Bachelor's project

NTNU
Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

Thea Ebbeline Nygård Haugen
Maren Rege

A quantitative analysis of the relationship between 5G technology and covid

Bachelor's project in economics
Supervisor: Costanza Biavaschi
May 2021

NTNU
Norwegian University of Science and Technology

Thea Ebbeline Nygård Haugen
Maren Rege

A quantitative analysis of the relationship between 5G technology and covid

Bachelor's project in economics
Supervisor: Costanza Biavaschi
May 2021

Norwegian University of Science and Technology
Faculty of Economics and Management
Department of Economics

NTNU
Kunnskap for en bedre verden

Table of contents

1. Introduction

1.1 Motivation

In relation to the coronavirus pandemic, there has been a rise in conspiracy theories. Such theories can be harmful and misleading, they also contribute to a more polarized society.

> *"Conspiracy theories cause real harm to people, to their health, and also to their physical safety. They amplify and legitimize misconceptions about the pandemic, and reinforce stereotypes which can fuel violence and violent extremist ideologies."* UNESCO Director-General (2020)

On this basis, we have decided to examine the conspiracy theory that links 5G technology to the spread of coronavirus. The theory suggests that 5G deployments increase the likelihood of contracting covid. We will first study if there is an association between the two. Thereafter we expand our model to examine whether the observed relationship is causal or spurious. This is important to address as conspiracies spread uncertainty and suspicion regarding government policies, like vaccinations, lock downs and facemasks. It thereby acts as a constraint on the reopening of our society. This results in increased unemployment rates, running expenditures and increased future costs related to the aftereffects of the pandemic. Overall, there is an economic cost tied to this corona conspiracy. In this paper we will therefore address this issue and convey whether the relationship is causal or spurious.

1.2 Research question

We decided to look at this specific conspiracy theory because it is preposterous. It has been debunked by health and technology professionals, but still, it has a large number of believers. At the same time, to the best of our knowledge, no comprehensive statistical analysis exists on the relationship between 5G and covid. Our research question is the following:

"Is the relationship between 5G technology and the spread of coronavirus causal or spurious?"

This leads us to the main motivation behind this thesis, namely how statistics are used to further an agenda, whether it is for political or economic gains. Statistics are often misinterpreted or manipulated, specifically in regard to assuming causation when

correlation is observed. It is important to remember, correlation is not the same as causation. In this thesis we aim to shed light on such a relationship. To this end, we have assembled from several sources a unique dataset. It covers more than 180 countries and includes information on their Covid cases, 5G coverage and a number of other economic and covid related characteristics.

2. The empirical basis of conspiracies and 5G

Covid-19 is an infectious disease caused by a newly discovered coronavirus. Its first incidences were recorded in December 2019, from people who had attended a market in Wuhan, China. Due to its rapid spread, the World Health Organization (WHO) classified it as a global pandemic in March 2020. A pandemic refers to a disease that spreads across large geographical areas and affects a great number of people (Tjernshaugen et al., 2021).

Concerning the coronavirus pandemic, the European Commission and UNESCO have seen a rise in harmful and misleading conspiracy theories. According to the European Commission, a conspiracy theory is "The belief that certain events or situations are secretly manipulated behind the scenes by powerful forces with negative intent." (The Directorate-General for Communication, 2020). Such theories often have a "logical" explanation to events or situations that may be difficult to understand. In addition, they bring a false sense of control and agency. Conspiracy theories often start as a suspicion. One asks who is benefiting from the event or situation, thus identifying the conspirators. Any "evidence" is then forced to fit the theory. Conspiracy theories mostly spread online, and once they have taken root they can grow quickly. The theories are hard to debunk because any person who attempts to do so is seen as being a part of the conspiracy. People who spread conspiracy theories might do so because they believe they are true. Others do it because they want to provoke, manipulate or target people for political or financial reasons (The Directorate-General for Communication, 2020).

During the pandemic, a conspiracy theory has linked 5G to the spread of covid. 5G is the 5th generation mobile network, after 1G, 2G, 3G, and 4G networks. In comparison to the former networks, 5G delivers at higher performance and improved efficiency (Qualcomm, 2017).

The 5G and covid conspiracy first gained global momentum in social media. It is based on the idea that the radio waves sent by 5G technology causes weakening of the immune system, thus making it easier to contract the virus. Or that the virus can be transmitted through the use of 5G technology. The conspiracy has been furthered by the fact that Wuhan was one of the first cities that was introduced to 5G. As a result, it caused several 5G deployments globally to get caught on fire and exposed to vandalism (NTB, 2020).

The conspiracy has been denied from both a technological and medical standpoint. It is said that a connection between the virus and 5G is impossible. Simon Clark, the Associate Professor in Cellular Microbiology, University of Reading denies this conspiracy. He states that:

> "Viruses are tiny particles made up of genetic material, wrapped in a layer of proteins and fats. ... In the case of this coronavirus, it infects cells in human lungs in order to replicate, damaging them and also causing a harmful immune reaction in the process. 5G radio signals are electromagnetic waves, very similar to those already used by mobile phones. Electromagnetic waves are one thing, viruses are another, and you can't get a virus off a phone mast." (Science Media Center, 2020)

To the best of our knowledge, no quantitative analysis has carefully analyzed the relationship between 5G and covid (Science Media Centre, 2020). Yet, such an analysis seems important in light of what was discussed above. Such conspiracy theories have brought and can bring unnecessary economic costs to involved firms (e.g. firms that have seen their places destroyed) as well as individuals (anxiety due to increased concerns). This paper aims to provide the first cross-national evidence on this question.

3. Simple data set

There is no single source that comprehensively combines information on 5G networks and Covid cases. For this reason, we assembled a new dataset, by putting together data from a number of different sources. We start off by presenting our simple data set. Our simple data set is cross-sectional. In such an analysis one can ignore any minor timing differences in collecting the data.

3.1 Presentation- simple data set

We will now introduce the variables that we have included in our simple data set. A complete list of every variable in the data set and an explanation of their value are included in appendix. 1.

fiveg. The variable *fiveg* is a continuous explanatory variable that represents all 5G rollouts in cities across the world. The data is collected from the Ookla 5G map. Ookla requires materials verifying the deployment type, including online sources or a press release detailing the deployment, for the 5G rollouts to be added to the data set (Ookla, 2021). The data set differentiates between 5G operators and 5G deployments. An operator is a company that provides 5G. Deployments are the software that enables 5G. In our data set, we only included 5G deployments for all cities available. In addition, we included all available countries without 5G, these take the value of 0.

Figure 1 – Map of all 5G deployments

Figure 1 shows a map of all 5G deployments (Ookla, 2019). We see that there is a great variation of deployments across continents. One can see that the majority of deployments are in more high-income areas. For instance, we observe that Africa and South America have hardly any 5G deployments compared to Europe.

confirmed and ***incidence_rate.*** These are both continuous explained variables, that give us the covid case statistics for the observed country. They are, respectively, the total number of confirmed corona cases and the covid incidence rate per 100 000 capita. Both are divided into the smallest geographic areas that we could find. This means that for some countries, like the US, we have information at the county level, while for others, like Norway, the same information is given at the country level (see appendix 2 for an overview of how the countries are divided into regions). The data is retrieved from John Hopkins University's Covid-19 map (Johns Hopkins University, 2021). Hopkins University uses a great number of different sources to collect the required data, like WHO, news sites, ECDC, etc. We have purposely chosen to only use the data from the start of the pandemic until the date of 01.12.2020. This is due to the vaccination process, which first started commercially on the 14.12.2020 (Guarino et al., 2020). We did not want the vaccination process to affect our results.

Figure 2 – Map of covid cases

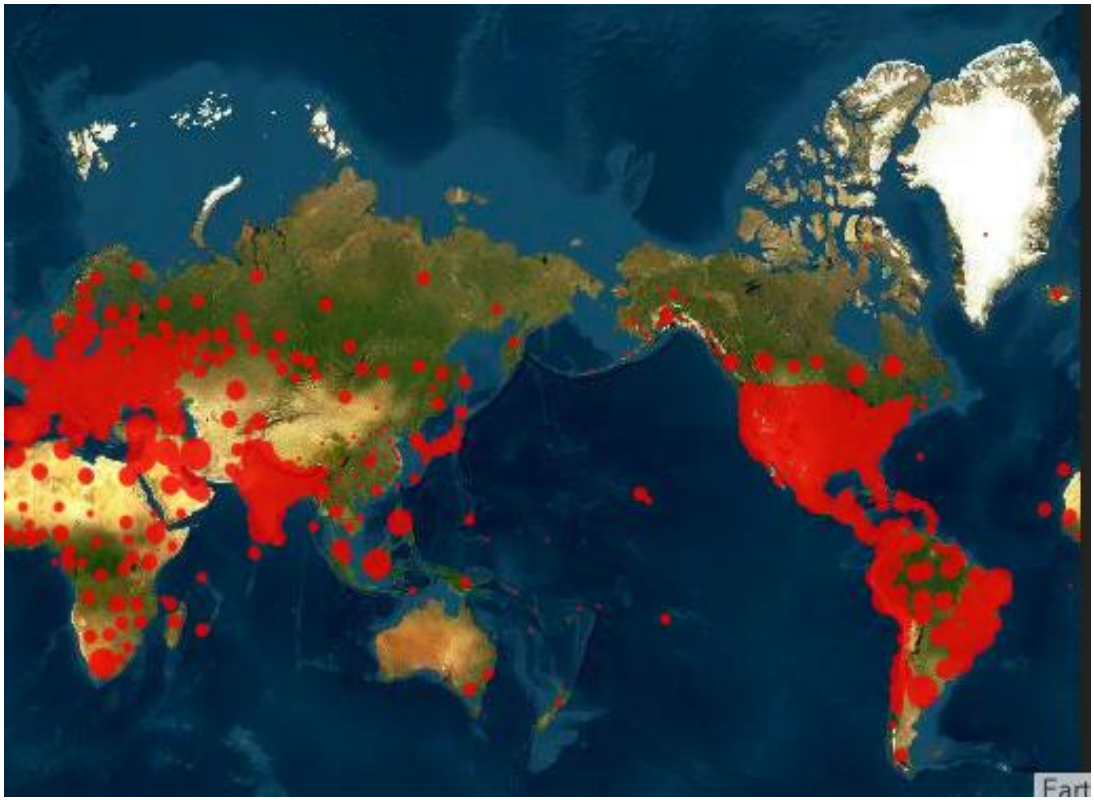


Figure 3 – Map of incidence rate

Figure 2 shows a map of confirmed covid cases (Johns Hopkins University, 2021), while figure 3 shows the incidence rate per hundred thousand (Johns Hopkins University, 2021). The colored dots indicate the scope of covid cases and the incidence rate. It can be a bit difficult to distinguish the different sizes of the dots. We have chosen to include the maps regardless, as they indicate that there is an observable correlation between 5G and Covid, when comparing figure 1 with figure 2 and 3. For instance we see how low-covid places like Africa also have a low number of 5G deployments. While high-covid places like Europe, have a high number of 5G deployments. To find whether this relationship is causal or spurious is the aim of the thesis.

