Learning Influenza Forecasting

Annalise Cramer

I have two potential ideas for this project. The first one would be more challenging for me, but the second I have more skills for currently, please give me feedback on what you think is reasonable.

This uses MS Word as output format. [See here](https://quarto.org/docs/output-formats/ms-word.html) for more information. You can switch to other formats, like html or pdf. See [the Quarto documentation](https://quarto.org/) for other formats.

[1] "/Users/annalisecramer/MADA/CRAMER-MADA-Project"

**Authors**

* Annalise Cramer (ORCID: 0009-0002-8718-3593)

**Author affiliations**

1. College of Public Health, University of Georgia, Athens, GA, USA.
2. Center for Ecology of Infectious Diseases, Athens, GA, USA.

Corresponding author: atc10260@uga.edu

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect those of the University of Georgia.

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

## 2.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 2.2 Description of data and data source

*Describe what the data is, what it contains, where it is from, etc. Eventually this might be part of a methods section.*

## 2.3 Questions/Hypotheses to be addressed

Option 1:

Can I make a flu forecast for this season? When I compare my work to the actual data, will it hold up?

Option 2:

What is the prevalence of poor mental health? Is poor mental health associated with a positive covid test?

Factors studied will include hours of sleep, reported depressed mood, reported trouble sleeping, recent exercise, and reported social/emotional support.

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in (1,2).

# 3. Methods

*Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.* Option 1:

Data for this project would be using positive infleunza case counts, which are listed weekly from CDC’s [FluView](https://www.cdc.gov/fluview/index.html). This data is updated weekly, and extends years backwards. Other useful measures are available, such as hospitalizations. I would need to obtain number from each prior week going backawards, I’m not sure if I’d have to do this manually or if there’s a consolidated form somewhere. Several people in my lab do forecasting, and I’m hoping to learn enough that I can be caught up to speed on their reserach projects. I plan to use and ARIMA model and a weighted interval score to evaluate my work. I’m currently taking Time Series in the statistics department, so I’m hoping as the semester progresses I will learn more about this modeling technique.

Option 2:

Data for this potential project would be using survey data from the National Health Interview Survey, produced by IPUMS through the University of Minnesota. This data was collected among United States residents of all ages and genders from 2019 - 2022 in the United States. There’s 151406 observations and 47 vairables. I would select variables I need, including creating new ones, and using the survey design variables. I would create models (basic and adjusted) to study association (generating prevelance ratios with CIs) between mental health factors (hours of sleep, feeling depressed mood, amount of social and emotional support, trouble sleeping, recent exercise) with a postiive covid test result. I will generate tables and plots for all results.

## 3.1 Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). [Figure 1](#fig-schematic) is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the assets folder.

|  |
| --- |
| Figure 1: A figure that is manually generated and shows some overview/schematic. This has nothing to do with the data, it’s just a random one from one of our projects I found and placed here. |

## 3.2 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 3.3 Data import and cleaning

*Write code that reads in the file and cleans it so it’s ready for analysis. Since this will be fairly long code for most datasets, it might be a good idea to have it in one or several R scripts. If that is the case, explain here briefly what kind of cleaning/processing you do, and provide more details and well documented code somewhere (e.g. as supplement in a paper). All materials, including files that contain code, should be commented well so everyone can follow along.*

data\_original <- readRDS(here("data", "NHIS\_COVID.rds"))  
data2 <- labelled::remove\_val\_labels(data\_original)  
summary(data2)

YEAR STRATA PSU PERNUM   
 Min. :2019 Min. :100.0 Min. : 1.00 Min. :1.000   
 1st Qu.:2019 1st Qu.:111.0 1st Qu.: 8.00 1st Qu.:1.000   
 Median :2020 Median :125.0 Median : 23.00 Median :1.000   
 Mean :2020 Mean :125.4 Mean : 31.13 Mean :1.174   
 3rd Qu.:2021 3rd Qu.:139.0 3rd Qu.: 48.00 3rd Qu.:1.000   
 Max. :2022 Max. :151.0 Max. :153.00 Max. :2.000   
   
 SAMPWEIGHT AGE SEX SEXORIEN   
 Min. : 396.2 Min. : 0.00 Min. :1.000 Min. :0.000   
 1st Qu.: 4417.0 1st Qu.: 24.00 1st Qu.:1.000 1st Qu.:2.000   
 Median : 7054.3 Median : 45.00 Median :2.000 Median :2.000   
 Mean : 8601.8 Mean : 45.74 Mean :1.532 Mean :1.783   
 3rd Qu.: 10733.0 3rd Qu.: 64.00 3rd Qu.:2.000 3rd Qu.:2.000   
 Max. :122902.9 Max. :999.00 Max. :9.000 Max. :8.000   
   
