Project 1: Regression Analysis

This project is about doing a "complete" regression analysis, and presenting your results in an IDMRAD paper. Your IDMRAD paper should have all the elements that we discussed in lecture 02, week 01. ***(Please review the guidelines and extra materials from week 01 as well as your work (and the TA's comments) on the D, M, R and technical appendix parts of an IDMRAD paper from homework 04.)***

You should follow the general approach suggested in homework 04:

- First, write the technical appendix showing all of the work that you need to do (R code, output, graphs, tables, and comments explaining what you did and why) to answer the questions below. This should not contain false starts and side-tracks, but should contain all the work in R you need to justify your results/answers for the questions below.
- Then, write the "results" section, listing each question, and the highlights from the technical appendix that are needed to answer the question. Be sure to refer to specific pages in the appendix at each place in the "results" section where the reader may want more detail.
- Then write the rest of the IDMRAD paper. Write the title and abstract *last*.

The Data

Nancy H. L. Leung and her colleageus recently published the paper "Respiratory virus shedding in exhaled breath and efficacy of face masks" (Leung et al, 2020); a copy of this paper is included in the file `s41591-020-0843-2.pdf` in the `project01` folder in the files area for our class on canvas. They recruited about 246 patients who were ill from various respiratory viruses, and randomly assigned half of the patients to wear a surgical facemask; the other half did not wear face masks. All patients were asked to breathe normally (including coughing etc.—they were sick!) for 30 minutes, and the researchers measured the "viral load", which is the number of viral particles per ml of air at a fixed distance from each subject, from one or two of the subject's breaths. They considered many factors, including the gender and age of each patient, the type of viral infection the patient had, the amount of virus found in nose and throat cultures before the "breathing" experiment, and whether the viral material was carried by larger water droplets in the air, or smaller aerosols. The full data set and data dictionary are contained in the files `G2resp_data_200324.csv` and `G2resp_dictionary_200324.csv` in the `project 01` folder in the files area for our class, for your reference. A reduced version of that data set is contained in the file `maskdata.csv`, also in the `project 01` folder, consisting of the 114 subjects with human coronavirus[1] (hcov), influenza virus (flu) or rhinovirus (cold) infections (see Table 1a in Leung et al., 2020). The variables in this reduced data set are listed in Table 1 on page 2.

References

Leung, N.H., Chu, D.K., Shiu, E.Y., Chan, K.H., McDevitt, J.J., Hau, B.J., Yen, H.L., Li, Y., Ip, D.K., Peiris, J.M. and Seto, W.H. (2020), "Respiratory virus shedding in exhaled breath and efficacy of face masks," *Nature: medicine, 26(5),* 676–680. Obtained October 5, 2020 from https://doi.org/10.1038/s41591-020-0843-2.

[1] Not COVID-19, which hadn't been discovered yet, but other milder coronaviruses including NL63, OC43, HKU1 and 229E.

Variable Name	Values	Description
study	"cold" or "flu"	The "flu" study ran from 2013 to 2014; the "cold" study ran from 2014 to 2016.
age	integer	Age of subjects, ranging from 12 to 84
female	0 or 1	1 = female
allo.wm	0 or 1	random allocation of subject to "mask" or "nomask" group; 1 = "mask"
mask.class	"ctrl", "mask", or "both"	"ctrl" = no mask, "mask" = wore mask; "both" = was included in both "ctrl" and "mask" samples
breath.num	integer	number of exhaled breaths collected (to measure viral load)
cough.num	integer	number of coughs during exhaled breath collection
virus	"hcov", "flu", or "cold"	type of virus measured: hcov = coronavirus, flu = influenza virus, cold = rhinovirus
nasal	real number	viral load (copies per ml) in nasal swab before experiment
throat	real number	viral load (copies per ml) in throat swab before experiment
aerosols	real number	viral load (copies per ml) in exhaled aerosols
droplets	real number	viral load (copies per ml) in exhaled droplets

Table 1: Variables in the file `maskdata.csv`. A value of 2 was used for the viral load when the amount of virus was too close to zero to be measured. NA's (missing data) were recorded whenever a measurement was not made for some reason, or the measurement was lost.

