PROJECT PROPOSAL

Nicholas Mallis

2021-09-15

## PROJECT PROPOSAL

For the analysis project, I propose to examine the relationship between education level and COVID-19 vaccination rate by county for all counties in the Unites States. Kaiser Health News and other sources have reported this relationship on the individual level using data from surveys and qualitative research. KHN released a report on June 11 that outlines the profile of the unvaccinated. They state that “compared to those who have received a covid-19 vaccine, unvaccinated adults are younger, less educated, more likely to be republicans, people of color, and uninsured.” You can see the whole report here. <https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-profile-of-the-unvaccinated/> Although this has been examined using surveys from individual participants, there are no current reports or papers that have studied this relationship using aggregate data from all counties in the US. The proposed analysis could provide more information on this possible association.

In order to explore this relationship between education level and vaccination rates at the county level, I will pull in data from several different sources and merge them together. Most county level data are tagged with a Federal Information Processing Standards (FIPS) number. This will be the ID variable for merging these datasets.

Main Outcome: My main outcome will be percent of eligible people (age 12+) in a county that have been fully vaccinated. I will obtain this from the CDC’s open-source dataset, COVID-19 Vaccinations in the United States, County. Here is the link to this dataset -> [COVID-19 Vaccine Data](https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh) This dataset “represents all vaccine partners including jurisdictional partner clinics, retail pharmacies, long-term care facilities, dialysis centers, Federal Emergency Management Agency and Health Resources and Services Administration partner sites, and federal entity facilities.” It is updated regularly and rapidly changing so I will need to decide on a cut-off date for my analysis at some point.

Main Exposure: The main exposure will be percent of adults in each county who have a bachelor’s degree or higher. The most current county level data on education is available on the USDA’s Economic Research website and can be found if you click here -> [County Education Data](https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/) This dataset includes information on education level for all years, but I will subset the data to only include observations from the most current year. The most current calculations of education level by county are 5- year averages based on 2015-2019 and they come from the Census Bureau’s American Community Survey.

Possible Confounders: There are other possible confounders that I am interested in examining including poverty, unemployment, median household income, locality (urban or rural) and health insurance rate. This data is not available in the current datasets so I will need to pull in data from a few more sources. The first three variables that I list are all available through the USDA Economic Research Service as well, but in different datasets. They can be found here on the USDA website [USDA data on poverty, unemployment, and median household income](https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/) The county poverty estimates are model-based estimates from the U.S. Census Bureau's Small Area Income and Poverty Estimate (SAIPE) program. The county unemployment rates are from the Bureau of Labor Statistics (BLS) Local Area Unemployment Statistics (LAUS) program. The county median household income comes from the U.S. Census Bureau's Small Area Income and Poverty Estimate (SAIPE) program. The locality (urban or rural) variable will come from the Rural-Urban Continuum Codes Dataset which is also managed by USDA. This can be found here [USDA data on Urban and Rural Counties](https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/)

Unfortunately, I am having trouble finding data on health insurance rate by county in the US, but I think it would be useful so I will keep looking.

There will be close to 3,000 observations in the dataset (one observation for each county).

Brief Analysis Plan: After merging, cleaning, and wrangling the data I will produce summary statistics for each variable. I will plot the main exposure (percent of adults in each county who have a bachelor’s degree or higher) and the main outcome (percent of eligible people (age 12+) in a county that have been fully vaccinated) in a scatterplot to examine the possible linear relationship. I will use linear regression modeling to calculate beta estimate for the bivariate relationship between county education level and county COVID-19 vaccine rate. I will test each potential confounder to see if they are qualified to be included in the model based on confounding criteria. Eventually, I will have a multivariate model that provides a beta estimate for the relationship between education level and COVID-19 vaccine rate while controlling for confounders.

Note on the Analysis Plan: Due to possible statistical issues, I am a little worried about using a percentage as my outcome. I am considering using some sort of transformation on the outcome (maybe arcsin) or possibly categorizing the outcome into two levels (1=High Vaccination Rate, 0=Low Vaccination Rate). I realize that we lose a lot of information in categorizing, but it could be the way to go. I have been reading about possible issues modeling percent as an outcome. Let me know if you have any suggestions on this.

Last Note: I realize that there is a whole set of issues that can occur with an ecological study. Like most people right now, I have vaccination rates on the mind and really wanted to do something with them as the outcome and this was the best way I could come up with. Originally, I wanted to do an analysis that had county “voting majority Trump” as the predictor, but with a quick internet search, I realized that has been run into the ground already.

# Summary/Abstract

*Write a summary of your project.*

# Introduction

## General Background Information

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

## 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.*

## Questions/Hypotheses to be addressed

*State the research questions you plan to answer with this analysis.*

# Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

## 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.*

## 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.*

## Exploratory analysis

*Use a combination of text/tables/figures to explore and describe your data. You should produce plots or tables or other summary quantities for the most interesting/important quantities in your data. Depending on the total number of variables in your dataset, explore all or some of the others. FIgures produced here might be histograms or density plots, correlation plots, etc. Tables might summarize your data.*

*Continue by creating plots or tables of the outcome(s) of interest and the predictor/exposure/input variables you are most interested in. If your dataset is small, you can do that for all variables. Plots produced here can be scatterplots, boxplots, violinplots, etc. Tables can be simple 2x2 tables or larger ones.*

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. t-tests, simple regression model with 1 predictor, etc.) 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.*

Table 3.1 shows a table summarizing the data.

