By: Yoke Hong Si

1005815806

Eco 375: Applied Econometrics

University of Toronto

Department of Economics

Assignment 1

February 2021

Abstract:

This paper explores the effect of vaccination on COVID-19 infection in Israel using simple and multi linear regression models in our analysis. From our results, we were unable to determine a statistically significant effect in our simple linear regression analysis. Upon controlling for several confounding variables, we were able to determine a significant effect in our multilinear regression analysis. However, there still exists omitted variables that would lead to bias in our results thus, we cannot conclude that there is purely a causal effect of vaccination on COVID-19 infection.

1. Introduction

In this paper, we will be examining the effect of vaccination on COVID-19 infection in Israel. We will be utilising econometric methods such as simple linear regression and multiple linear regression models to analyse this relationship empirically using the data provided.

2. The Context and Data

Israel is currently a leading country for COVID-19 vaccination and from their vaccination campaign, medical data will be collected from their population of 9 million people. We will analyse the medical data in this empirical paper. To introduce a brief discussion of the summary statistic of our variables we can look at the size of estimated effects in context. From Table 1, we determined the mean COVID-19 cases per week to be around 2598 per million. Hence, a 3% decrease in cases will reflect a decrease of around 78 cases per million per week. This is approximately a decrease of 702 cases per week for Israel's 9 million population. Furthermore, we observe a difference in COVID-19 rates across age groups as shown in Figure 1, where people who are aged 80+ have the lowest COVID-19 cases and people who are aged 0-14 have the highest COVID-19 cases.

3. Regression analysis

3.1. Simple Linear Regression

The estimated simple linear regression of the log of COVID-19 cases regressed on the two-week lag of first dose vaccination (measured in percent) is as follows:

$$\widehat{case\_log} = 7.323544 + 0.0045166\ lagvacc\_per$$
$$(.0818457) \quad (0.0051737)$$
$$n = 198,\ R^2 = 0.0039$$

Firstly, if lagvacc_per = 0, the predicted case_log will be the intercept which is 7.323544. If we write the predicted change in case_log as a function of change in lagvacc_per:

$\Delta\widehat{case\_log} = 0.0045166\ (\Delta lagvacc\_per)$, it means that if the lagvacc_per increases by one percentage point, $\Delta lagvacc\_per = 1$, then case_log is predicted to change by about 0.0045166. In practical terms, this implies that the number of weekly COVID-19 cases (per million people) increased by 0.45166% for every additional unit increase in lagvacc_per.

The t-statistic gives a value of 0.87 which falls within the interval of [-1.96, 1.96] that corresponds to a significance level of 0.05. Therefore, the effect of vaccination on COVID-19 infections is not statistically significant. Hence, in our sample, we determined the value

of lagvacc_per to be 0.0045166 however, the value of lagvacc_per is not sufficiently great in order to conclude a statistically significant effect at the population level. Thus, we fail to reject the null hypothesis.

We are uncertain if SLR.1 holds since according to the Simple Linear Regression model, the equation $\widehat{case\_log} = 7.323544 + 0.0045166\, lagvacc\_per$ follows SLR.1 in the form of $y = \beta_0 + \beta_1 x + u$. However, due to the theoretical mechanism, SLR.1 does not hold since there would be a nonlinear effect of vaccination rate on the COVID-19 cases due to the positive externality of the vaccine. Eventually, herd immunity will be achieved and there will be zero effect of further vaccination on case rates. SLR.2 does not hold since the dataset is obtained from different age groups and weeks in Israel for a particular time period. Therefore, since there is correlation across time, there would be a positive correlation across observations which causes an underestimation in our standard errors as we are not accounting for the correlation in the data. SLR.3 holds since from the dataset provided, we know that the values of lagvacc_per are not all the same value thus, there is variation amongst the values of lagvacc_per. SLR.4 indicates that we assume $E(u|x) = 0$. However, given the current context, this will not hold. For example, occupation (that is part of u) correlates with lagvacc_per (x) since front-line workers (medical staffs) will tend to have a higher rate of vaccination compared to ordinary people. SLR.5 indicates that we assume $Var(u|x) = \sigma^2$. However, given the current context, this will not hold. For instance, consider $Var(\text{agegroups}|lagvacc\_per)$ where we look at the variability of age groups given the vaccination rate. The older age group tend to have a higher vaccination rate since they are given priority to the vaccine thus, this leads to a low dispersion across the age groups. On the other hand, people with lower vaccination rates would include people such as infants or middle-aged adults, leading to a high variability across the age groups. The graph displaying heteroscedasticity of COVID-19 Cases on Vaccination can be found in Figure 2.