3.2 Descriptive statistics- simple data set

Table 1 – descriptive statistics for our simple dataset

	Obs	**Mean**	**Std. Dev.**	**Min**	**Max**
incidence_rate	2313	3914.25	2639.845	0	15916.05
confirmed	2313	23646.46	105032.9	0	2231344
fiveg	2313	8.81885	48.95654	0	1184

Note: The descriptive statistics shows the number of observations, mean, standard deviation, minimum and maximum values of the variables incidence_rate, confirmed and fiveg (see 3.1 for variable sources).

Table 1 shows the descriptive statistics for our explained variables, both *incidence_rate* and *confirmed*, and explanatory variable *fiveg*. All variables have the same number of observations. This means that the regressions will include 2313 observations.

The mean shows the average value for all the observations for the given variable. The average value indicates the standard and is used to determine if the county, province, or country is above or below the average. If we compare the mean for the three variables, with the given interval of observations, it indicates that there are some extreme observations that are more clustered at lower values. This is because the mean is not close to the middle value in the given interval of observations, it is consistently below. This is supported by the standard deviation.

The standard deviation shows the variables' average deviation from the mean. A relatively low standard deviation tells us that the data is clustered around the mean, while a relatively high standard deviation indicates that the data are more spread out. A high standard deviation can also be a result of one or several outliers and should always be interpreted with the help of the interval for the observations. There are a large number of high-value outliers in our data set. This will be discussed further, later in the paper.

3.3 Criticism - simple data set

As this is a data set we have combined ourselves, with the help of two different data sets, there might be a greater risk of human error. The data set that we gathered from The Ookla map had the areas listed on a city level. While the data gathered from John Hopkins covid-19 map sometimes had the cases listed by county level and other times on a country level. To combine the two, we summarized the 5G deployments in each area as it is given by the John Hopkins covid-19 map. Even though we have done our due diligence while doing this, by regularly controlling for mistakes, there is always the possibility of human error. For example, by summarizing wrong, or omitting a value that should have been included.

4. Results from the simple data set

We perform a test of the conspiracy theory, where our goal is to infer the effect 5G has on covid. We simply wish to find the association between the two variables. We are using the Ordinary Least Squares (OLS) method of estimation on the two Single Linear Regression (SLR) models, with Cross-Sectional data.

4.1 Investigating the association between 5G and confirmed covid cases

A simple regression model can be used to study the relationship between two variables. We set up a simple single linear regression model (SLR) that studies whether the number of 5G deployments affects confirmed covid cases.

Simple linear regressions are defined by only having one independent variable. Confirmed covid cases is the dependent variable and 5G deployments is the independent variable. The variable *u* is the error term. It represents all factors other than 5G deployments that affect confirmed covid cases. To derive a conclusion regarding the conspiracy, we have to be sure that the relationship we are studying captures the effect of *5G* on covid cases. It should not cofound the causal effect of 5G in regard to other variables. Our first SLR model takes the form:

$$confirmed = \beta_0 + \beta_1 fiveg + u \qquad \text{(model.s1)}$$

β_1 is the slope parameter. It explains the relationship between confirmed covid cases and 5G deployments, when holding all other factors in u fixed. β_0 represents the intercept. The linearity of the SLR implies that a unit change in 5G deployments has the same effect on confirmed covid cases, regardless of the initial value of 5G deployments.

From Stata we obtain the following sample regression functions for the SLR model:

Table 2 – Relationship between the confirmed covid cases and 5G deployments

	Model. s1 **confirmed**
fiveg	753.7***

	(18.04)
_cons	16999.7***
	(8.18)
Observations	2313
R²	0.123

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: A regression of model.s1 (see 3.1 for sources). Regressed the dependent variable confirmed on the independent variable, fiveg. t statistics are given in parentheses.

In model.s1, when regressing *confirmed* on *fiveg,* we obtain our predicted value for covid cases for any level of 5G deployments. From the coefficient, we find that with the inclusion of one more 5G deployment it is estimated that confirmed covid cases will increase by 753.8, all else equal. This result substantiates the conspiracy theory, as it suggests a positive relationship between 5G deployments and confirmed covid cases. In order to test if these results are statistically significant, we perform a hypothesis test.

Hypothesis testing is used to determine whether the OLS-estimator corresponds with a given significance level. The significance level represents the likelihood of rejecting the null hypothesis. A hypothesis test is used to test if the estimate is statistically different from the true parameter. The value of the true parameter represents the null hypothesis (H_0) and is chosen in regard to the test we are performing. The alternative hypothesis (H_A) represents a specified deviation from the null hypothesis. We either reject or fail to reject the null hypothesis.

Using a 5% significance level, we test if the estimated value of the parameter β_1(*fiveg)* is statically higher than zero. We wish to provide evidence on the plausibility of the null hypothesis. Our null hypothesis is equal to zero, as we assume that there is no positive linear relationship between 5G and confirmed cases. Since the conspiracy suggests a positive linear relationship between the two variables, we set the alternative hypothesis to be that $\beta_1$is greater than zero. This gives us a one-tailed test. We formulate the hypothesis:

$H_0: \beta_1 = 0$

$H_A: \beta_1 > 0$

We are using an estimate of the standard deviation of the sampling distribution, it is therefore appropriate to use a t-statistic. We assume that assumptions for inference have been met. Then the t-distribution is standardized and equal to $\sim t_{n-k-1} = t_{df}$. df represents the degrees of freedom of the t-distribution. It is equal to the number of observations (n) minus the number of slope parameters (k) minus the intercept (1). Making the degrees of freedom for model.s1 2313 - 1 - 1 = 2311.

Intuitively we reject the null hypothesis if the observed test statistic is far from zero. We defined the rejection region using a 5% significance level. This means that we reject as long as TS is in the right tail of the distribution, which should occur 5% of the time for this t-distribution. If TS falls in the tail, we are adequately assured that we have enough evidence to reject H_0. We search in the t-table of the critical value for which P (Z > c) = 0.05. For a t_{2311} the critical value is roughly 1.645. Hence, we will fail to reject H_0 if TS falls below 1.645 and we will reject if TS is above 1.645 (Thomas, 2005, s. 587).

We test the null hypothesis, by finding the test statistic (TS) associated with the statistic for $\widehat{\beta_1}$. For the TS we take the value of the slope of the regression line ($\widehat{\beta_1}$) and subtract it by the slope assumed in the null hypothesis (β_1) , then we divide it by the standard error of the sampling distribution (se($\widehat{\beta_1}$)).

$$TS = \frac{(\widehat{\beta_1} - \beta_1)}{se(\widehat{\beta_1})} = \frac{(\widehat{\beta_1} - 0)}{se(\widehat{\beta_1})} = \frac{\widehat{\beta_1}}{se(\widehat{\beta_1})}$$

The TS for model.s1 (see appendix 3 for the standard error):

$$TS = \frac{753.7805 - 0}{41.75701} \approx 18.05$$

For the model we get, TS > critical value, 18.05 > 1.671, we therefore reject H_0. This means that the data is not compatible with a zero- relationship between 5G deployments and confirmed covid cases. On the contrary it suggests that there is a positive relationship between 5G and covid.

According to the single linear regression model and the hypothesis test, it is evident that there is a positive relationship between 5G and covid. Based on this we could understand

why people might believe in the conspiracy theory. However, this analysis alone does not imply causation.

4.2 Investigating the association between 5G and the covid incidence rate

In model.s1 we looked at the number of confirmed covid cases and not the more statistically accurate measure for comparison, the incidence rate. It is more accurate, seeing that it is more profitable for companies to develop 5G in areas where there is a larger customer base. Since confirmed cases are not adjusted to the population size, there is a greater chance of discovering a connection. This might explain why people believe in the theory. Model.s1 is therefore, somewhat, imprecise. To adjust for this we set up a second SLR model that studies the relationship between the covid incidence rate per hundred thousand and 5G deployments instead. We wish to explain the incidence rate in terms of 5G and study how it varies with changes in 5G deployments.

We now get our second simple model, where we regress the incidence rate on 5G. The second SLR model takes the form:

$$incidence\_rate = \beta_0 + \beta_1 fiveg + u \qquad \text{(model.s2)}$$

From Stata we obtain the following sample regression functions for the SLR model:

Table 3 – Relationship between the Covid incidence rate and 5G deployments

	Model.s1 **confirmed**	**Model.s2** **incidence_rate**
fiveg	753.7*** (18.04)	-3.493** (-3.12)
_cons	16999.7*** (8.18)	3945.1*** (70.87)
Observations	2313	2313
R^2	0.123	0.004

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: regression for simple model.s1 and model.s2 (see 3.1 for sources). Regressed the dependent variable, confirmed, on the independent variable, fiveg, for model.s1. Regressed the dependent variable, incidence_rate, on the independent variable, fiveg, for model.s2. t statistics in parentheses.

