The Effectiveness of COVID-19 Vaccination in Curbing the Spread of the Virus

By: William Wang

1004278818

Eco 375: Applied Econometrics

University of Toronto

Department of Economics

Assignment 1

February 2021

Abstract

The aim of this study was to determine whether the vaccination rate in Israel has any significant affects on curbing the growth of COVID-19 infection rates and whether herd immunity can be achieved at a certain level of vaccination. We conducted both a simple linear regression and multiple linear regression to determine whether there were any indications of a negative relationship. By controlling for time trends and age group, our analyses of the results determined that vaccination rates been associated with a statistically significant and practically significant decrease to the number of infections in Israel. Moreover, we were able to identify a possible non-linear relationship between the rates of vaccination and the number of weekly infections, finding evidence that may support a claim of herd immunity. However, due to the limitations of our models and data, further research and advancements may be needed.

1. Introduction

With Israel being one of the leaders of COVID-19 vaccination research, the opportunity to measure the effects of the COVID-19 vaccine is very important in determining the safety and effectiveness of such a vaccine. As such, a study of the vaccine's effects on the population of Israel and whether a certain percentage of vaccination can achieve "herd immunity" is going to be important if the world wants to get past this pandemic. In this paper, we hope to be able to find evidence of such.

2. The Context and Data

The dataset that will be looked at in particular is COVID-19 data in Israel from September 2020 to the end of January 2021. This set of data contains the vaccination rate, in percent, of all age groups in any given week, with a 2-week lag applied since the vaccine does not work immediately. As well, it contains the number of weekly cases of COVID-19 per million, various dummy variables and some other variables that need to be controlled for such as time and age. In the dataset, we can look at the descriptive statistics for the variables to get a better understanding of the data. We can refer to Table 1 to find the descriptive statistics for all of the continuous variables in the data set. In particular, one can see that on average there are approximately 2598 $\pm$ 2065 weekly cases per million, with a minimum of roughly 262 and a max of 9049, so the range of cases provides us a varied set of data. Additionally, the percentage of vaccinations also has a wide range, with an average rate of 4.43% $\pm$ 15.22% and a range from 0% - 86.5% vaccination. The remaining continuous variables are transformations of the two already mentioned and these were transformed to possibly correct for some issues with the originals, such as correcting for skewness in the cases or non-linearity in vaccination rates.

3. Regression analysis

3.1. Simple Linear Regression

The basic initial model analyzed is a simple linear regression (SLR) model outlined below where β_0 and β_1 are the intercept and slope coefficients for the model, respectively.

$$\widehat{case\_log} = \beta_0 + \beta_1 lagvacc\_per$$

$$\widehat{case\_log} = \underset{(0.0832)}{7.3235} + \underset{(0.0033)}{0.0045}\, lagvacc\_per$$

$$n = 198 \quad , R^2 = 0.0039$$

In our model, if there are no vaccinations for a particular week, then one would expect the log of the weekly cases to be 7.3235 ± 0.0832. Moreover, one can see that the lagged vaccination percentage variable does not really affect the outcome of the natural log of weekly cases too much, adding only 0.0045 ± 0.0033 to the log case value for every unit increase of this variable, i.e., every additional percent of 2-week lagged vaccination. Additionally, we can see that our model does not really fit the data too well since we have a R^2 value of 0.0039, which is nowhere near 1, which would mean the goodness of fit of our model and the data is not very good. Conducting a hypothesis test on whether the effect of vaccination on COVID-19 infection is zero, one can find a p-value of 0.1735. This would indicate that we should not be rejecting the hypothesis that vaccination rates do not have an effect on COVID-19 infection, as this value is greater than the 0.05 significance level for 95% confidence, and thus we could conclude that there is the possibility that this effect is zero.

