

Data Analytics and Mathematical Statistics

Group Project

Since PM2.5 is a component of PM10, some studies (notably the WHO ambient outdoor air pollution for cities database) have applied a factor to relate the annual mean PM10, based on the annual mean PM2.5, or vice versa, where data for one or the other is not available. Carry out a linear regression modelling study to investigate if an increase in overall particulate matter (PM10) causes fine particulate matter ($\leq 2.5 \mu\text{m}$) to increase, for the following countries, which would justify the application of such a factor:

- a. Cities in Germany
- b. Cities in France

You may choose one country, or both. Refer to the WHO air pollution for cities database (filename: "WHO_DB_FR_GER_cleaned.csv"), as uploaded on Canvas, for your data source.

Group Presentation:

Work in your assigned groups. Each presentation should be around **20 minutes**. You should include the following in your presentation:

1. Introduce the problem – background of the topic and why it is important to study this problem. You may include summaries of any research your group has done on the topic.
2. Application of the quantitative reasoning framework to the study of the problem:
 - o Frame a question to describe your problem
 - o Formulate a hypothesis
 - o Operationalise any concepts in your hypothesis that are not quantifiable, into measurable variables
 - o Deduce a prediction involving these variables
3. Is the data that you have for the study reliable and representative? You may discuss this question by commenting on the source of data, period of sampling, geographical spread etc.
4. Carry out initial exploratory analysis of the data using your choice of descriptive statistics and visualisation methods. What are the insights obtained? Are the data normally distributed? Any trends between variables? Any outliers? Should these outliers be removed, or included? Provide justification for your answers.
5. Fit a linear regression model through the data, estimating the slope and intercept. What does the hypothesis test on the slope tell you?
6. Use residual plots to check the assumptions of regression. Are the assumptions valid? What are the implications if they are violated? What other assumptions do you think you have made in your study, e.g. with regards to the data that you have analysed?
7. Make your conclusion about your topic based on the results of your analyses. Include any recommendations to address the issues concerned, and/or an explanation on the outcome of your analyses, where applicable.

All group members will give the presentation. Each presentation will be followed by a 5-minute question-and-answer session. The session will also be open to audience members. **Upload a copy of your presentation slides + R notebook to the submission folder on Canvas by 2359 the day before the presentation date. Name your files after your group, e.g. "Group 1" etc. Just one group member per team needs to do the upload, on behalf of the other members.**

The group presentation will be graded based on the following:

- Content of presentation
- Delivery of presentation
- Good time management and ability to answer questions