As for model.s2, we obtain our predicted value for the covid incidence rate for any level of 5G deployments. We find that with the inclusion of one more 5G deployment it is estimated a 3.487 decrease in the incidence rate per hundred thousand, all else equal. As suggested, when adjusting for the incidence rate, the overall results change. We test if these results are statistically significant.

We have the same hypothesis that was formulated earlier:

$H_0: \beta_1 = 0$

$H_A: \beta_1 > 0$

We defined the rejection region using a 5% significance level. For a t_{2311} the critical value is 1.671. We will fail to reject H_0 since TS falls below the critical value, -3.12 < 1.671. This means that the data do not suggest that there is a positive relationship between 5G and covid.

When controlling for the incidence rate, instead of confirmed cases, the relationship between covid and 5G switches from being positive to negative. In the first hypothesis test we found a positive relationship between 5G and covid. However, when controlling for the incidence rate we can no longer suggest this. We rather observe a negative relationship. This seems strange, as the experts have clearly stated that there should be no relationship at all. This needs to be further addressed.

4.3 Examining results

To find an unbiased estimate of β_1, for both models, several assumptions need to be satisfied.

The first assumption requires linearity. For this bachelor thesis, we will assume that this assumption holds.

The second assumption requires random sampling. In the data set, we have included the number of confirmed cases/incidence rates at a regional, county, or country level. Every country/region with publicly available data have been included in the John Hopkins map. There might be data at regional or county level that are missing. Some regions or counties

that have not been recorded might have outliers. Therefore, the remaining sample is not random. So, the second assumption might not hold.

The third assumption requires enough variation in 5G deployments, meaning the variance not being equal to zero. From the descriptive statistics we find that the variance in 5G deployments is not zero. Therefore, this assumption is satisfied.

The fourth assumption requires a zero-conditional mean. It means that the covariance between the error term and the independent variable, 5G deployments, must be zero, $(u|5G) = 0$. Due to the simplicity of the model, we might have omitted variable bias, then this assumption is likely to not hold.

4.4 Concerns of Omitted Variable Bias

The crux of our research question is whether the observed relationship between 5G and covid is causal or spurious. When investigating the association between 5G and confirmed cases we can suggest a positive relationship. However, when adjusting for the incidence rate we found a negative relationship between the two. There might be a problem with both analysis as we ignored other determinants of the dependent variable that correlate with the independent variable. Influences on the dependent variable, which are not captured by the model, are collected in the error term. As addressed above, the error term might be correlated with the independent variable and the omitted variable a determinant of the dependent variable. This might induce an estimation bias, where the mean of the OLS estimator is no longer equal to the true mean. Model.s1 might, therefore, wrongly suggest a causal effect on covid for one additional 5G deployment. This issue is called omitted variable bias. Omitted variable bias is the bias in the OLS estimator that arises when the independent variable is correlated with an omitted variable.

There are several variables that may cause omitted variable bias when not included in the model. A highly relevant variable could be the covid testing rate, as it is impossible to know if someone has contracted covid without a covid test. This means that the number of confirmed cases depends on the testing rate. GDP could also be a relevant variable, as it is plausible that places with higher GDP can better respond to the pandemic given that they

have more resources. When not controlling for omitted variable bias we risk wrongly estimating a causal relationship between 5G and covid when it might be spurious. To give a more accurate analysis of the relationship we therefore extend our data set to include more explanatory variables.

5. Extended data set

As we have just discussed, the simple data set suggests that the changes in the incidence rate can be explained by other factors not included in our model. In order to examine if the relationship we have discovered is actually robust, we will now include several different control variables. This is in order to reduce concerns of endogeneity. Endogeneity occurs when there is a correlation between the explanatory variables (x) and the error term (u) in a model. An endogeneity problem is one aspect of the broader question of selection bias discussed earlier.

5.1 Presentation of the extended data set

Our extended data set is cross-sectional. The data structure consists of a sample of countries, taken at a given point in time. In such an analysis one can ignore any minor timing differences in collecting the data. The data used in this analysis has been retrieved from several different sources. See appendix 4 for a complete list of all variables in the extended data set and their explanations.

pop_density. This is a continuous explanatory variable that represents the population density for a given country. In order to get this data, we have marked each available area as "geography" in excel. Excel then has a function that retrieves information from a geographical, this is collected from "data.worldbank.org". It finds the latest available data for each area. We used this function to get the corresponding area and population. Then we divided the population on the area to get the population density. We have included this variable as covid is a highly contagious virus, making it more likely that densely populated countries have a greater infection rate.

median_age. This is a continuous explanatory variable. It gives the middle age in the population when the ages are arranged from lowest to highest. In order to find this we used

the “geography” marker in excel, which finds the latest available data for each country. We have chosen to include this variable seeing that countries with an overall older population might impose stronger restrictions to lower the number of covid related deaths.

gdp_percapita. The variable is a continuous explanatory variable and represents the observed country’s gross domestic product (GDP) per capita, in 2020. GDP per capita is the sum of a country's total domestic output of all goods and services divided by its population. In order to find the GDP we used the “geography” marker in excel. We then divided GDP on the population for the given country. We have included this variable as it is more likely that a country with a higher GDP per capita will have more resources to respond to the virus.

education. Is a continuous explanatory variable, calculated from 2019. It shows the education index provided by the United Nations (United Nations Development Program, 2020). It is calculated by taking the mean years of education received by the population over 25 years old, where the maximum years is 15. In addition, it uses the expected years of schooling, which is calculated by the number of years a child is expected to attend any form of education, with a maximum of 18 years. Both mean years and expected years of schooling are weighted 50%, and the index is given on a country level. Each country gets a score between 0 and 100, where a higher value says that the country educates a larger proportion of their inhabitants. It gives us a good indicator of how well educated the observed country's population is.

GINI. GINI is a continuous explanatory variable that represents the Gini index. The Gini index measures the relative degree of income inequality. It does so by determining the ratio of the area between the line of equality and the Lorenz curve. The Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest households. The GINI index is measured on a scale from 0 to 100, where 100 is equal to perfect inequality and 0 is equal to perfect equality. This implies that countries with highly unequal income distributions have a higher Gini coefficient. For our variable, we collected the latest available measurement for each country. The data is collected from our world data (Roser and Ortiz-Ospina, 2013). We included this variable as it gives an indicator of the county’s overall living standard.

test_per1000. This is a continuous explanatory variable. It represents the number of citizens per hundred thousand who have gotten tested for covid (Hasell, J., Mathieu, E., Beltekian, D. et al, 2020). The data is represented on a national level. In our data set, we have included the data of covid tested for all available countries. We purposely included tests per hundred thousand from the start of the pandemic until 01.12.2020, due to the vaccination process. It is an important variable as it is impossible to know if someone has contracted covid without a covid test. This means that the number of confirmed cases depends on how much a country tests.

corruption. This is a continuous explanatory variable that gives a score from 0 to 100 indicating how corrupt a country is. The more corrupt a country is, the lower that country's score will be. The data is gathered from transparancy.org and is given for 2019 (Transparency International, 2020). We have decided to include this variable as it gives an indication on how much the observed country's citizens trust its government. Thereby indicating how persistent the inhabitants are when it comes to following the government's covid restrictions.

stringency_index. The stringency index records the strictness of government policies and is a continuous explanatory variable. It gives countries a ranked score between 0 and 100, where a score of 100 equals the strictest response. The index is developed by The Oxford Coronavirus Government Response Tracker (OxCGRT). It is a mean composite of the following nine metrics: school closures, workplace closures, cancellation of public events, restrictions on public gatherings, closures of public transport, stay-at-home requirements, public information campaigns, restrictions on internal movements, and international travel controls. It is important to note that this does not illustrate the appropriateness or effectiveness of a country's response (Our World In Data, 2021). We gathered the data from the date of 01.12.2020.

We have also changed the *fiveg* variable to give the number of 5G deployments at a country level (see appendix 5 for an overview for the total number of 5G deployments in each country). We have done this in order to include other variables that possibly affect the covid

cases, which were only available by country. If the geographical granularity for the observations still varied, it would have given the other explanatory variables for countries divided into smaller sections, a greater weight. This would result in the estimated coefficient being unreliable. The explained variable *incidence_rate* have also been changed to give us the values for each country instead.

5.2 Descriptive statistics for the extended data set

Table 4 - Descriptive statistics for our extended data set

	Obs	Mean	Std. Dev.	Min	Max
incidence_rate	181	1205.116	1515.497	.3256	8787.938
fiveg	181	111.5801	649.5756	0	7337
gdp_per_capita	181	16476.7	25995.6	261.2475	184397
median_age	173	30.35202	9.080535	15.1	48.2
education	175	66.02971	17.49594	24.9	94.3
gini	148	38.91892	8.125188	25.6	63.4
corruption	171	43.50877	19.02888	9	87
pop_density	181	309.1087	1555.043	2.061974	19289.11
stringency_Index	167	54.71	18.04089	8.33	87.04
test_per1000	94	256.9337	359.0344	3.782	2196.626

Note: The descriptive statistic shows the number of observations, mean, standard deviation, minimum and maximum values of the variables incidence_rate, fiveg, gdp_per_capita, median_age, education, gini, corruption, pop_denisty, stringency_index and test_per1000. See 5.1 for variable sources.

This table shows the descriptive statistics for our explained variable, the covid incidence rate per hundred thousand, our interest variable 5G deployments, and our control variables. By control variables we mean the variables we have chosen to include to "control" if we are examining the relationship between our explained variable and our interest variable, or if the relationship that we have previously confirmed is actually spurious.

We can see that the maximum number of observations is 181 and is only applicable for four variables. While tests per hundred thousand have below 100 observations. This might be a

problem when it comes to a complete regression analysis, seeing that Stata will only use the observations that have registered values for every variable.

For the incidence rate, 5G, GDP per capita, population density and number of tests we have a large interval of observations. Here the mean is far lower than the middle value of the interval. Combining this with a high value for the standard deviation, points to our data set having clustered observations for the lower values. With one or more outliers taking higher values.