3.2. Multiple linear regression

From table 2 column 1, we determine the coefficient of the lagvacc_per to be .0045166. This implies that the number of weekly COVID-19 cases (per million people) increased by 0.45166% for every additional unit increase in lagvacc_per. In columns 2, 3 and 4, we begin to control for a set of week and age group dummy variables. In column 2, the number of weekly COVID-19 cases decreased by 0.57974% for every additional unit increase in lagvacc_per. In column 3, the number of weekly COVID-19 cases decreased by 1.34144%

and increased by 0.00967% for every additional unit increase in lagvacc_per and lagvacc_per2 respectively. Lastly, in column 4, the number of weekly COVID-19 cases decreased by 45.50193%, 54.30955% and 67.61747% for every additional unit increase in lagvacc_0_10, lagvacc_10_20 and lagvacc_20_100 respectively.

By adding a set of control variables, the statistical significance of the coefficient of lagvacc_per changes from statistically not significant to being significant at the 1 percent level. Therefore, in practical means, we can infer that there is a statistically significant difference in the number of weekly COVID-19 cases under the effect of vaccination. The data in table 1 indicates that the number of weekly COVID-19 cases decreased by 0.57974% for every additional unit increase in lagvacc_per when controlling for a set of week and age group dummy variables. Thus, we reject the null hypothesis.

As we move from specification (1) to (2), we are effectively taking the year-week and age group out of the error term and including it explicitly in our equation. Since our coefficient of interest is the coefficient of lagvacc_per, we would want to fix all other factors that affects it. Year-week and age group are examples of factors that can affect the result of the coefficient. The vaccination rate tends to be higher when the year-week progresses and for elderly individuals. Therefore, we would expect the year-week and age group to be part of the confounding time trend in both infection and vaccination. Since we measure the effect of vaccination rate on COVID-19 cases holding the year-week and age group fixed, the coefficient of lagvarr_per in (2) yields a more accurate result. From a theoretical standpoint, vaccination is an economic good that provides a positive externality, which, by nature, yields decreasing returns up to the point of herd immunity. Hence, we would expect a non-linear effect of vaccination rate on the COVID cases. Evidence of a non-linear effect from the estimation can be found in specification (3). The coefficients of lagvacc_per and lagvacc_per2 are determined to be -0.0134144 and 0.0000967, respectively. Since the coefficients of lagvacc_per is negative and lagvacc_per2 is positive, the quadratic of specification (3) will have a parabolic shape with an absolute minimum. In addition, the coefficients also implies that at low vaccination rates, an additional unit increase in vaccination rate has a negative effect on the log of weekly cases. However, at some point, the effect becomes positive and the shape of the quadratic means that the semi-elasticity of weekly cases with respect to vaccination rate is increasing as vaccination rate increases. Realistically, we would not expect weekly COVID-19 cases to increase thus, for practical purposes, the part of the quadratic where it is increasing can be ignored.

From specifications (2) – (4), we control for a set of week and age group dummy variables. Therefore, this can address a part of the assumptions SLR.4 and SLR.5 since we take week and age group out of the error term and include them explicitly in the equation. Hence, by controlling for these variables, we can ensure that these variables do not correlate with the independent variable of interest which helps satisfy the assumptions. However, since we are unable to account for all variables that could potentially affect the desired variable, the specifications may exclude relevant variables which could lead to omitted variable bias.