 RACENEW HISPETH EDUC OWNERSHIP LOWRENT   
 Min. :100 Min. :10.00 Min. : 0.0 Min. :10.00 Min. :0.0000   
 1st Qu.:100 1st Qu.:10.00 1st Qu.:103.0 1st Qu.:10.00 1st Qu.:0.0000   
 Median :100 Median :10.00 Median :301.0 Median :10.00 Median :0.0000   
 Mean :194 Mean :14.23 Mean :251.6 Mean :16.27 Mean :0.3423   
 3rd Qu.:200 3rd Qu.:10.00 3rd Qu.:400.0 3rd Qu.:20.00 3rd Qu.:1.0000   
 Max. :999 Max. :93.00 Max. :999.0 Max. :99.00 Max. :9.0000   
   
 HINOTCOVE HIPRIVATEE HICHIPE HIMILITE   
 Min. :1.000 Min. :1.000 Min. :10.00 Min. :10.00   
 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:10.00 1st Qu.:10.00   
 Median :1.000 Median :2.000 Median :10.00 Median :10.00   
 Mean :1.097 Mean :1.655 Mean :10.42 Mean :11.17   
 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:10.00 3rd Qu.:10.00   
 Max. :9.000 Max. :9.000 Max. :99.00 Max. :99.00   
   
 HIOTHGOVE HIMCAIDE HIMCAREE HIHSE   
 Min. :10.00 Min. :1.000 Min. :0.000 Min. :1.000   
 1st Qu.:10.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000   
 Median :10.00 Median :1.000 Median :1.000 Median :1.000   
 Mean :10.36 Mean :1.168 Mean :1.074 Mean :1.032   
 3rd Qu.:10.00 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:1.000   
 Max. :99.00 Max. :9.000 Max. :9.000 Max. :9.000   
   
 ASTHATAKYR ASTHERYR IMSPCHC CIGDAYMO   
 Min. :0.0000 Min. :0.0000 Min. :0.00 Min. : 0.00   
 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00 1st Qu.:96.00   
 Median :0.0000 Median :0.0000 Median :1.00 Median :96.00   
 Mean :0.1656 Mean :0.1406 Mean :0.73 Mean :94.24   
 3rd Qu.:0.0000 3rd Qu.:0.0000 3rd Qu.:1.00 3rd Qu.:96.00   
 Max. :9.0000 Max. :9.0000 Max. :9.00 Max. :99.00   
 NA's :41190   
 SMOKFREQNOW SMKLSEV SMKLSFREQNOW ECIGEV   
 Min. :0.000 Min. :0.000 Min. :0.0000 Min. :0.000   
 1st Qu.:0.000 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:1.000   
 Median :0.000 Median :1.000 Median :0.0000 Median :1.000   
 Mean :0.466 Mean :1.014 Mean :0.2871 Mean :1.052   
 3rd Qu.:1.000 3rd Qu.:1.000 3rd Qu.:0.0000 3rd Qu.:1.000   
 Max. :9.000 Max. :9.000 Max. :7.0000 Max. :9.000   
   