For the median age, stringency Index, education index, GINI and corruption the mean seems to be, to varying degrees, close to the middle value of the observation interval. Combining this information with the somewhat smaller standard deviations, we get variables with less clustered and more evenly spread observations.

5.3 Criticism of the extended data set

In the extended data set, we look at every observation on a country level. This means that the theoretical maximum number of observations is 195, seeing that there are only 195 countries in the world. Out of these 195 countries, there is not always data available for every variable in every country. As a result, some of our variables have quite few observations, making any result we might find less reliable. However, we do not have any other choice when it comes to carrying out a quantitative analysis that depends on a geographical level consisting of countries. Seeing that this is a sample size, it is indeed a larger proportion of the full population. However, the sample size might not be random.

Considering it is more likely to lack data in low-income then high-income countries, the data set becomes less representative. This is because high-income countries usually consists of relatively similar institutions, which makes it more likely that their explanatory variables are somewhat homogeneous. It suggests that our regression might not be an accurate representation of the whole world.

Another criticism of the data set is the incidence rate, and how objective it is. There have been several news articles talking about different countries' tendencies to underreport their

covid cases. Some governments want to seem successful in their response to the pandemic. This leads to them pressing reporters and hospitals to suppress the number of cases they report. We have for example seen this in India recently (Gettleman et al., 2021). This is a threat to our data set as the John Hopkins Covid-19 map uses the media in different countries to update their data. They do try to exclude the statistics in countries where they suspect underreporting, but this might be difficult to catch. Therefore, this data set may be affected by human error, leading any estimations to be incorrect. John Hopkins University goes back in their data and corrects the reporting that has been proven to be wrong. As the data we have used is from 01.12.2020, and we gathered this data in March 2021, this gives John Hopkins University a window of 3 months to correct for any mistakes. This makes it more likely that the data that we have retrieved is still accurate.

6. Results from the extended data set

For our first model, we are using the Ordinary Least Squares (OLS) method of estimation of a Single Linear Regression model. We use a sample, our extended data set, to estimate something about the population. The model describes the relationship between the incidence rate and 5G deployments at a country level.

For our second and third models, we are using the Ordinary Least Squares (OLS) method of estimation on Multiple Linear Regression models. We use a sample, our extended data set, to estimate something about the population. The models describe the relationship between the variables of interest.

6.1 Investigating the association between 5G and the covid incidence rate - country level

Since we have changed our *incidence_rate* and *fiveg* variable to no longer count for different geographical granularity, we will first show the changes in the regression model. We now wish to explain the incidence rate in terms of *5G* at a country level. We get our first extended model, where we regress the incidence rate on 5G. The MLR model takes the form:

$$incidence\_rate = \beta_0 + \beta_1 fiveg + u \qquad \text{(model.e1)}$$

Table 5 - Relationship between the Covid incidence rate and 5G deployments

	Model.e1 **Incidence_rate**
fiveg	0.467**
	(2.73)
_cons	1153.0***
	(10.27)
Observations	181
R^2	0.040

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: regression model.e1 (see 5.1 for sources). Regressed the dependent variable, incidence_rate, on the independent variable, fiveg, for model.e1. t statistics in parentheses.

We find that β_1 has change from - 3.493 to 0.467 form model.e1 to model.s2. Meaning that the effect that one additional 5G deployment have on the covid incidence rate has increased in value. This can be a result of our data set decreasing from 2313 observations to 181, at a country level. Given our data set's interval for the incidence rate, [0.326, 8787.938], this is not a significantly great increase. It is however interesting how the coefficient changed from a negative to a positive value. Seeing that our previous data set has observations on a lower geographical level, we would assume it to be more accurate. The point of showing this again is to see if our new geographical level will affect our result.

We have the same hypothesis that was formulated earlier:

$H_0: \beta_1 = 0$

$H_A: \beta_1 > 0$

We defined the rejection region using a 5% significance level. For a t_{179} the critical value is roughly 1.645. We will reject H_0 since TS falls below the critical value, 2.73 > 1.645. This means that the data do suggest that there is a positive relationship between 5G and covid. We can, therefore, still not debunk the idea that 5G affects the spread of covid.

On the other hand, we do see that the R-squared for this model is quite low, 0.04. The R-squared is the statistical measure for how close our data is to the regression line. A value of 1 indicates that the model explains 100% of the variability of the response data around its mean. Meaning that the variance in 5G only explains 4% of the variance in the incidence

rate. It is generally recommended that the R-squared value should be at least 0.10, preferably higher, in order for the model's explanatory power to be deemed adequate. As the R-squared is below 0.10, it is too early to draw any conclusions on whether the relationship is casual or spurious. We will therefore expand our model to see if different control variables can explain a larger part of the variance in our explained variable.

6.2 Investigating a causal relationship by including economic variables

A multiple linear regression model (MLR) is a model that allows us to explore how multiple independent variables are related to the dependent variable. The dependent variable is defined by Y. We define the independent variables as x_1, x_2 , …, x_m, where the subscript *m* indicates the number of variables. The subscript *i* indicates any such variable. The variable u is the error term. It represents factors other than the independent variables that affect the dependent variable. We define the beta coefficients as β_1, β_2, …, β_m, where the subscript *m* indicates the number of coefficients. In general, the MLR takes the form:

$$Y = \beta_1 x_1, \ \beta_2 x_2, \ \ldots, \beta_m x_m$$

β_i is the slope parameter. It measures the expected change in the dependent variable for a unit change in x_i, all else equal. In other words, it explains the relationship between the dependent variable and the given independent variable when holding all other factors fixed. β_0 represents the intercept. It measures the expected value of the dependent variable when all independent variables are equal to zero, $x_i = 0$.

When adding more variables to the regression, it acts as additional controls for the previous SLR model. A MLR model is likely to give a better indication of what influences the covid incidence rate compared to the previously estimated model. Firstly, we control for multiple economic variables, this includes GDP per capita, education index, GINI index, and corruption. The MLR takes the form:

$$incidence\_rate = \beta_0 + \beta_1 fiveg + \beta_2 gdp\_per\_capita + \beta_3 education + \beta_4 GINI + \beta_5 corruption + u \qquad \text{(model.e2)}$$

We run the estimation of the parameters in Stata:

Table 6 - Relationship between the Covid incidence rate and 5G as well as economic variables

	Model.e1 **incidence_rate**	**Model.e2** **incidence_rate**
fiveg	0.467**	0.138
	(2.73)	(1.07)
GDP_per_capita		0.0165*
		(2.09)
education		36.53***
		(5.04)
GINI		-0.564
		(-0.05)
corruption		-8.496
		(-0.98)
_cons	1153.0***	-1119.9
	(10.27)	(-1.68)
Observations	181	143
R^2	0.040	0.395

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: regression model.e1 and model.e2 (see 5.1 for sources). Regressed the dependent variable, incidence_rate, on the independent variable, fiveg, for model.e1. Regressed the dependent variable, incidence_rate, on the independent variable, fiveg and all economic variables, for model.e2. t statistics in parentheses.

As we have included more variables, we wish to check their significance as well. In order to do so we use the p-value instead of a t-test. By using the p-value for a test it is possible to know the smallest significance level at which the null hypothesis would be rejected, given the observed value of the t statistic. The p-value is a probability and it is valued between 0 and 1. Small p-values suggest there is evidence against H_0 , while larger values suggest little evidence against it. In order to determine the significance, we look at the number of "stars" given in the regression table. These indicates the significance level for which we will reject the null hypothesis. It is given by * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The "stars" denotes: * significant at 5% level, ** significant at 1% level and *** significant at 0,1% level. These are all given for a two tailed test.

For the control variables coefficients, all other things being equal, the following holds:

- As GDP per capita increases by one dollar, we expect the incidence rate per hundred thousand to increase by 0.0165. The relationship is significant at 5% level, as indicated by $p < 0.05$.
- When the education index score increases by one point the incidence rate per hundred thousand is expected to increase by 36.53. The relationship is significant at 0,1% level, as indicated by $p < 0.001$.
- When the GINI index score increases by one point the incidence rate per hundred thousand is expected to decrease by 0.564. However, the relationship is not significant given $p < 0.05$.
- As the corruption score increases by one point the incidence rate per hundred thousand is expected to decrease by 8.496. However, the relationship is not significant given $p < 0.05$.

We see that the effect of 5G has continued to reduce in value, from 0.467 to 0.138, all else equal. The estimated effect that 5G has on the covid incidence rate has reduced. In addition, the R-squared has increased from 4% to 39.5%. This is a relatively large increase, indicating that this model better explains the changes in the covid incidence rate. We do however not know if the estimated effect still holds its significance and will therefore perform another t-test on β_1.

We have the same hypothesis that was formulated earlier:

$H_0: \beta_1 = 0$

$H_A: \beta_1 > 0$

We defined the rejection region using a 5% significance level. For a t_{137} the critical value is roughly 1.645. We will fail to reject H_0 since TS falls below the critical value, 1.07 < 1.645. This means that the data is compatible with a zero-relationship between 5G deployments and the covid incidence rate per hundred thousand, rather than suggesting that there is a positive relationship between the two.

These results suggest that the previously confirmed relationship between 5G and covid is spurious. Yet, a few of the estimations made in model.e2 raises some questions. For

instance, how an increase in the education score increases the covid incidence rate. One would assume that a more educated population would lead to a lower number of cases. Or how an increase in the GDP per capita leads to an increase in the covid incidence rate. Our theory is that these odd coefficents can be explained by other variables that this model does not include. Intuitively one can assume that a more educated, high income country will test more inhabitants. There is no way of proving covid unless the individual is being tested. Thus, the more a country tests, the more positive cases it will have. Given this theory of other important explanatory variables being omitted, we expand our data set further.