4. Limitations of results

Despite controlling for year-week and age group in the previous section, our MLR econometric models in Table 1 may be underspecified as it cannot capture all omitted variables that may lurk in the residuals. This could include variables such as income that could affect the vaccination rates as we would expect people with higher income to be able to afford the vaccine compared to lower income individuals. Thus, omitted variables could underestimate or overestimate the coefficient of our variable of interest, leading to omitted variable bias. The bias could take forms of positive or negative bias and it depends on the coefficient of the omitted variable as well as the sign of the correlation between the vaccination rate and omitted variable. Due to this bias, we are unable to interpret the results of COVID cases as purely a causal effect by vaccination.

Additionally, there are several other factors that would affect the validity of our regression results. One being that we are unable to conduct randomised sampling to obtain the data. Due to the lack of random sampling, it would lead to open backdoor paths which fails the conditional independence assumption leading to selection bias in the experiment. Furthermore, due to the nature of the benefits from vaccination, an accurate econometric model representing the effect of vaccination rate on COVID-19 cases will be nonlinear. These will in turn lead to biasness of the OLS estimators for the population parameters.

5. Conclusion

In conclusion, through SLR, we are unable to determine a statistically significant effect of vaccination on COVID-19 infections. However, with MLR, we controlled the year-week and age groups thus, a statistically significant effect was achieved, illustrating a decrease in infections as we increase vaccination rate. Ultimately, correlation does not prove causation thus, we are unable to determine a causal effect of the decrease in infections due to the vaccine since there could be bias from any omitted variables.

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: Table of descriptive statistic

Variable	**Observations**	**Mean**	**Standard Deviation**	**Minimum**	**Maximum**
Start Date of Week	22			30th Aug 2020	24th Jan 2021
Year-Week	22			2020 Week 35	2021 Week 4
Age Groups	9	40-49		0-14	80+
Weekly COVID-19 Cases (per million people)	198	2598.147	2065.757	262.2702	9049.648
Natural Log of Weekly COVID-19 Cases	198	7.343567	1.104857	4.691348	9.657331
Two Week Lag of 1st Dose Vaccination (in percent)	198	4.433254	15.22428	0	86.4881
Two Week Lag of 1st Dose Vaccination (in percent) Squared	198	250.2618	1094.51	0	7480.191
Dummy: Two Week Lag of 1st Dose Vaccination					
0% of Population	166	.8383838	.3690314	0	1
(0%-10%) of Population	12	.0606061	.2392111	0	1
[10%-20%) of Population	7	.0353535	.1851399	0	1
[20%-100%] of Population	13	.0656566	.2483086	0	1

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

	(1)	(2)	(3)	(4)
Lag Vaccination (in percent)	.0045166	-.0057974***	-.0134144***	
	(.0051737)	(.0010479)	(.0036669)	
Lag Vaccination (in percent) squared			.0000967**	
			(.0000446)	
Lag Vaccination: (0%-10%)				-.4550193***
				(.0870872)
Lag Vaccination: [10%-20%)				-.5430955***
				(.0926803)
Lag Vaccination: [20%-100%]				-.6761747***
				(.087064)
Age:				
15-19		-.4400148***	-.437979***	-.3588792***
		(.043834)	(.0433679)	(.0438842)
20-29		-.0302479	-.0233398	.0466429
		(.0438458)	(.0434865)	(.0438842)
30-39		-.3023669***	-.2922038***	-.2207126***
		(.0438638)	(.0436402)	(.0436532)
40-49		-.3980066***	-.3828524***	-.3067294***
		(.0439233)	(.0440059)	(.0434309)
50-59		-.6554999***	-.6363954***	-.5753782***
		(.0440983)	(.0445023)	(.0434309)
60-69		-1.015151***	-.9975742***	-.9478091***
		(.0447388)	(.0449908)	(.0436201)
70-79		-1.65868***	-1.653272***	-1.606745***
		(.0453784)	(.0449547)	(.0436201)
80+		-2.042018***	-2.032193***	-1.983133***
		(.0450697)	(.0448102)	(.0436201)
Year-Week Dummies	No	Yes	Yes	Yes
Adjusted R-Squared	-0.0012	0.9827	0.9831	0.9849
N	198	198	198	198

The values in Table 2 are computed using Stata. The dataset used to compute the values is obtained from an Israeli dataset on COVID-19 infection and vaccination by age and week. The time period of this dataset is between 30th of August 2020 to the 24th of January 2021.