 ECIGED MOD10DMIN VIG10DTP PA18AER   
 Min. :0.0000 Min. : 0.00 Min. :0.00 Min. :0.00   
 1st Qu.:0.0000 1st Qu.: 0.00 1st Qu.:0.00 1st Qu.:1.00   
 Median :0.0000 Median : 20.00 Median :0.00 Median :2.00   
 Mean :0.3221 Mean : 35.78 Mean :0.39 Mean :1.97   
 3rd Qu.:0.0000 3rd Qu.: 60.00 3rd Qu.:1.00 3rd Qu.:3.00   
 Max. :9.0000 Max. :999.00 Max. :9.00 Max. :8.00   
 NA's :78933 NA's :78933 NA's :78933   
 PA18AERSTR HRSLEEP DEPFREQ PHQSLEEP CMPSUPPORT   
 Min. :0.00 Min. : 0.00 Min. :0.0 Min. :0.00 Min. :0.00   
 1st Qu.:1.00 1st Qu.: 6.00 1st Qu.:4.0 1st Qu.:0.00 1st Qu.:0.00   
 Median :1.00 Median : 7.00 Median :5.0 Median :0.00 Median :2.00   
 Mean :2.02 Mean : 8.38 Mean :4.1 Mean :1.89 Mean :1.77   
 3rd Qu.:3.00 3rd Qu.: 8.00 3rd Qu.:5.0 3rd Qu.:3.00 3rd Qu.:3.00   
 Max. :8.00 Max. :99.00 Max. :9.0 Max. :9.00 Max. :9.00   
 NA's :78933 NA's :78933 NA's :75101 NA's :76305   
 CVDDIAG CVDTEST CVDTESTRSLT CVDSYMP   
 Min. :0.00 Min. :0.00 Min. :0.00 Min. :0.0   
 1st Qu.:1.00 1st Qu.:1.00 1st Qu.:0.00 1st Qu.:0.0   
 Median :1.00 Median :1.00 Median :1.00 Median :0.0   
 Mean :1.01 Mean :1.15 Mean :0.73 Mean :0.4   
 3rd Qu.:1.00 3rd Qu.:2.00 3rd Qu.:1.00 3rd Qu.:0.0   
 Max. :9.00 Max. :9.00 Max. :9.00 Max. :9.0   
 NA's :41190 NA's :76305 NA's :41190 NA's :41190   
 CVDDLYCARE CVDDNGCARE CVDVIRAPPCVD CVDSHT   
 Min. :0.00 Min. :0.0 Min. :0.00 Min. :0.00   
 1st Qu.:1.00 1st Qu.:1.0 1st Qu.:0.00 1st Qu.:1.00   
 Median :1.00 Median :1.0 Median :0.00 Median :2.00   
 Mean :0.96 Mean :0.9 Mean :0.48 Mean :1.46   
 3rd Qu.:1.00 3rd Qu.:1.0 3rd Qu.:1.00 3rd Qu.:2.00   
 Max. :9.00 Max. :9.0 Max. :9.00 Max. :9.00   
 NA's :76305 NA's :76305 NA's :76305 NA's :78548   
 CVDSHTNUM   
 Min. :0.00   
 1st Qu.:0.00   
 Median :2.00   
 Mean :1.39   
 3rd Qu.:2.00   
 Max. :9.00   
 NA's :78548

data\_using <- data2 %>%  
 select(STRATA, PSU, PERNUM, SAMPWEIGHT, MOD10DMIN, VIG10DTP, PA18AER, PA18AERSTR, HRSLEEP, DEPFREQ, PHQSLEEP, CMPSUPPORT, CVDTESTRSLT)

## 3.4 Statistical analysis

*Explain anything related to your statistical analyses.*

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

[Table 1](#tbl-summarytable) shows a summary of the data.

Note the loading of the data providing a **relative** path using the ../../ notation. (Two dots means a folder up). You never want to specify an **absolute** path like C:\ahandel\myproject\results\ because if you share this with someone, it won’t work for them since they don’t have that path. You can also use the here R package to create paths. See examples of that below. I generally recommend the here package.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1: Data summary table.   | skim\_type | skim\_variable | n\_missing | complete\_rate | factor.ordered | factor.n\_unique | factor.top\_counts | numeric.mean | numeric.sd | numeric.p0 | numeric.p25 | numeric.p50 | numeric.p75 | numeric.p100 | numeric.hist | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | factor | Gender | 0 | 1 | FALSE | 3 | M: 4, F: 3, O: 2 | NA | NA | NA | NA | NA | NA | NA | NA | | numeric | Height | 0 | 1 | NA | NA | NA | 165.66667 | 15.97655 | 133 | 156 | 166 | 178 | 183 | ▂▁▃▃▇ | | numeric | Weight | 0 | 1 | NA | NA | NA | 70.11111 | 21.24526 | 45 | 55 | 70 | 80 | 110 | ▇▂▃▂▂ | |

## 4.2 Basic statistical analysis

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

[Figure 2](#fig-result) shows a scatterplot figure produced by one of the R scripts.

|  |
| --- |
| Figure 2: Height and weight stratified by gender. |

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example [Table 2](#tbl-resulttable2) shows a summary of a linear model fit.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2: Linear model fit table.   | term | estimate | std.error | statistic | p.value | | --- | --- | --- | --- | --- | | (Intercept) | 149.2726967 | 23.3823360 | 6.3839942 | 0.0013962 | | Weight | 0.2623972 | 0.3512436 | 0.7470519 | 0.4886517 | | GenderM | -2.1244913 | 15.5488953 | -0.1366329 | 0.8966520 | | GenderO | -4.7644739 | 19.0114155 | -0.2506112 | 0.8120871 | |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (3) discusses types of analyses.

These papers (1,2) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References

1. McKay B, Ebell M, Billings WZ, et al. [Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery.](https://doi.org/10.1093/ofid/ofaa494) *Open forum infectious diseases*. 2020;7(11):ofaa494.

2. McKay B, Ebell M, Dale AP, et al. [Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients.](https://doi.org/10.1098/rspb.2020.0496) *Proceedings. Biological sciences*. 2020;287(1927):20200496.

3. Leek JT, Peng RD. [Statistics. What is the question?](https://doi.org/10.1126/science.aaa6146) *Science (New York, N.Y.)*. 2015;347(6228):1314–1315.