6.3 Investigating a causal relationship by including economic and corona related variables

We expand model.e2 by adding the following covid related variables, population density, tests per hundred thousand, stringency index and median age. We then obtain model.e3 which takes the following form:

$$incidence\_rate = \beta_0 + \beta_1 fiveg + \beta_2 gdp\_per\_capita + \beta_3 education + \beta_4 GINI + \beta_5 corruption + \beta_6 pop\_density + \beta_7 tests\_per1000 + B_8 stingency\_index + \beta_9 median\_age + u \quad \text{(model.e3)}$$

We run the estimation of the parameters in Stata:

Table 7 - Relationship between the Covid incidence rate and 5G, economic as well as corona related variables

	Model.e1 Incidence_rate	**Model.e2 Incidence_rate**	**Model.e3 Incidence_rate**
fiveg	0.467**	0.138	0.232
	(2.73)	(1.07)	(1.57)
GDP_per_capita		0.0165*	0.00411
		(2.09)	(0.37)
education		36.53***	11.27
		(5.04)	(0.77)
GINI		-0.564	27.29
		(-0.05)	(1.46)
corruption		-8.496	-16.35
		(-0.98)	(-1.42)
pop_density			-0.176
			(-0.24)
Tests_per1000			1.789**
			(3.26)

stringency_index			16.89 (1.79)
median_age			48.23 (1.89)
_cons	1153.0*** (10.27)	-1119.9 (-1.68)	-2777.2* (-2.33)
Observations	181	143	82
R^2	0.040	0.395	0.495

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Regression extended model.e1, model.e2 and model.e3 (see 5.1 for sources). *Regressed the dependent variable, incidence_rate, on the independent variable, fiveg, for model.e1. Regressed the dependent variable, incidence_rate, on the independent variable, fiveg and all economic variables, for model.e2. Regressed the dependent variable, incidence_rate, on the independent variable, fiveg, economic and corona related variables, for model.e3. t statistics in parentheses.*

We see that the R-Squared value has increased here as well, from 39.5% to 49.5%. Again, indicating that the model's explanatory power, in regard to the changes in the covid incidence rate, has increased.

In countries where all independent variables, 5G deployments, GDP per capita, population density, education, GINI, corruption, tests per hundred thousand, stringency index and median age, are equal to zero we expect the covid incidence rate per hundred thousand to be reduced by 2777.2. In the real world, these results are not possible as no country can have all the independent variables mentioned, equal to zero. Which explains the constant's negative value, as it is also not possible to have a negative incidence rate.

All other things being equal, the following holds:

- Whenever the 5G increases by one deployment, we expect the incidence rate per hundred thousand to increase by 0.232.
- As GDP per capita increases by one dollar, we expect the incidence rate per hundred thousand to decrease by 0.00411. However, the relationship is not significant given $p < 0.05$.
- When the population density increases by one additional citizen per km^2 the incidence rate per hundred thousand is expected to decrease by 0.176. However, the relationship is not significant given $p < 0.05$.

- When the education index score increases by one point the incidence rate per hundred thousand is expected to increase by 11.27. However, the relationship is not significant given $p < 0.05$.
- When the GINI index score increases by one point the incidence rate per hundred thousand is expected to increase by 27.89. However, the relationship is not significant given $p < 0.05$.
- As the corruption score increases by one point the incidence rate per hundred thousand is expected to decrease by 16.89. However, the relationship is not significant given $p < 0.05$.
- As the number of tests per hundred thousand increases by one additional test, the incidence rate per hundred thousand is expected to increase by 1.789. The relationship is significant at 1% level, as indicated by $p < 0.01$.
- When the stringency index score increases by one point the incidence rate per hundred thousand is expected to increase by 16.89. However, the relationship is not significant given $p < 0.05$.
- When the median age increase by one year the incidence rate per hundred thousand is expected to increase by 48.23. However, the relationship is not significant given $p < 0.05$.

These results do not support the claims of the conspiracy theory, as it suggests that the relationship between 5G density and the incidence rate is almost zero. More so we find that most independent variables have a more prominent relationship to the incidence rate than 5G density. We test if these results are statistically significant.

We have the same hypothesis that was formulated earlier:

$H_0 : \beta_1 = 0$

$H_A : \beta_1 > 0$

We defined the rejection region using a 5% significance level. For a t_{72} the critical value is roughly 1.658. We will fail to reject H_0 since TS falls below the critical value, $1.57 < 1.658$. Again, this suggests that the data is compatible with a zero relationship between 5G density and the incidence rate of covid cases per hundred thousand.

We have controlled for both economic and covid related variables in model.e3. In addition, we tested the significance of our results. Our finding suggests no relationship between 5G and the spread of covid and that these results are not significant enough to support the claims of the conspiracy theory. However, we also find that all other variables, except for the testing rate, are not significant in regard to the changes in the incidence rate. To investigate this, we will therefore preform an F-test on these variables (except for 5G and tests).

6.4 F-test on all independent variables except 5G and tests

An f-test is used to test whether a group of variables does not affect the dependent variable. For this type of test, we set up two models, a restricted model, and an unrestricted model. When removing the variables we want to test for, from the unrestricted model, we get the restricted model.

To find the F statistic, we need to adjust for the numerator- and denominator degrees of freedom. The numerator degrees of freedom is equal to the degrees of freedom in the restricted model minus the degrees of freedom in the unrestricted model. This number should be equal to the number of restrictions in the null hypothesis. This is denoted with *q*, where *q* is the number of restrictions. The F-statistic will decrease if we add more restrictions to our test, this is not a problem if the variables are truly significant.

$$q \ = \ numerator\ degrees\ of\ freedom \ = df_r - df_{ur}$$

The denominator degrees of freedom is equal to the number of observations (n) minus the number of slope parameters (k) minus the intercept (1).

$$n - k - 1 \ = \ denominator\ degrees\ of\ freedom \ = \ df_{ur}$$

The test is constructed around R-squared (it can also be constructed around the sum of squared residuals). The R-squared will decrease as we restrict the model. The subscript *ur* represents the unrestricted model, while *r* represents the restricted model. This gives the R-square version of the F-test:

$$F - stat = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} * \frac{n - k - 1}{q}$$

For the null hypothesis (H_0) we constitute several exclusion restrictions. If the null hypothesis is true, then the given variables do not affect the independent variable. This is a set of multiple restrictions because we are putting more than one restriction on the parameters. The null hypothesis puts *q* exclusion restrictions on the model. The alternative hypothesis (H_A) states that the null hypothesis is false. This means that at least one of the parameters listed in the null hypothesis is different from zero. We either reject or fail to reject the null hypothesis. According to the rejection rule, once the critical value has been obtained, we reject H_0 in favor of H_A at the chosen significance level, if F > c.

The crux of our research question is whether the observed relationship between 5G and covid is causal or spurious. At a 10%, 5% and 1% significance level, we therefore test the null hypothesis that GDP per capita, population density, education score, GINI, corruption score, stringency score and median age do not affect covid. For this F-test, H_0 states that all independent variables except for 5G and tests per hundred thousand do not affect covid. This gives us the alternative hypothesis, H_A, which states that these independent variables are related to the spread of covid. We perform the following test:

H_0: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_8 = \beta_9 = 0$

H_A: *not* H_0

We have five restrictions in the F-test, therefore q=7. According to Stata the unrestricted degrees of freedom is df_{ur} = 72. Due to having several missing observations in our sample, our unrestricted regression dropped them from the regression. When we omit these variables in the restricted regression, they are included in the regression as they no longer have missing variables. This leads to the restricted regression having 94 observations, compared to the unrestricted regression having 82 observations. Since we are using the R-squared to generate the F-statistic it is not critical to correct for the different number of observations. However, to make sure that this will not affect the result of our test we perform the correction in Stata by using the command drop for GDP per capita, education, GINI, corruption, and population density. This changes the R-squared for the restricted model from 0.3229 to 0.3442 (see appendix 6 and 7) The r-squared for the unrestricted model is 0.4950 (see table 7). We run the F-test in Stata:

Table 8 – f-test on all independent variables, except for 5G and tests

f- stat (6, 60)	3.07
Prob > F	0.0069

Note: f-test of the null hypothesis that H0: $\beta 2 = \beta 3 = \beta 4 = \beta 5 = \beta 6 = \beta 8 = \beta 9 = 0$ for model.e3. The f-test is performed in Stata, see appendix 8.

We find that the F-statistic is 7.34 and that the p-value is 0.0069. This means that we could reject the null hypothesis down to a 0,69% significance level. This implies that at standard significant levels of 1%, 5% and 10% we reject the null hypothesis and conclude that GDP per capita, population density, education score, GINI, corruption score, stringency score and median age have joint significance with regards to the incidence rate. Form this we cannot conclude that all the given variables are of no significance to our analysis, as they are all joint significant. We have now observed that other explanatory variables are significant in regard to the incidence rate. These results strengthen the theory that the previously confirmed relationship between 5G and covid is spurious.

7. Robustness of results

The MLR-prerequisites set strong restrictions for our model. This is also a data set that we have constructed ourselves, therefore we do more thorough inspections of its limitations. To find an unbiased estimate for the independent variables, the following conditions need to be satisfied. We will examine the robustness for our last expanded model, model.e3, as this is the model we base our conclusions on.

7.1 Linearity

The first condition requires linearity. For this bachelor thesis, we will assume that this assumption holds.

7.2 Random sampling

The second condition requires random sampling. In the data set, we are looking at the incidence rate at a country level. The data was collected from the John Hopkins map, where every country with publicly available data has been included in the map. This includes a sample of 82 countries where the total population is 195. It is therefore likely that some

regions or counties that have not been recorded might have outliers. Therefore, the remaining sample is not random, which implies that the second assumption might not hold.

7.3 Multicollinearity

The third assumption requires enough variation and no perfect collinearity. Here we will test for multicollinearity, we do so by using the Stata command corr. We find that there is no perfect collinearity between any of the variables, which we have when the correlation equals 1. The correlation between some variables is relatively high, yet none are equal to 1. See appendix 9 for correlation statistics.

We also use the variance inflation factor (VIF) to check for multicollinearity. Multicollinearity does reduce the statistical significance of the independent variables. A high VIF on an independent variable indicates a high collinear relationship to the other variables in the model. In general, a VIF value greater than 10 needs to be further discussed. 1/VIF value lower than 0.1 is comparable to a VIF of 10. We use Stata to find the variance inflation factors for model.e3. As the highest VIF value is 5.11 this suggests that none of the variables are possibly redundant. Overall, this shows that multicollinearity does not occur. We have not put in too many variables that measure the same thing.