The Research Questions

A medical scientist who is not a member of Leung's research team has read the Leung et al. (2020) paper and realizes there could be much more in the data set, than is reported in Leung's paper. She has asked you to look into it, using only, or primarily, the data in `maskdata.csv`:

1. As a "sanity check" for the data set `maskdata.csv`, can you exactly or approximately reproduce the results in Figure 1 of Leung et al. (2020)? If only approximately, explain why you are, or are not, worried about not getting exactly the same results.

2. In this study, does wearing a mask make a difference? Is the answer different for aerosols, than it is for droplets?

3. Is there a difference between coronavirus, influenza virus, and rhinovirus, in the efficacy of masks? Is the answer different for aerosols, than it is for droplets?

4. Find the multiple regression model that makes the best tradeoff between the following criteria:

   - Best fit, as measured by *adjusted* R^2, AIC, or BIC.
   - Best satisfies the assumptions of the linear model.
   - Best for interpreting and explaining to a medical collaborator or client.

   No matter what you do, you are likely to be unhappy with some or all of these criteria; the better you make one criterion, the worse another is likely to get. So you will have to find a compromise or

tradeoff between the three criteria. Explain how you decided to make the tradeoff(s) you made.

5. Provide a careful and easy-to-follow interpretation of your final model for question 4.

6. Is there anything else interesting to say about the data?

Further Directions And Hints

- For questions 1–4, I expect to see linear regression models[2] of some kind, and possibly other methods. Feel free to use transformations, interactions, etc. as needed, for each of the research questions.
- For question #4, please feel free to consider partial F tests for groups of variables, vif's, and any other tools, to construct your best "tradeoff" model, and possibly to help explain your findings.
- Here are some model-building suggestions:
  - Think about the problem and the meanings (both verbal/scientific and mathematical) of the variables, to choose a good set of variables and a good path through the model space, to work with, and ***look*** at the raw data, transformed data, diagnostic plots, etc., early and often.
  - Don't transform for the sake of showing that you know how to make transformations. Choose transformations that (a) help with modeling in some way; and (b) are still explainable to the medical scientist. If there are no transformations that satisfy these criteria, don't transform.
  - Generally, main effects (the original input variable) are easier to explain than interactions (products of input variables). Two way interactions (product of two variables) are easier to explain than three way interactions (product of three variables), which are easier to explain than 4-way interactions, etc. So don't go too wild with interactions unless you are really getting somewhere that the medical scientist will underastand.
  - Review the advice for model building at the end of lecture 11, from Gelman & Hill.
  - Please do not present 10 different models. Ideally, present your one best model. If necessary, discuss one or two close competitors.
- You should turn in a single pdf containing a complete IDMRAD paper, including title, abstract and technical appendix.
  - Review materials from week 01 on IDMRAD papers. Some of those materials are also in the subfolder in the `Project 01` folder on canvas:
    * `IDMRAD outline from lecture 2.pdf`
    * `ASA Style Guide.pdf`
    * `10 rules for better organized papers.pdf` (see especially the C-C-C ideas in Rule 3.)
    * `menu pricing IDMRAD - version 2 with appx.pdf`
  - You can make the paper in LaTeX, Word or rmarkdown.
    * If you are using LaTeX or word, you will probably need to make a separate pdf of the technical appendix, and then attach that to the pdf for your main IDMRAD report.

[2] Leung et al. (2020) use Fisher's exact test for some of their results, and a variation of the standard regression model called Tobit regression for other results. Feel free to google those methods if you are curious. For this project however, please use standard linear regression.

* If you prefer using rmardown, the TA Ben LeRoy has been working on a .Rmd template that will help you remember all the parts of a good IDMRAD paper. it is also in the Project01 folder on canvas.