Taking these results into account, this likely means that this simple linear model may not be the best model to estimate our outcome variable, perhaps due to shortcomings with the SLR assumptions made. In a SLR model, there are 5 main assumptions that need to be satisfied. These assumptions are the model must be linear in the parameters, the sample is chosen randomly, the outcomes of the explanatory variables are not the same, the expected value of the errors is zero for any given value of explanatory variable and the variance of the errors is the same given any value of the explanatory variables. In the case of this model, many of these SLR assumptions may not be satisfied. We can refer to Figure 1 to see that the data is most certainly not linear, especially at the lower vaccination rates, so the first assumption is not satisfied. Moreover, if we think about this model in relation to our research question, if herd immunity is indeed related to COVID-19, then we would expect there to be less infections as the total vaccination rate of the population increases, which would make the model parameters non-linear. Additionally, with the same figure, we can see that it is likely that the fourth and fifth assumptions are also not satisfied as the error of the vaccination rates does not look like to be zero, and variance of errors are not constant. Furthermore, our sample is not random as this is data for all age groups within a certain time period, so individuals were not randomly chosen. The third assumption seems to be the only one that is satisfied as we do have different rates of vaccination. Thus, these failed assumptions along with the R^2 points to the SLR model not being a viable way to estimate the effect of vaccination rates on COVID-19 cases.

3.2. Multiple linear regression

Since it is established that the SLR model may not be sufficient to properly estimate the effect of vaccination on COVID-19 infection, a multiple linear regression (MLR) will also be conducted with various specifications, taking into account factors such as time trends and correcting for non-linearity. With an MLR, there are certain assumptions that need to be made for the model to be valid, the three of which are the same as the SLR assumptions, being linearity in the parameters, random sampling and the zero conditional mean assumptions. In addition to these assumptions there is now also an assumption regarding collinearity, stating that none of the independent variables are constant, and the independent variables have no exact linear relationships.

Each of the three additional regressions done in Table 2 will attempt to solve some of the problems with the SLR model. In column (2) of Table 2, we attempt to control for the week and age using dummy variables, since the number of infections and vaccinations seem to be confounding with the time trend, as both increase as time passes. As well, the age of an individual infected may be related to the likelihood they get infected, since the survival rate of younger cases is much higher than the older, these younger people may not take the pandemic as seriously, leading to more cases with more risky behaviour. Thus, controlling for these may improve our model, this results in the following model

$$\widehat{case\_log} = \beta_0 + \beta_1 lagvacc\_per + \beta_2 D_{week} + \beta_3 D_{age}$$

$$\widehat{case\_log} = 7.9332 - 0.0058\, lagvacc\_per + \beta_2 D_{week} + \beta_3 D_{age}$$

(0.0777) (0.0011)

n= 198 , R^2 = 0.9853

Looking at the R^2 value, we see it increases substantially, going from 0.0039 to 0.9853 once we control for these variables, as such it seems that an omitted variable bias may be in part why our SLR model was not very good. In addition, looking at the p-value, the coefficient for vaccination is statistically significant, being 0, thus we would be able to conclude that vaccination would have a statistically significant negative effect on the number of cases, seeing a 0.58% decrease in cases for every 1 unit increase in the vaccination rate, which is also practically significant as long as the vaccination rate is high enough, since the outcome variable is exponential. However, with this model, we have not really solved the problem with linearity, thus the regression in column (3) of Table 2 will attempt to fix that. In this model, the square of the lagged vaccination rates is added to the model since one may suspect a non-linear quadratic relationship. This results in the model below

$$\widehat{case\_log} = \beta_0 + \beta_1 lagvacc\_per + \beta_2 lagvacc\_per^2 + \beta_3 D_{week} + \beta_4 D_{age}$$

$$\widehat{case\_log} = 7.9237 - 0.0134 lagvacc\_per + 0.000097 lagvacc\_per^2 + \beta_3 D_{week} + \beta_4 D_{age}$$

(0.077) (0.0038) (0.000041)

n= 198 , R^2 = 0.9857

This model has a slightly better R^2 than the second model as now the R^2 is 0.9857, meaning that this model is a better fit for our data than before. Moreover, looking at the p-value for the new square variable, we see that it is statistically significant. Meaning that as we increase the vaccination percentage, at some point, the semi-elasticity of the number of weekly cases with respect to vaccination increases as vaccination rates increases. For the last regression, we take a look at the log of the cases regressed on vaccination dummy variables for specific ranges of vaccination rates. For this, the base group is 0% vaccination with other dummy variables for the range of 0%-10%, 10%-20% and 20%-100% vaccination rates. These results are found in column (4) of Table 2, and we can see that all three ranges are statistically significant, meaning that at all ranges of vaccination rate, we see a decrease in the number of COVID-19 cases, in relation to the 0% group. This result supports the idea that vaccination reduces cases, since there is a decrease in cases when done. Moreover, these estimates are also practically significant since even at the lowest vaccination range, 0%-10%, we see a 45.5% decrease in log cases compared to 0% vaccination, with the other ranges having even greater effects.