Table 9 – Checking multicollinearity by using the variance inflation factor

	VIF	**VIF/1**
fiveg	1.18	0.850147
gdp_per_capita	5.11	0.195687
median_age	4.28	0.233779
education	4.83	0.207017
gini	1.62	0.618357
corruption	3.65	0.274133
pop_density	1.09	0.915460
srtignecy_index	1.27	0.789929
test_per1000	2.39	0.418026

Note: Table showcasing all VIF and VIF/1 values for all independent variables in model.e3. The test was performed in Stata.

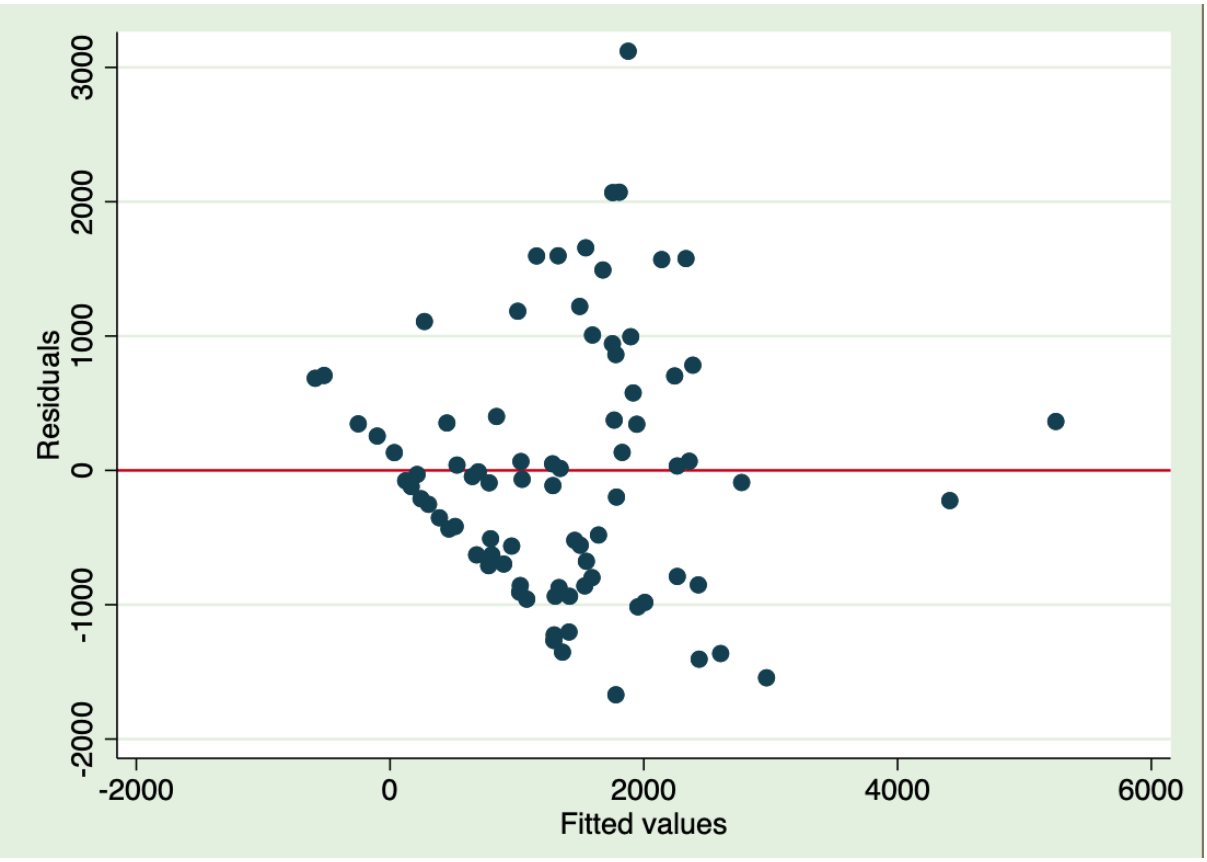
7.4 Zero conditional mean

The fourth assumption requires a zero conditional mean of errors, (E[u|x_i] = 0 which implies Cov(u,x)= 0). This means that the error term, u, has an expected value of zero, conditional on any values of the independent variables.

This assumption is rarely satisfied. We have tried to reduce this error, throughout the paper, by adding additional variables to the model, thus removing them from the error term. Yet, it is not unreasonable to think that some omitted variables are affecting 5G deployments, or the other control variables. This causes a violation of the zero conditional mean assumption. For example, how a country's geographic conformation affects its GDP or 5G. Mountains make it difficult to create roads and to develop economically, but also hard to build a good 5G network. Or how, where the covid test kits are produced affects a country's access to the test kits, hence the testing rate. This creates a correlation between the testing rate and the omitted variable, access to testing kits. So it is very unlikely that the fourth condition holds. Indicating that our least square estimators are somewhat biased.

This assumption is impossible to accurately test for. If the errors have a nonzero mean, when performing the regression, it would be absorbed by the constant. Meaning the residuals would on average be zero. We can therefore not test whether the residuals have a common mean that's not zero. However, we can get an indicator by checking whether the residuals, and by implication the errors that they estimate, have a constant mean. On average they would still be zero, but conditionally they may have a mean with some distance from zero. We check this by plotting residuals against the fitted values.

Figure 4 – Scatterplot of residuals against fitted values



Form figure.4 we can immediately see that the linearity assumption is rather suspect, and maybe some curved relationship is present. Indicating, again, that this assumption might not hold.

7.5 Homoscedasticity

The fifth assumption requires homoscedasticity. Which is defined as when observations of the residuals of a random variable are subtracted from a distribution that have constant variance. The opposite of homoscedasticity is heteroscedasticity. We have previously discussed how there is a large number of high-value outlier observations in our data set. This goes against the MLR prerequisites of homoscedasticity. It rather points to the existence of heteroskedasticity.

If the data has heteroscedastic noise it results in the formula for the standard deviation of the estimator being incorrect. For our conclusions on what affects the incidence rate to be correct, the error term needs to have a constant variance. It is also crucial to get the correct confidence interval and to perform hypothesis tests. If there is heteroskedasticity in our

data set this can, for example, point to an important explanatory variable(s) changing its significance. We will therefore test for heteroscedasticity in our model. There are several different ways to test for this, but we have chosen to perform a Breusch-Pagan test. The reason being that this can be done in Stata.

The Breusch-Pagan test method involves examining whether the variance in the noise joints from a linear regression is conditioned by the values in the independent variables. W_i denotes variables that we think may have an impact on the variance. a_i denotes constants.

$$\text{Var}(u_i)=f(a_1+a_2W_2+\ldots+a_mW_m)$$

We will test our extended model.e3, to be able to examine whether the conclusions we have drawn are correct or not:

H_0: $a_2 = a_3 = \ldots = a_m = 0$, here the residual joints is homoscedastic
H_A : $a_2 = a_3 = \ldots = a_m \neq 0$, here the residual joints is heteroscedastic

The Breusch-Pagan test gives the following test statistic with a chi-square distribution with (*m-1*) degrees of freedom. *m* denotes number of constants (*a*):

$$TS = nR^2 \sim X^2(m-1)$$

The Breusch-Pagan test finds predicted values ($\hat{Y}$) and residuals. Then the residuals are squared and rescaled so that the average is 1. The squared residuals are then regressed for $\hat{Y}$. If the null hypothesis is true, there is no heteroskedasticity and the test has a chi-square distribution with one degree of freedom.

We will test a chi-square distributed null hypothesis that the variance is constant and that we have homoscedastic noise, in Stata:

Table 9 – Testing for homoscedasticity as lin lin

chi²	4.12
Prob > chi²	0.0424

Note: The results of the Breusch-Pagan test performed in Stata to check for homoscedasticity in our lin lin model.e3. See appendix 10.

With a chi-square value equal to 4.12, we get a p-value of 0.0424 which means that we reject the null hypothesis to a 5% significance level and conclude that the variance is different. We thus test positive for heteroskedasticity, which indicates that the results found using OLS are invalid. As this is a breach of the prerequisite about constant variance in the residual term.

It could however be that part of this is the results of the presence of a few outliers in the incidence rate. To take this into account, we change our model to take the form of a log-lin model. Logs are convenient for transforming a highly skewed variable into a more normalized data set. This means that the incidence rate is now changed to show the estimated percentage change in cases per hundred thousand. We run a regression on the model.e3, as a log-lin, and perform another Breusch-Pagan test in Stata:

Table 10 – Testing for homoscedasticity as log lin

chi ²	0.79
Prob > chi ²	0.3753

Note: The results of the Breusch-Pagan test performed in Stata to check for homoscedasticity in our log-lin model.e3. See appendix 11.

With a chi-square value equal to 0.79, we get a p-value of 0.3753, which means that we no longer can reject the null hypothesis to a 5% significance level. Our data set is sufficiently homoscedastic, and the prerequisite is fulfilled.

7.6 Normality

The sixth assumption requires normality of the error term. This implies that the error term, u, is independent of the explanatory variables. In addition to u being normally distributed with a zero mean and variance of $\sigma^2 \sim N(0,\sigma^2)$. This assumption encompasses the fifth and fourth assumption as it is impossible to have a normally distributed error term if the error term was correlated with any explanatory variables, either in the error terms mean value or

in the error terms variance. Since the fourth assumption does not hold and the same goes for the fifth assumption given a lin-lin model, the sixth assumption will not hold either.

7.7 Robustness summary

It is rare having a model that fulfills all assumptions. If all assumptions are fulfilled it is considered the overall best estimator. This is not the case for our model, as assumptions four and six do not hold. However, this does not completely invalidate our findings. Seeing that our main point is not to examine the exact relationship between the variables, but to examine if 5G has a causal effect on the spread of covid.

