DSC 385 - Project 2 Linear Regression Modeling

Overview

The goal of this project is to use linear regression modeling to explore a question about environmental health. You will need to fit linear regressions, interpret model coefficients, and diagnose model fit. The final product of this project will be a PDF report (a rendered R Quarto file) containing your code and your results.

Topic

For this project you will look at data from the **Mouse Allergen and Asthma Cohort Study**, which was an observational study of children living in Baltimore, Maryland, USA. (The data have been perturbed slightly so that they are not the exact data from the study but still reflect the same statistical characteristics.) The primary aim of the study was to examine the relationship between the indoor home environment and a child's asthma morbidity. Participants in the study were all children aged 5–17 years with asthma and were followed by researchers for a year. For this project, the dataset contains data from each child's very first visit in the study (also known as the "baseline" visit). A representative research paper published from this study can be found on the Canvas web site for this course.

Getting the Data

The data can be obtained from the following GitHub repository:

https://github.com/Principles-of-Data-Science/Project2

You can pull the GitHub repository into RStudio as follows:

1. Click on File > New Project…
2. Click on "Version Control"

3. Click on "Git"
4. Paste in the URL **https://github.com/Principles-of-Data-Science/Project2** into the "Repository URL" field
5. Type in the "Project directory name" if needed
6. Set the directory if you don't want to use the default
7. Click "Create Project"

Dataset Description

The dataset for this project consists of a single CSV file `maacs.csv`. In the dataset you will find the following variables

- `id` a random string identifying a participant
- `age` age in years for the participant at the beginning of the study
- `gender` self-reported gender for the child
- `fev1` forced expiratory volumn in 1 second. This is the amount of air that the child could expel from their lungs in one second. It is often used as a measurement of lung function and lung health. Higher values of FEV1 are considered more healthy but it is dependent on the size of the person. The variable is measured in liters.
- `eNO` exhaled nitric oxide. This is a measure that is used to indicate the amount of inflammation in the lungs. Higher values indicate more inflammation. eNO is measured in parts per billion.
- `cough_days` the number of days in the past 2 weeks where the child experienced coughing or wheezing. This is a count of days and ranges from 0 to 14.
- `pm25` the level of particulate matter less than 2.5 μm in diameter inside the home. PM2.5 is sometimes referred to as "fine particulate matter" and is a measure of indoor air pollution. PM2.5 is measured in $\mu g/m^3$.
- `no2` the level of nitrogen dioxide inside the home. Nitrogen dioxide is an indoor pollutant that can be produced, for example, by burning natural gas. NO2 is measured in parts per billion.
- `mouse` the level of mouse allergen in the child's bed. The units of mouse allergen are not particularly interpretable, but higher values indicate higher amounts.
- `mouse_allergic` an indicator of whether the child is allergic to mouse allergen ("yes") or not allergic ("no"). Allergic status is determined using a skin prick test.

The first few rows of the dataset are shown below.

```
# A tibble: 150 x 10
   id         age gender  fev1   eNO cough_days  pm25   no2 mouse mouse_allergic
   <chr>    <dbl> <chr>  <dbl> <dbl>      <dbl> <dbl> <dbl> <dbl> <chr>
 1 fd171e2d  14.7 male    1.78   141          2  15.6  NA    2423 yes
 2 a66fc33a  13.6 female  2.12    68          8  18.9  12.8   939 yes
 3 fc038e68  14.5 male    2.73   210          2  17.8  NA     200 no
 4 8e28b8c2  14.0 female  2.36    23          0  14.0  NA      NA no
 5 b2699b54  16.6 female  3.13    18         14  17.9  31.0 10371 no
 6 7d4f3508  16.6 female  2.59   128          0  43.8  10.6  4789 yes
 7 3d8242a6  17.2 female  2.6     19          0  26.3  NA     760 yes
 8 f401998a  15.5 male    3.49    26          4  39.9  32.9   264 no
 9 357fdacb  12.7 female  2.29    17          1  27.1  25.7   419 no
10 2a722e16  16.0 male    2.37   134         14  64.6  18.0   187 yes
# i 140 more rows
```

The Report

The text of your report will provide a narrative structure around your code and outputs with R Quarto. Answers without supporting code will not receive credit and outputs without comments will not receive credit either. Write full sentences to describe your findings. All code contained in your final project document must work correctly (render early, render often)!

The report template provided in the GitHub repository contains prompts/questions that you will need to answer. Please follow the prompts in the template and answer all of the questions there.

Formatting

- Create the report using R Quarto knitted to a PDF file, with headers for each section and each question answered;
- Include comments describing your R code;
- Include any references (datasets, context), if needed.
- The final report should be no more than 20 pages including all code/graphics/output (the number of pages can vary greatly depending on the cleaning process).
- It is extremely important that you **select pages** when submitting on Gradescope (see more below). Points will be taken off if you do not select the appropriate pages for each question in the Gradescope outline.

Submission of the Report to Gradescope

This project report will be submitted on Gradescope for grading. Gradescope is a tool that enables the teaching staff to efficiently grade assignments like this one according to a defined rubric. **You will not be submitting this project on Canvas**. Anything submitted to Canvas will be ignored.

If you have never submitted anything to the Gradescope web site, please watch this video demonstration of how to do so.

To submit your project report, please follow these steps:

1. First render your project report into a PDF file. This can be done by either rendering directly to PDF in RStudio or by rendering to an HTML file and then "Printing" to a PDF file. Either way, **you must have a PDF file to submit to Gradescope**.
2. Go to the course Canvas page and click on the "Gradescope" link on the navigation bar on the left hand side.
3. When the Gradescope page loads, click on the assignment titled "Project 1: Exploratory Data Analysis".
4. You should be prompted with a window allowing you to submit a PDF file of your assignment.
5. After uploading your PDF file, you will be prompted to select pages of your PDF file that correspond to questions in the "Question Outline". Please make sure to do this carefully, as it essential for allowing us to grade your project efficiently.
6. After selecting the pages, submit the assignment.

Late Policy

As per the Syllabus, projects will not be accepted late. There are no exceptions; please do not contact the instructor or TA to request an exception.