In addition to the vaccination rate, we see that in all MLR models age does seem to have an association with the number of cases, as all age ranges except the 20 to 29-year-olds have statistically significant results. This suggests that those not within the age of 20-29 are less likely to contract COVID-19, as they all have a statistically significant negative coefficient. With our specifications in columns (2) – (4), we have satisfied some of the failed assumptions such as the non-linear parameters. However, there still may be some shortcomings with our model, which will be discussed next.

4. Limitations of results

Our models using MLR are a significant upgrade to the SLR model outcomes, as seen with the substantial change in goodness of fit, however, there may also be other problems with our MLR model. First, even using our new MLR model, we have not really satisfied the assumption that we are randomly sampling, as we cannot experimentally choose and randomize a sample, which may threaten the validity of our results since this assumption is not satisfied. Moreover, since within groups, our treatment is not exogenous, it may be difficult to attribute any causal effects from our results. Furthermore, there may still be some things we have not taken into account such as limited testing which may affect the number of cases actually reported, this would have a confounding effect on our overall results and may require an additional variable to account for this. If we are missing any other variables in our model, this would mean our zero conditional mean assumption may not be satisfied thus threatening the validity of our results.

5. Conclusion

Based on our preliminary results of our MLR models, we found that vaccination may have a statistically significant effect on the number of cases of COVID-19 in Israel, finding that as vaccination rates increase, the number of cases decreases. Moreover, certain factors such as age or time may have an effect on the number of infections or vaccination rates for specific age groups. This may be the result of the virus being more deadly to those who are older, resulting in less risky behaviour and thus less cases. In the case of the younger age range of 15-19, a decrease may be the result of these individuals being students and having less freedom in taking risky behaviour. Finally, we found that there may be a non-linear relationship with the vaccination rates and the number of infections, which may support the theory of herd immunity. However, with the limitations of our models, further research may be required to identify the causal chain of COVID-19 vaccination, such as the addition of other potential confounders, such as the limitation of testing.

References:

Roser, Max, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell (2020) - "Coronavirus Pandemic (COVID-19)". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]

Israeli Ministry of Health. (2020) REAL-WORLD EPIDEMIOLOGICAL EVIDENCE COLLABORATION AGREEMENT. Accessed February 3, 2021.
https://govextra.gov.il/media/30806/11221-moh-pfizer-collaboration-agreement-redacted.pdf.

Table 1: Descriptive Statistics of Continuous Variables

	Observations	Mean	Std. Dev.	Min	Max
cases_rate	198	2598.147	2065.757	262.2702	9049.648
cases_log	198	7.343567	1.104857	4.691348	9.657331
lagvacc_per	198	4.433254	15.22428	0	86.4881
lagvacc_per2	198	250.2618	1094.51	0	7480.191