8. Discussion and Limitations

8.1 Main results

In this paper, we have performed several tests on the conspiracy theory. The goal was to find whether 5G has a causal or spurious effect on covid. Our main finding suggests that there is no relationship between 5G and the spread of covid.

For our simple model.s1, we observed a positive connection between confirmed covid cases and 5G. We used a sample, our simple data set, to estimate something about the population. We found that with the inclusion of one more 5G deployment it is estimated an increase in confirmed covid cases of 753.8, all else equal. This result substantiates the conspiracy theory, as it suggests a prominent relationship between 5G deployments and confirmed covid cases. When performing a hypothesis test, at a 5% significance level, we observed a positive relationship between 5G and covid. These results were significant enough to support the claims of the conspiracy theory. What we did not know yet, was whether the relationship had a causal connection.

Since model.s1 did not consider different population sizes, we may have gotten a significant result due to the skewed data. We, therefore, set up a second SLR model that studied the relationship between the incidence rate of covid cases per hundred thousand and 5G deployments. As for model.s2 we found that with the inclusion of one more 5G deployment, it is estimated to decrease the incidence rate per hundred thousand by 3.487, all else equal.

When performing a hypothesis test, at a 5% significance level, the data did not suggest a positive relationship between 5G and covid.

Both model s1 and s2 are simple linear models that do not include other explanatory variables and have a low R-squared. This indicated that neither one of the models did explain every variation in the explained variable. By including more explanatory variables we could control for this. In addition, when not controlling for omitted variable bias we risked wrongly estimating a causal relationship between 5G and covid when it might be spurious. To give a more accurate analysis of the relationship, we therefore further extended our dataset to include more explanatory variables. In this data set, we look at every observation on a country level.

Since we changed our variables to a country level, we first did an SLR model regressing the covid incidence rate at a country level on 5G. We did this to make sure our previous results still held after making the changes. In addition, we wanted to be able to make a comparison of the data when expanding our model further. In model.e1 we found that there is an increase of 0.467 cases per hundred thousand for one additional 5G deployment in the country, all else equal. For this model we rejected the null hypothesis at a 1% significance level, and could therefore, not debunk the idea that 5G affects the spread of covid.

For our model.e2 we controlled the incidence rate for 5G and multiple economic variables. We did this in order to try and improve our models' explanatory power. The expansion included GDP per capita, education index, GINI index, and corruption index. When all other things remain equal, we found that the impact of 5G has reduced in value, from 0.467 to 0.138 from model.e1 to model.e2, all else equal. Using a 5% significance level we found that the data is compatible with a zero-relationship between 5G density and the incidence rate covid cases per hundred thousand, rather than suggesting that there is a positive relationship between the two.

For model.e3 we expanded model.e2 by adding the following covid related variables, population density, median age, tests per hundred thousand, and the covid stringency index. We found that most other independent variables have a more prominent relationship

to the incidence rate than 5G. Using a 5% significance level we found that the data is compatible with a zero-relationship between 5G and the covid incidence rate per hundred thousand. We saw that the R-Squared value had increased, from 39.5% to 49.5% for model e2 to model e3. This indicates that the model's explanatory power, in regard to the changes in the covid incidence rate, had increased. These results suggests that the previously confirmed relationship between 5G and covid is spurious, as we have now observed that other explanatory variables are significant in regard to the incidence rate.

Several of the independent variables for model.e3 were of no significance. We therefore decided to perform an f-test where we looked at the joint significance of these variables. At a significant level of 0,69% we would reject the null hypothesis and conclude that GDP per capita, population density, education score, GINI, corruption score, stringency score and median age have joint significance with regards to the incidence rate. The results strengthened the theory that the previously confirmed relationship between 5G and covid is spurious.

To summarize, we first demonstrated that there is a large and significant correlation. Then we investigated if this is spurious or causal. We did so in the final model, where we controlled for several explanatory variables. The results implied that there were other explanatory variables that had a greater impact on the incident rate. So, our main finding suggests that the relationship between 5G and covid is a spurious correlation.

8.2 Interpretations

We have spent the whole paper examining the relationship between covid-19 and 5G technology, and its limitations. Now the question is, how are our results useful?
As stated, we have found no causal relationship, but this is still a conspiracy theory that is widely believed. The issue is why and how these types of conspiracy theories get their foothold. From model.s1 and e.1, we saw how conspiracy theorists can force any real evidence to fit their theory. When expanding the model with control variables, we saw that a causal relationship was no longer observed. One has to thoroughly examine the relationship to understand why there is no causation between the two. Based on what we have examined in the paper we have seen that the tool used to make these theories

believable, is not to ameliorate for different population sizes or to omit important variables. So, the statistics aren't inherently wrong, but misleading.

This idea of misinterpreting correlation with causation is the main reason why we would propose that countries implement a politically independent statistics committee. Their responsibility would be to fact check public officials. Public officials should not be able to use their status or rank to spread information that can be misleading or inherently false. They are hired by the country's population and should be upheld to a higher standard when it comes to the information they choose to share, or use to implement political changes. When implementing this committee, it would be important that the committee is completely objective, and their focus should be solely on facts. This is important because conspiracy theories are generally hard to debunk in such a way that the individuals who believe in them stop believing. Any person who attempts to do so is seen as being part of the conspiracy. Meaning that the committee has to be believed, at least, by the majority of the population, to be effective. When achieving this, such committees can reduce the economic costs tied to conspiracies.

8.3 Ambiguities and limitations

Our data set can contribute to ambiguities of our analysis. We have used several different sources to collect the data. When doing this, we risk collecting data that might be biased or collected based on different conditions. It is known that several countries have wrongly reported the number of covid transmissions, in addition, there are a great number of unreported covid cases (Gettleman et al., 2021). When adjusting for this it could have an impact on the result of the analysis. We also used two different data sets. This can make it hard to compare the result of our analysis. We attempted to control for this by doing a regression on 5G and the incident rate at a country level so we can compare the data when including more explanatory variables.

The data set also limits the scope of our analysis. The number of observations varies for several variables. For instance, tests per hundred thousand and corruption have a great number of unreported results. When doing a regression, and there is a lack of data for a given variable, Stata will omit the given country. As a result, this will only give us an analysis

of the countries that have a value for all explanatory variables. Since model.e3 has over 30 observations this is considered sufficient. Yet this acts as a limitation to our analysis, as the countries that might have been omitted in our regression can cause different results. This could be true if the omitted countries have a relatively high or low incident rate. In addition, we could have included more explanatory variables in our data set. It is likely that other variables better explain the incident rate. This could for instance be mobility during covid, the number of smokers, or air pollution.

8.4 Continuance of research

To make a more accurate analysis, further research should focus on including every variable for every country. As mentioned, our data has more observations in high-income countries as their statistics are more widely accessible. This makes our results skewed. In theory, the effects could be different in low-income countries, especially if you consider the effect our control variables have on the spread of covid. It would also be beneficial to include other explanatory variables we have excluded. As mentioned under the fourth condition there might be a correlation between our explanatory variables and the error term. Like how access to testing kits affects the testing rate. When including these types of variables, further research will effectively remove the effect of the variable from the error term, and into the model, making any estimates more accurate. It would also be an idea to look at the effect given different continents. Creating dummy variables illustrating which continent the observed country is located in. This can be done to understand how different geographics might impact our results.

On a more general basis, one natural continuance of our research would be to examine which mediums are most commonly used when spreading conspiracy theories. This is in order to most efficiently reduce the stronghold that some conspiracies have. Learning where these theories come from, and where people believing in them have learned them, makes it easier to know how to reduce their believability. To find the true cost to society from these theories, it would also be natural to examine how often statistics are misused professionally. We know that this happens on a large scale among private individuals, but how many public figures use these types of misconstructions is unknown. One can assume this happens on a large scale in the political field.

9. Conclusion

In this paper, we have analyzed the conspiracy theory regarding 5G and covid. We first demonstrate that there is a large and significant correlation. Then we investigated if this is spurious or causal. We did so in the final model, e3, where we controlled for several explanatory variables. Here we estimated that whenever 5G increases by one deployment, we expect the incidence rate per hundred thousand to increase by 0.232, all else equal. These results were not significant, thereby not supporting the claims of the conspiracy theory. We therefore conclude that there is no link between 5G deployments and the incidence rate. The results rather suggest that the other explanatory variables have a greater impact on the incident rate. Our main finding suggests that the relationship between 5G and the spread of covid is a spurious correlation. By conveying this in a credible manner, people can understand why the conspiracy is not true. This showcases the importance of our findings, as it can help reduce the economic costs tied to corona conspiracies.

10. Sources

Gettleman, J., Yasir, S., Kumar, H., Raj, S. and Loke, A. (2021). As Covid-19 Devastates India, Deaths Go Undercounted. *The New York Times*. [online] 24 Apr. Available at: https://www.nytimes.com/2021/04/24/world/asia/india-coronavirus-deaths.html [Accessed 2 May 2021].

Guarino, B., Cha, A.E., Wood, J. and Witte, G. (2020). “The weapon that will end the war”: First coronavirus vaccine shots given outside trials in U.S.. *Washington Post*. [online] 14 Dec. Available at: https://www.washingtonpost.com/nation/2020/12/14/first-covid-vaccines-new-york/ [Accessed 1 Mar. 2021].

Hasell, J., Mathieu, E., Beltekian, D. et al (2020). *Coronavirus (COVID-19) Testing - Statistics and Research*. [online] Our World in Data. Available at: https://ourworldindata.org/coronavirus-testing [Accessed 30 Mar. 2021].

Johns Hopkins University (2021). *Johns Hopkins Coronavirus Resource Center*. [online] Johns Hopkins Coronavirus Resource Center. Available at: https://coronavirus.jhu.edu/map.html [Accessed 17 Mar. 2021].

NTB (2020). *Konspirasjonsteorier Om Korona Får Folk Til Å Sette Fyr På 5G-master*. [online] www.aftenposten.no. Available at: https://www.aftenposten.no/verden/i/opbP7W/konspirasjonsteorier-om-korona-faar-folk-til-aa-sette-fyr-paa-5g-master [Accessed 9 May 2021].

