollutants), in pg/g
- POP\_furan1--4: concentration of 4 different furans (a class of persistent organic pollutants), in pg/g
- Other covariates
  - male : 0= female, 1=male
  - ageyrs : age in years
  - edu\_cat : 1 = Less Than 9th Grade or 9-11th Grade (Includes 12th grade with no diploma); 2 = High School Grad/GED or Equivalent; 3 = Some College or AA degree; 4 = College Graduate
  - race\_cat : 1 = Other Race (Including Multi-Racial); 2 = Mexican American; 3 = Non-Hispanic Black; 4 = Non-Hispanic White
  - yrssmoke : how many years smoked cigarettes
  - smokenow : 0=does not currently smoke; 1=currently smokes
  - BMI : body mass index
  - ln\_lbxcot : log of cotinine in ng/mL
  - whitecell\_count : white blood cell count
  - lymphocyte\_pct : percent lymphocytes (out of white blood cells )
  - monocyte\_pct : percent monocyte (out of white blood cells )
  - eosinophils\_pct : percent eosinophils (out of white blood cells )
  - basophils\_pct : percent basophils (out of white blood cells )
  - neutrophils\_pct : percent neutrophils (out of white blood cells )

## Report

Your 7–10 page report must contain the following components:

1. Summary:
  - A maximum of 200 words describing the objective of the report, an overview of the statistical analysis, and summary of the main results.
2. Objective:
  - Describe your goals for the analysis.
3. Exploratory Data Analysis:
  - Conduct exploratory data analyses: report summary statistics, visualize data (histograms, scatter plots, etc.). Report on any interesting findings and comment on how these inform the rest of your analysis.
4. Methods:
  - Describe your statistical analysis: What is your model? Did you use any transformations or extensions of the basic multiple linear regression model? How did you select a model? Does the model fit the data well? Are the necessary assumptions met? Be sure to explain and justify your decisions.

5. Results:

- Report on the findings of your analysis

6. Discussion:

- Comment on your findings/conclusions; describe any limitations of your analysis.

## Grading

- Project grades will consider the following:
  - All required components are included.
  - Ideas are well organized (please use the sections as described above, but you can further divide material into subsections as appropriate)
  - Ideas are clearly expressed, and written in complete sentences.
  - Subjective analysis decisions are reasonable and well-justified.
  - Statistical challenges are well described and addressed appropriately
  - The most important/relevant results and findings are shown and discussed in the report (optionally, any supporting analyses or results you wish to show can be included in the Appendix).
  - Results are presented and interpreted correctly
  - Analysis limitations are acknowledged
  - Conclusions are insightful and well-justified
  - Choice of tables and figures is effective
  - Tables and figures are well presented: captions and labels are informative; axes/scales are appropriate; no needless digits (3 or 4 significant digits is usually ok); no wasted space
  - R Code is clear, well-commented and reproducible