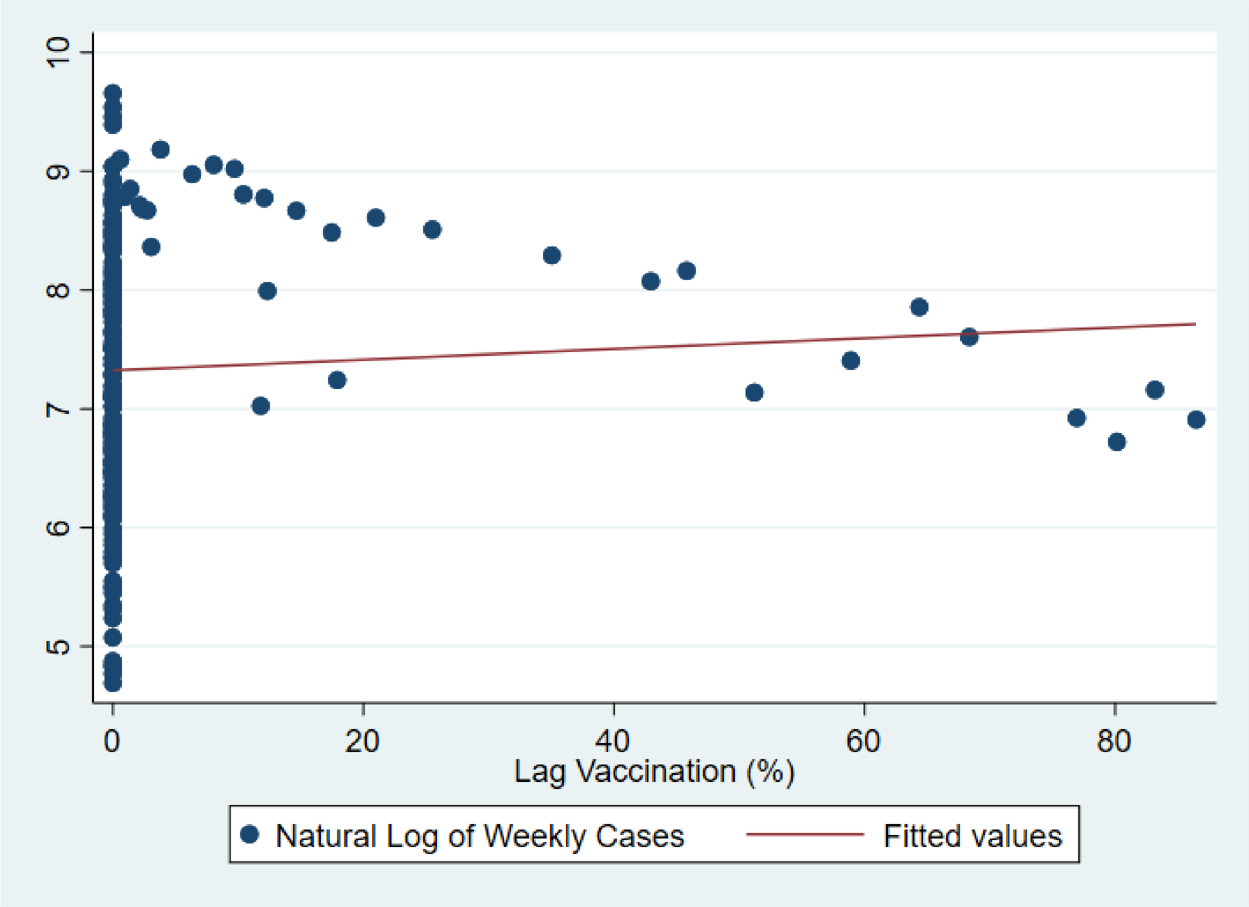
Table 2: Regression Analysis of log COVID-19 infection cases and vaccination

	(1)	(2)	(3)	(4)
Lag Vaccination (in percent)	0.0045	-0.0058***	-0.0134***	
	(0.0033)	(0.0011)	(0.0038)	
Lag Vaccination (in percent) squared			0.000097**	
			(0.000041)	
Lag Vaccination: (0%-10%)				-0.455***
				(0.0538)
Lag Vaccination: [10%-20%)				-0.5431***
				(0.0516)
Lag Vaccination: [20%-100%]				-0.6762***
				(0.0559)
Age:				
15-19		-0.44***	-0.438***	-0.3589***
		(0.0616)	(0.0598)	(0.0587)
20-29		-0.0302	-0.0233	0.0466
		(0.0463)	(0.0433)	(0.0392)
30-39		-0.3024***	-0.2922***	-0.2207***
		(0.0452)	(0.0413)	(0.0379)
40-49		-0.398***	-0.3829***	-0.3067***
		(0.046)	(0.0421)	(0.0391)
50-59		-0.6555***	-0.6364***	-0.5754***
		(0.0449)	(0.0414)	(0.038)
60-69		-1.0152***	-0.9976***	-0.9478***
		(0.0427)	(0.0398)	(0.0374)
70-79		-1.6587***	-1.6533***	-1.6067***
		(0.0525)	(0.0506)	(0.0482)
80+		-2.042***	-2.0322***	-1.9831***
		(0.0638)	(0.0624)	(0.0601)
Year-Week Dummies	No	Yes	Yes	Yes
R-Squared	0.0039	0.9853	0.9857	0.9874
N	198	198	198	198

Note: The standard errors reported in this table are in fact robust standard errors, as there may be problems with homoskedasticity of the errors, see the section on SLR assumptions in section 3.1 for more details. As a result of the robust regression analysis, this means that the R-Squared are also reported, rather than adjusted R-Squared values.

***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Figure 1: Plot of the log cases vs the lagged vaccination rates.


Lag Vaccination (%)
● Natural Log of Weekly Cases
Fitted values