Lag Vaccination refers to the vaccination rate (the cumulative percent of people in that age group vaccinated with the first dose of the Pfizer vaccine) from two weeks before the current week. Lag Vaccination (0%-10%) refers to a dummy variable that indicates whether if the value of Lag Vaccination is between 0%-10% (excluding 0% and 10%). Lag Vaccination [10%-20%) refers to a dummy variable that indicates whether if the value of Lag Vaccination is between 10%-20%

(including 10% and excluding 20%). Lag Vaccination [20%-100%] refers to a dummy variable that indicates whether if the value of Lag Vaccination is between 20%-100% (including 20% and 100%). The quantities in parentheses below the estimates are the standard errors.
***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

Other tables and figures if needed.

Figure 1

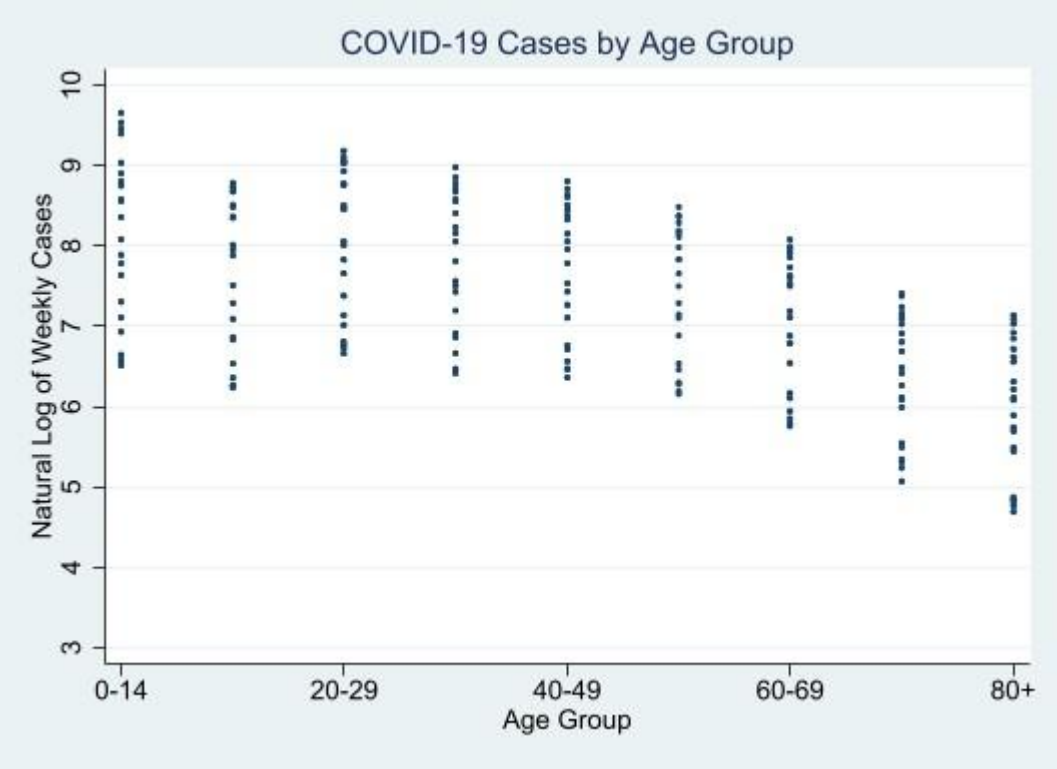

Figure 2