1. **Introduction**

**Research question: How do the stringency index and vaccination rate affect COVID-19 infection in Malaysia**

**Motivation**

The originality of the research comes from the fact that research on Malaysia’s COVID-19 situation has not been thoroughly conducted, especially how the stringency index (lockdown measures) affects COVID-19 infections. As a Malaysian, we were unable to obtain vaccines as soon as possible and had poor lockdown measures due to the lack of proactiveness from the Malaysian government. Thus, the importance of the research would indicate the benefits of vaccination as well as implementing lockdown measures on reducing the spread of COVID-19. My interest in this area is from the fact that due to the pandemic, there has been a myriad of negative effects such as online schooling, unemployment, increased hospitalizations, and deaths. Hence, with my research, I hope to highlight the importance of having an effective National Pandemic Strategy to be prepared for any future virus outbreaks.

**Background/Literature:**

A search on the National Center for Biotechnology Information and The New England Journal of Medicine yielded three important research papers titled “The impact of vaccination on COVID-19 outbreaks in the United States”, the “Effect of Vaccination on Household Transmission of SARS-CoV-2 in England” and the “Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine”. The first paper indicated that vaccination reduced the overall infection rate from 9.0% to 4.6% over 300 days for the Pfizer and Moderna vaccines. On the other hand, the second paper indicated that the likelihood of household transmission was roughly 40 to 50% lower in households that have been vaccinated when compared to households that are unvaccinated. Lastly, in the third paper, the results highlight a 95% protection against COVID-19 infections.

This would indicate that in our data analysis, we would expect the overall infection rate to decrease in Malaysia as vaccination rates increased. The results of these papers relate to this question because they explore the effects of vaccination on COVID-19 infections. The research question I will study relates to this because I will be exploring vaccines affect the COVID-19 situation in Malaysia. Additionally, I will be exploring how lockdown measures (stringency index) affect COVID-19 infections which are different to the research papers mentioned.

A point to note is that although not specified in the dataset I will be exploring, it is known that 51% of vaccines used in Malaysia is the Sinovac vaccine as of Sept 9, 2021, which could affect our case rates differently when compared to the results of the research conducted in the United States and the United Kingdom as mentioned previously.

1. **Methods**

The data analysis started off with viewing and cleaning the data where I checked for missing observations and omitting them.

I proceeded to begin my analysis by first randomly dividing up the dataset into train and test datasets with 50/50 proportions to validate my results at the end of my analysis. I conduct most of my analysis using the training dataset only by first obtaining summary statistics and conducting an EDA, allowing me to assess the assumptions of using a linear regression model. I then constructed my model and begin assessing the assumptions formally by first checking that the additional conditions are satisfied. Since condition 2 failed, I applied the Box-Cox transformation to satisfy the condition. With the transformed model and having normality satisfied, condition 2 is satisfied and I begin to interpret the residual plots ensuring that my assumptions are satisfied.

I begin to identify problematic observations including leverage and influential points as well as outliers using their measures and cut-offs. I determined the VIF of my model to ensure that there is no multicollinearity present in my predictors. Lastly, I conducted an F-test to test whether a significant linear relationship exists overall as well as t-tests on my predictors in my model.

Now that I have a potential model, I will fit the preferred model in my test data set and compare their properties to those in the training dataset. If the models look very similar (i.e. similar hypothesis test significance, estimates, VIF values) to how they performed in the test dataset, I can conclude that my model is validated.

1. **Results**

**Data source:**

The main variables of my analysis include vaccination rate (“data\_MYS\_vaccine”), COVID-19 cases (“data\_MYS\_cases”), stringency index (“data\_MYS\_strin”) and vaccine effectiveness (“data\_vaccine\_effect”).

The vaccination rate represents the total number of people who received at least one vaccine dose per 100 people in the total population. COVID-19 cases represent the new COVID-19 cases per million people in the population. The stringency index represents the stringency index which is a composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest response) The vaccine effectiveness denotes a binary variable that represents a cut-off at 60% of the vaccination rate where the cases begin to decrease. Since the effect of the vaccinations is not immediate as it requires a threshold of the population to be vaccinated before cases starts to decrease, the vaccine effectiveness acts as an interaction term with the vaccination rate.

The following describes my analysis using my training dataset.