You are to do this project on your own, without collaborators. If you are unsure of what something means, feel free to look it up on the web or elsewhere, but you may not post questions on discussion websites or blogs like stackexchange, etc. Questions on Piazza should be **private** *to the instructors.* ***You are welcome to discuss this project with me or the TA (office hours are also fine), but no one else.*** *Please remember to cite all the sources that you used, including webpages, in the reference list at the end of your report.*

Due date

Optional: I will hold extra office hours on Friday October 9, 9:30-11:00am (Pittsburgh time). You may use these office hours to discuss your initial work on the project with me.

Required: You must submit a final pdf on Canvas **(not gradescope!)** by midnight (11:59pm, Pgh time) Friday October 16. "Grace" for late submissions until Saturday at 11:59pm.

Grading

On the next page is a summary of what I will be looking for. See the materials in the week01 folder for more detail. *I have collected the main materisl you should look at in the subfolder* `IDMRAD rules and examples` *within the* `Project 01` *folder on Canvas.*

The percentages in the table on the next page assume that all parts of the paper are there. If one or more parts is missing, it may result in a much lower grade than the percentages suggest.

Part	Looking For...	Percent
Title	*Clear, interesting, focused.*	5%
Author/Contact Info	*Your name & email addr!*	∞!!
Abstract	*Summarize I, D, M R and D sections of the paper* (typically one sentence each).	5%
Introduction	Brief, clear, to the point; context for the problem; What is the problem/aim of the study? *Why would anyone want to read this paper? What questions will be addressed?*	10%
Data	*What data set was used in this study?* Typically, include variable definitions, sample size, quick numerical summaries of the variables and initial EDA, but *no model fitting or analysis*.	5%
Methods	*What did you do, to address these questions?* List the methods and/or analyses that will be used to answer eacb question stated in the **Introduction**. *No data analysis, graphing, model fitting, etc. appears here*; you just say what methods and analyses you will use with which variables, to answer each question.	5%
Results	*Statistical analysis & results* in order parallel to **Introduciton** and **Methods** sections. Here you *finally* get to show the data analyses (model fitting, graphics, etc.) that you did, and what the results were. Don't overload the reader: put the highlights here so the reader understands what you did and why, and refer the reader to specific pages or sections of the **Technical Appendix** for more details. It should be clear which data analyses and results go with which question from the **Introduction**. *Every analysis that is presented here should have been mentioned in the* **Methods** *section.*	10%
Discussion	*What does it all mean? Recap findings; address main problem/question; strengths & weaknesses; implications, unanswered questions, future research.* Typically you will say, for each question from the **Introduction**, how the analyses that you did the **Results** section answers that question. You might also mention EDA and so forth from the **Data** section if that makes clearer to the reader what answers you found for one (or more) of the questions. Then you will talk about the big picture, what future work or generalizations of your work might look like, and any limitations of your study. But *there should be no additional analyses or results in this section; just use the analyses you did for the* **Results** *section (and possibly the* **Data** *section).*	10%
Mechanics	Follows C-C-C[3]as much as possible (sentences, paragraphs & sections); *Grammatical; Complete sentences and paragraphs; Easy to follow.*	5%
Statistical Content	*Correctly and appropriately uses technical and non-technical material* we have learned in class. Easy to follow; Analyses makes sense/not crazy (roughly 6–7% per research question)	40%
References & Citations	*Follow ASA guide*[4]*, "The Reference List".* *Follow ASA guide, "Reference Citations".* Be sure to cite **all** sources!	5%
Technical Appendix	*Contains complete versions of the analyses listed in the* **Methods** *section and presented in the* **Results** *section*: R code, output, graphs, tables, and comments explaining what you did and why. There may be additional analyses here (e.g. to support the **Data** section of the paper, or to show why the methods and analyses that you chose for the paper were the right ones). Make it easy for me to follow.	0%[5]

[5]See Rule 3 in `10 rules for better organized papers.pdf`.

[5]See `ASA Style Guide.pdf`.

[5]You will get credit for this as part of a hw assignment instead.