Table 3.1: Data summary table.

|  |  |  |
| --- | --- | --- |
|  | Height | Weight |
| Min. | 133.00 | 45.00 |
| 1st Qu. | 155.25 | 54.25 |
| Median | 166.00 | 73.00 |
| Mean | 165.50 | 72.00 |
| 3rd Qu. | 177.25 | 87.50 |
| Max. | 192.00 | 110.00 |

Figure 3.1 shows a scatterplot figure produced by one of the R scripts.

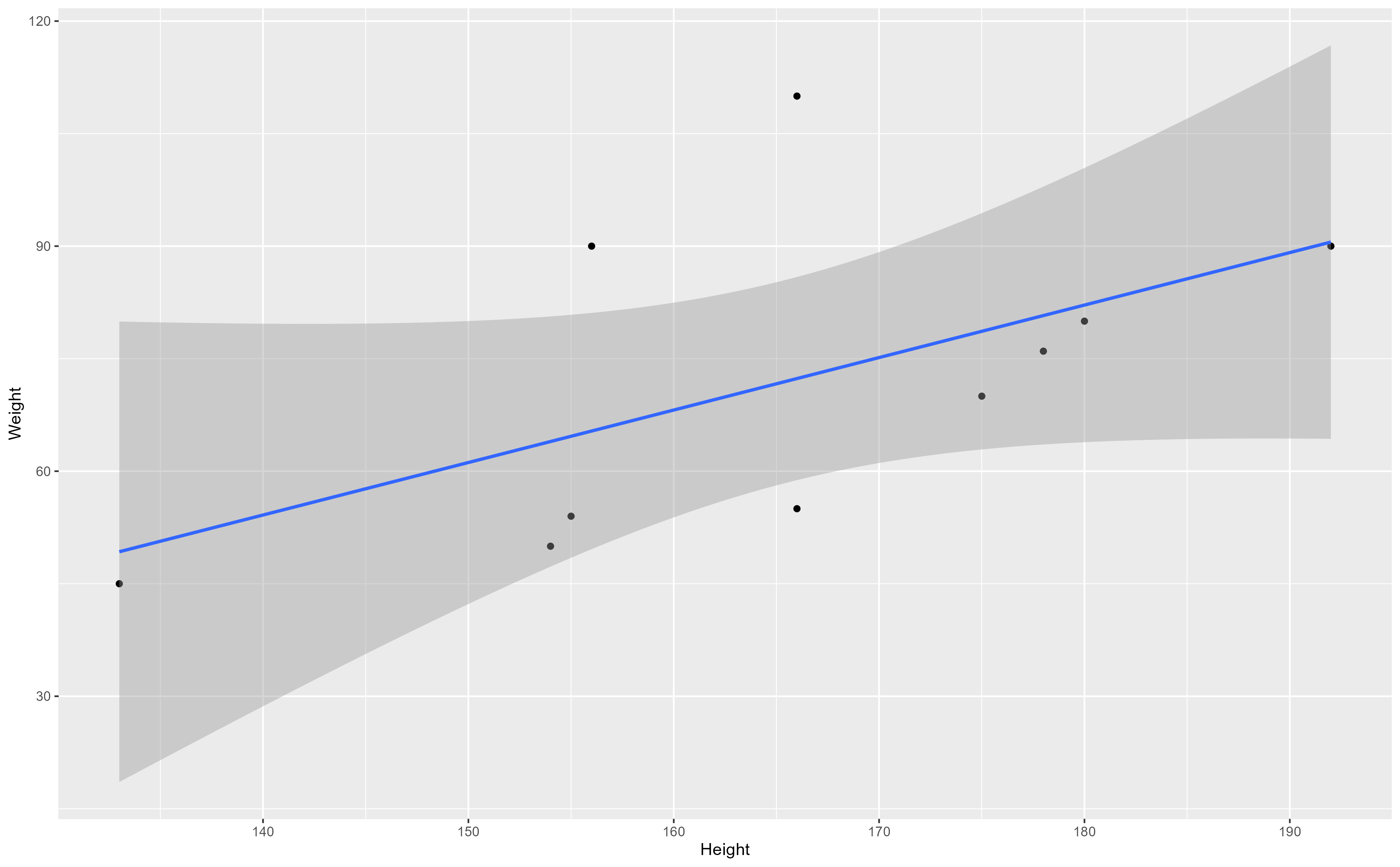


Figure 3.1: Analysis figure.

## 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 3.2 shows a table summarizing a linear model fit.

Table 3.2: Linear model fit table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -43.7883068 | 61.1150617 | -0.7164896 | 0.4940713 |
| Height | 0.6996272 | 0.3675692 | 1.9033889 | 0.0934786 |

# Discussion

## Summary and Interpretation

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

## Strengths and Limitations

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

## 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 (Leek & Peng, 2015) discusses types of analyses.

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, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# References

Leek, J. T., & Peng, R. D. (2015). Statistics. what is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>