**Exploratory data analysis:**

Table 1: Descriptive Statistics including Mean and Standard Deviation

|  |  |  |
| --- | --- | --- |
| Variable | Mean | Standard Deviation |
| COVID-19 Cases  Vaccination Rate  Stringency Index  Vaccine Effectiveness | 251.832  36.473  68.315  0.360 | 188.374  31.949  10.191  0.482 |

The descriptive statistics can be found in table 1. When conducting my EDA, I observed much missing data from some of my variables. Thus, I omitted the rows containing the NA values. The limitation of omitting missing variables will be found in the discussion.

Figure 1: Histograms and Scatterplots of Variables

Chart, histogram

Description automatically generated

Chart, scatter chart

Description automatically generated

In figure 1, the histograms depict that COVID-19 cases as well as vaccination rate have distributions that are right-skewed. On the other hand, the histogram for the stringency index does not appear to be normally distributed either displaying skews. Hence, normality will likely be violated since the predictors and the response does not appear normal. Additionally, in the scatterplots, we notice that there isn’t a linear relationship between the COVID cases and vaccination rate. COVID cases rises as vaccination rate rises until a cut-off of around 60% before the cases begin to decrease. This is expected since we do not expect the effect of vaccination to be immediate as most of the population is still unvaccinated and susceptible to contract the virus. Thus, vaccination effect acts as an interaction term with vaccination rate as mentioned previously. Therefore, from our histograms and scatterplots, we observe that there may be issues with non-linearity and non-normality.

The model that is chosen for the analysis is as follows:

I now begin to determine whether there are violations of model assumptions using residual plots. Before using the residual plots, I checked for the additional conditions to ensure that my residual plots provide me with reliable results.

Figure 2: Checking additional conditions to ensure reliability of residual plots (top: before transformation, bottom: after transformation)

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

As observed in the plot of the cases against the fitted values, the points are randomly scattered around the line of best fit which indicates that condition 1 holds. On the other hand, in the pairwise plots between the predictors, we do not observe a linear relationship between them hence, condition 2 does not hold. Since condition 2 does not hold, we apply the Box-Cox transformation to obtain a more desirable model satisfying the conditions.

Upon applying the transformation, the model obtained is as follows.

After the Box-Cox transformation, we observe that the points are randomly scattered around the line of best fit in the plot between the cases and the fitted values which indicates that condition 1 holds. On the other hand, in the pairwise plots between the predictors, there does not appear to be any non-linear relationship between the vaccination rate and stringency index. Therefore, condition 2 is now satisfied.

Figure 3: Residual Plots and Q-Q plot of our predictors and fitted values of the transformed model

Chart

Description automatically generatedChart, line chart

Description automatically generated

Whilst interpreting the residual plots of our fitted values and predictors, we observe that there is no discernible pattern seen and that the residuals are uniformly scattered around 0 for the fitted values, stringency index and vaccination effect. The Q-Q plot appears relatively normal with slight deviations at the end. There is an outlier in the top of the plots which cannot be removed since we do not have contextual reasons to do so hence, we will discuss it when analysing our problematic observations. The main concern is that the vaccination rate still depicts a slight pattern indicating non-linearity which I am unable to improve further despite attempting numerous transformations thus, the impact of it will be discussed in the limitations.

Checking for multicollinearity in our model using the VIF, we obtained 3.965169 for the vaccination rate and 2.765220 for the stringency index which is below the severe multicollinearity cut-off at 5. On the other hand, we obtained very high values of multicollinearity between the vaccination effect and our vaccination rate interacted with the vaccination effect. This is expected since the vaccination effect acts as an interaction term with the vaccination rate. Therefore, there is some amount of multicollinearity between the vaccination rate and stringency index, but it is not severe.

Proceeding with our analysis, we identify problematic observations including leverage and influential points as well as outliers.

Figure 4: Scatterplots of predictors vs. response displaying leverage points (top) and outliers (bottom)

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

As displayed in the following plots, we observe leverage points and outliers in our transformed model. Furthermore, using cook’s distance, DFFITS and DFBETAS, we also observe influential points in our transformed model. Since we do not have contextual reason to remove these points and outliers, we leave them in the model and discuss its limitations in the following discussion.

Conducting a stepwise selection process would lead us to select the same model previously with an AIC of -332.08. Therefore, our final model is as follows.

