Jonathan Amdur

Project Topic 1

11/18/2021

Final Project Outline , Findings, and Questions to Answer

1. Exploring data + Problem Statement
   1. **Goal**
      1. To model the daily new covid cases using data at <https://ourworldindata.org/coronavirus>
   2. Questions about datasource and how to define data for modeling
      1. How many countries to include and how to choose them?
         1. Can consider study design from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7604061/ that used the countries with the earliest known cases to model.
         2. Having one country could help to provide a better model fit, but including multiple could also help since we will have more varied datapoints and can help the model better generalize
         3. Could also utilize only countries that have their first reported case in the dataset accounted for to ensure we have the full time period to model

**Choice:** Chose 5 countries, all of which had their 1st confirmed case occurring during the data time period. Aimed to get a range of locations as well in order to have diversity in the data. Starting choices were chosen from the study linked below. Chose USA, South Korea,

* + 1. What datapoints are useful?
       1. From EDA, certain variables (ie. GDP, population under 65, etc….) are only useful if using multiple countries. If going the route of a single country, these will no longer be useful.
    2. What date ranges are useful?
       1. Not all countries had data from day one and large chunks of the most useful data are missing. May be worth cutting the date range at the beginning to ensure all countries have data. Could also drop countries that had late tracking.
       2. Could transform dates into days since first positive covid infection in country. This could help to put the countries on even ground since they will be at different stages on the same date.
       3. This dataset is updated and maintained so could be worth setting an end date. Keeping with mid-November could be good to take into account the recent effects of boosters

**Choice:** Chose to keep data range from 1/1/2020 – 11/17/2021. This ensured the most data possible and allowed for new potential relationship examination of newer features,such as boosters given. Some later data can be brought in to see how the models predictive power works on new data since it is time series based.

* 1. Findings
     1. Dataset has 134,015 rows and 67 columns
     2. Date Range: 1/1/2020 – 11/17/2021
     3. Number of countries: 237
        1. Some rows are worldwide or income based rather than a specific location. These will need to be removed or specific countries pulled in for modeling.
     4. The **new cases count is our modeling variable**
        1. It is found to display overdispersion with a significantly larger variance than mean
        2. Some countries reported negative counts, likely adjustments for false positive tests on previous days. These countries or results may need to be tossed out since Poisson expects solely positive values, as would the interpretation of the new cases variable

1. Model Type and Research
   1. Since this is a count model, linear regression likely will not be the best fit. We can run that to compare against other possible choices to illustrate this though.
   2. By <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7604061/>, it was found that the negative binomial GLM with log link was the best fit in this experiment. Part of why, is due to the fact that it can better handle over overdispersion. This is continued in this article <https://towardsdatascience.com/adjust-for-overdispersion-in-poisson-regression-4b1f52baa2f1>.
   3. The Poisson model with log link is also a candidate for model for fitting to count data. This is displayed in both *Introduction to Linear Regression Analysis* and the paper *A Flexible Regression Model for Count Data*. From the paper, we are also given the COM-Poisson model that can be used in place of this due to the overdispersion of the data
   4. **Choice:** Run at least the Negative Binomial GLM, Poisson GLM, and Linear Regression models. The Com-Poisson model will be attempted as well, but will need more care and depending on availability of code.
2. Variable Handling
   1. Cleaning
      1. Ensure all variable types matchup and no extraneous characters in any fields
      2. Handling negative new case count values by either removing them or removing the country from the model
      3. Potentially standardize all numeric variables. This decreases the interpretability of the coefficients, but allows us to assume all variables are normally distributed.
      4. Total variables include the days new count as well (ie. Total\_Cases for day n is Total\_Cases for day n-1 + Total New Cases for day n). This will have to be removed to use as a predictor for the current day by transforming it into total cases from previous day.
   2. Null Value Imputing
      1. From *Introduction to Linear Regression Analysis,* we need to have no null values for our models to work. We can impute the null values by using the mode for any characteristic variables and using the median for numeric, possibly calculating at each specific location for both methods for greater accuracy.
      2. For some variables, it makes sense to fill with 0’s since NaN represents where a value just wouldn’t exist. For example, total cases before the 1st are labeled as NaN because there were no cases, but this can be filled with 0. Things like total death and new deaths are similar.
   3. New Variable Creations
      1. Days Since First Case – Difference between reporting date and first date a reported case existed/first date virus was present in location. This can replace the date variable, which is not ingestible by the model, and might allow for a better model since it will allow us to more directly compare countries that are at different stages in the progression of the virus. Further, it will help us better study and understand the effect changes in vaccinations and social distancing scores have on the new case values.
      2. Month of cases – Used to account for seasonality. All countries chosen were in the northern hemisphere so the seasons will be the same for all. The southern hemisphere will likely show opposite trends that would cancel out seasonality affects.
      3. Year of cases – Used to account for seasonality. Will keep as numeric to account for increased/decreased affects as years continue.
3. Run Full Model
   1. Following the procedure laid out in *Introduction to Linear Regression Analysis* and the paper *A Flexible Regression Model for Count Data,*  we first run the models on all variables. We can begin to evaluate what model might be best and compare the approaches.
4. Model Evaluation and Validation
   1. Following the procedure laid out in *Introduction to Linear Regression Analysis* and extending it with the descriptions in the paper *A Flexible Regression Model for Count Data*
   2. Will perform residual analysis, but will utilize the deviance residuals in place of typical residuals for the GLM approaches. Evaluate using the Q-Q plot and Residuals vs Predicted response to see if any assumptions have been violated or transformations are needed
   3. Run variable tests of significance, R^2 values, and PRESS statistics to view models ability to fit the data
5. Deviance Residuals Distribution Estimation
   1. We don’t expect the deviance residuals to necessarily follow a normal distribution. By the paper, we can utilize the normal probability plots still if we first estimate the distribution of the deviance residuals. We can use bootstrap to estimate the residuals and create a QQ plot using this.
   2. In need of literature for how to implement this
6. Feature and Best Model Selection
   1. By *Introduction to Linear Regression Analysis* we can utilize the AIC, BIC, and MSE to evaluate which variables to keep
   2. We also can use stepwise regression methods to choose the appropriate variables
7. Evaluation and Validation for Selected Model
   1. Repeat process from section V. We will do this for each type of model and appropriately compare each against each other
   2. Decide between maximizing description of previous happenings, so fit to historical data, or maximizing the predictive power of the data, fit to unseen and future data
      1. Note, a test train split will be difficult in time series based data
8. Conclusion + Considerations
   1. Draw appropriate connections between final model and data running it on. Interpret the model to explain how the many variables can affect new case loads each day.
   2. Considerations
      1. Not all data is received daily. Some countries have moved towards reporting every few days. This could skew the results a bit.
      2. Reported numbers may not accurately reflect true numbers, as some countries have under/over reported in the past. The validity of the data is assumed for this use case.
9. FINDINGS WHILE MODELING
   1. Lots of collinearity (b/c of things coming from the same country but not ever changing, like GDP or % of population under 65).
   2. Drop new tests b/c it’s going to have a clear relationship with new cases. Other argument would be to keep total tests b/c the more tests given, the better tracing they might have.