With this final model, we conducted an F-test with the null hypothesis indicating that none of the predictors are linearly related to the response vs. at least one predictor is related to the response. The F-test provided a result of 253.4 with a p-value of thus, we reject the null hypothesis meaning that at least one of our predictors is linearly related to COVID cases. Furthermore, I conducted t-tests on the predictors which indicates statistical significance at the 0.001 level. Therefore, we can conclude that the predictors are all linearly related to COVID cases.

Lastly, using the preferred transformed model, I fit the model in my test dataset and proceed to compare their properties to those in my training dataset. Comparing both the models yielded from the training and testing dataset, I obtained the same significance on all the predictors, F-test as well as similar adjusted values and estimates of my predictors. Additionally, I also obtained similar results for the VIFs of both datasets. Observing the problematic observations in my test dataset, I obtained a similar amount of these observations when compared to my training dataset. Lastly, I checked my linear model assumptions, and no alarming new patterns are found. Therefore, I can conclude that my model is validated.

1. **Discussion**

The interpretation of our final model is as follows.

If we change vaccination rate to the power of 0.33 by 1 unit, we expect the COVID cases to change by percent (log-level). If we change vaccination effect by 1 unit, we expect the COVID cases to change by percent (log-level). Lastly, if we change the stringency index by one percent, we expect COVID cases to change by 1.52% (log-log).

Since we require the interaction term of vaccination effect to denote when the COVID cases begin to decrease, we are only able to analyse the partial effects of vaccination rate on covid cases and vaccination effect on covid cases. For instance, if there is an effect of vaccination on the covid cases where the percentage of the population vaccinated is greater than 60%, the vaccination effect will equal to 1 which indicates that there is a negative relationship between vaccination rate and covid cases. On the other hand, if there is no effect of vaccination on the covid cases where the percentage of the population vaccinated is less than 60%, the vaccination effect will equal to 0 which indicates that there is a positive relationship between vaccination rate and covid cases.

Based on our results, there are several limitations of our model. The first being that there are missing values in our dataset therefore, it would lead to a reduce in statistical power as well as representativeness of the sample since our model may not be able to capture the full effect of the predictors on the response. Additionally, we are not able to satisfy linearity fully since we still observe a slight pattern in our vaccination rate residual plot despite using transformations. Therefore, to correct for this issue, we utilised the interaction term of vaccine effect with our vaccination rate. Lastly, we also observe problematic observations in our model which could potentially affect the interpretation and coefficients of our model since it displaces our regression line.

**References**

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., & Rodés-Guirao, L. (2021, May 10). *A global database of COVID-19 vaccinations*. Nature News. Retrieved October 22, 2021, from https://www.nature.com/articles/s41562-021-01122-8.

Yahoo! (n.d.). *Covid-19 deaths among vaccinated rare, mostly Sinovac recipients*. Yahoo! News. Retrieved October 22, 2021, from https://malaysia.news.yahoo.com/covid-19-deaths-among-vaccinated-060400766.html.

Moghadas, S. M., Vilches, T. N., Zhang, K., Wells, C. R., Shoukat, A., Singer, B. H., Meyers, L. A., Neuzil, K. M., Langley, J. M., Fitzpatrick, M. C., & Galvani, A. P. (2021, January 2). *The impact of vaccination on covid-19 outbreaks in the United States*. medRxiv : the preprint server for health sciences. Retrieved October 22, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7709178/.

Others, D. M. and, Others, M. B. and, & S. D. Anker and Others. (2021, October 6). *Effect of vaccination on household transmission of SARS-COV-2 in England: Nejm*. New England Journal of Medicine. Retrieved October 22, 2021, from https://www.nejm.org/doi/full/10.1056/NEJMc2107717.

Polack, F. P., Al., E., for the C4591001 Clinical Trial Group\*, Author AffiliationsFrom Fundacion INFANT (F.P.P.) and iTrials-Hospital Militar Central (G.P.M.), Longo, E. J. R. and D. L., Others, A. J. S. and, Others, J. M. and, E. Mendes de Oliveira and Others, Others, D. M. and, Others, M. B. and, & S. D. Anker and Others. (2020, December 31). *Safety and efficacy of the BNT162B2 mrna covid-19 vaccine: Nejm*. New England Journal of Medicine. Retrieved October 22, 2021, from https://www.nejm.org/doi/full/10.1056/nejmoa2034577.