Ookla (2021). *Ookla 5G Map - Tracking 5G Network Rollouts Around the World*. [online] Speedtest.net. Available at: https://www.speedtest.net/ookla-5g-map [Accessed 11 Mar. 2021].

Our World In Data (2021). *COVID-19: Government Response Stringency Index*. [online] Our World in Data. Available at: https://ourworldindata.org/grapher/covid-stringency-index [Accessed 5 May 2021].

Qualcomm (2017). *What Is 5G | Everything You Need to Know About 5G | 5G FAQ*. [online] Qualcomm. Available at: https://www.qualcomm.com/5g/what-is-5g [Accessed 21 Apr. 2021].

Roser, M. and Ortiz-Ospina, E. (2013). *Income Inequality*. [online] Our World in Data. Available at: https://ourworldindata.org/income-inequality [Accessed 10 Mar. 2021].

Saybrook University (2020). *Conspiracy theories: a Booming Business*. [online] Unbound. Available at: https://www.saybrook.edu/unbound/conspiracy-theories-a-booming-business/ [Accessed 22 Mar. 2021].

Science Media Centre. (2020). *expert reaction to people who think 5G causes coronavirus | Science Media Centre*. [online] Available at: https://www.sciencemediacentre.org/expert-reaction-to-people-who-think-5g-causes-coronavirus/ [Accessed 28 Apr. 2021].

The Directorate-General for Communication (2020). *Identifying Conspiracy Theories*. [online] European Commission. Available at: https://ec.europa.eu/info/live-work-travel-

eu/coronavirus-response/fighting-disinformation/identifying-conspiracy-theories_en [Accessed 2 May 2021].

Thomas, R. L. (2005). *Using Statistics in Economics.* Berkshire: McGraw-Hill Education.

Tjernshaugen, A., Hiis, H., Bernt, J.F. and Braut, G.S. (2021). *koronavirus-pandemien 2020-2021*. [online] Store Norske Leksikon. Available at: https://sml.snl.no/koronavirus-pandemien_2020-2021 [Accessed 29 Apr. 2021].

Transparency International (2020). *Corruption Perceptions Index 2019*. [online] Transparency.org. Available at: https://www.transparency.org/en/cpi/2020/index/nzl [Accessed 1 May 2021].

UNESCO (2020). *ThinkBeforeSharing - Stop the Spread of Conspiracy Theories*. [online] UNESCO. Available at: https://en.unesco.org/themes/gced/thinkbeforesharing [Accessed 5 May 2021].

United Nations Development Programme (2020). *Human Development Reports*. [online] Human Development Reports. Available at: http://hdr.undp.org/en/indicators/103706 [Accessed 2 Apr. 2021].

11. Appendix

Appendix 1 - An overview over all variables included in the simple data set with short explanations.

fiveg	Number of 5G deployments in the observed area
confirmed	Total confirmed cases for the observed area
incidence_rate	Total confirmed cases per 100 000 inhabitants, for the observed area

Appendix 2 - Number of countries with following regions for confirmed cases and the incidence rate of covid. Note, if the number of regions is equal to 1 we look at the incident rate/confirmed cases at a country level.

Country	**Number of regions**
Afghanistan	1
Albania	1
Algeria	1
Andorra	1
Angola	1

Antigua and Barbuda	1
Argentina	1
Armenia	1
Australia	6
Austria	1
Azerbaijan	1
Bahamas	1
Bahrain	1
Bangladesh	1
Barbados	1
Belarus	1
Belgium	11
Belize	1
Benin	1
Bhutan	1
Bolivia	1
Bosnia and Herzegovina	1
Botswana	1
Brazil	6
Brunei	1
Bulgaria	1
Burkina Faso	1
Burundi	1
Cambodia	1
Cameroon	1
Canada	7
Cape Verde	1
Central African Republic	1
Chad	1
Chile	16
China	29
Colombia	1
Comoros	1
Congo	1
Costa Rica	1
Croatia	1
Cuba	1
Cyprus	1
Czech Republic	1
Democratic Republic of the Congo	1
Denmark	1
Djibouti	1

Dominica	1
Dominican Republic	1
Ecuador	1
Egypt	1
El Salvador	1
Equatorial Guinea	1
Eritrea	1
Estonia	1
Ethiopia	1
Fiji	1
Finland	1
France	1
Gabon	1
Gambia	1
Georgia	1
Germany	16
Ghana	1
Greece	1
Grenada	1
Guatemala	1
Guinea	1
Guinea-Bissau	1
Guyana	1
Haiti	1
Honduras	1
Hungary	1
Iceland	1
India	35
Indonesia	1
Iran	1
Iraq	1
Ireland	1
Israel	1
Italy	20
Ivory Coast	1
Jamaica	1
Japan	26
Jordan	1
Kazakhstan	1
Kenya	1
Kosovo	1
Kuwait	1

Kyrgyzstan	1
Laos	1
Latvia	1
Lebanon	1
Lesotho	1
Liberia	1
Libya	1
Liechtenstein	1
Lithuania	1
Luxembourg	1
Madagascar	1
Malawi	1
Malaysia	1
Maldives	1
Mali	1
Malta	1
Marshall Islands	1
Mauritania	1
Mauritius	1
Mexico	32
Moldova	1
Monaco	1
Mongolia	1
Montenegro	1
Morocco	1
Mozambique	1
Myanmar	1
Namibia	1
Nepal	1
Netherlands	12
New Zealand	1
Nicaragua	1
Niger	1
Nigeria	1
Norway	1
Oman	1
Pakistan	7
Panama	1
Papua New Guinea	1
Paraguay	1
Peru	25
Philippines	1

Poland	1
Portugal	1
Puerto Rico	74
Qatar	1
Republic of North Macedonia	1
Romania	1
Russia	83
Rwanda	1
Saint Kitts and Nevis	1
Saint Lucia	1
Saint Vincent and the Grenadines	1
San Marino	1
São Tomé and Príncipe	1
Saudi Arabia	1
Senegal	1
Serbia	1
Seychelles	1
Sierra Leone	1
Singapore	1
Slovakia	1
Slovenia	1
Solomon Islands	1
Somalia	1
South Africa	1
South Korea	1
South Sudan	1
Spain	19
Sri Lanka	1
Sudan	1
Suriname	1
Swaziland	1
Sweden	18
Switzerland	1
Syria	1
Taiwan	1
Tajikistan	1
Tanzania	1
Thailand	1
Timor-Leste	1
Togo	1
Trinidad and Tobago	1
Tunisia	1

Turkey	1
Uganda	1
United Arab Emirates	1
United Kingdom	4
United States	1700
Uruguay	1
Uzbekistan	1
Vanuatu	1
Vatican City	1
Venezuela	1
Vietnam	1
Yemen	1
Zambia	1
Zimbabwe	1

Appendix 3 – Results regression model.s1

confirmed	**Coef.**	**Std. Err.**	**t**	**P>|t|**	**[95% Conf. Interval]**	
fiveg	753.6994	41.78418	18.04	0.000	671.761	835.6378
_cons	16999.7	2078.098	8.18	0.000	12924.57	21074.83

Appendix 4 - An overview over all variables included in the expanded data set with short explanations.

fiveg	Number of 5G deployments in each country
incidence_rate	Total confirmed cases per one million inhabitants, for the observed country
gdp_per_capita	The observed country's GDP per capita
pop_density	The observed country's population density
education	The observed country's score in UN's education index
median_age	The observed country´s median age
stringency_index	A score between 0 and 100, representing the government corona response
GINI	The observed country's GINI index
tets_per1000	The observed country's number of citizens, per hundred thousand, who have gotten tested for covid

corruption	A score between 0 and 100, on the corruption in the observed country.

Appendix 5 - Countries with number of 5G deployments.

Country	5G deployments
Afghanistan	0
Albania	0
Algeria	0
Andorra	0
Angola	0
Antigua and Barbuda	0
Argentina	0
Armenia	0
Australia	71
Austria	1184
Azerbaijan	0
Bahamas	0
Bahrain	49
Bangladesh	0
Barbados	0
Belarus	0
Belgium	83
Belize	0
Benin	0
Bhutan	0
Bolivia	0
Bosnia and Herzegovina	0
Botswana	0
Brazil	12
Brunei	0
Bulgaria	28
Burkina Faso	0
Burundi	0
Cambodia	0
Cameroon	0
Canada	197
Cape Verde	0
Central African Republic	0
Chad	0
Chile	0
China	103

Colombia	1
Comoros	0
Congo	0
Costa Rica	0
Croatia	17
Cuba	0
Cyprus	0
Czech Republic	427
Democratic Republic of the Congo	0
Denmark	9
Djibouti	0
Dominica	0
Dominican Republic	0
Ecuador	0
Egypt	0
El Salvador	0
Equatorial Guinea	0
Eritrea	0
Estonia	6
Ethiopia	0
Fiji	0
Finland	155
France	923
Gabon	0
Gambia	0
Georgia	0
Germany	4313
Ghana	0
Greece	18
Grenada	0
Guatemala	0
Guinea	0
Guinea-Bissau	0
Guyana	0
Haiti	0
Honduras	0
Hungary	22
Iceland	1
India	0
Indonesia	0
Iran	0
Iraq	0

Ireland	314
Israel	80
Italy	705
Jamaica	0
Japan	75
Jordan	0
Kazakhstan	0
Kenya	0
Kuwait	226
Kyrgyzstan	0
Laos	1
Latvia	7
Lebanon	0
Lesotho	0
Liberia	0
Libya	0
Liechtenstein	0
Lithuania	0
Luxembourg	6
Madagascar	2
Malawi	0
Malaysia	0
Maldives	6
Mali	0
Malta	0
Marshall Islands	0
Mauritania	0
Mauritius	0
Mexico	0
Moldova	0
Monaco	4
Mongolia	0
Montenegro	0
Morocco	0
Mozambique	0
Myanmar	0
Namibia	0
Nepal	0
Netherlands	1242
New Zealand	15
Nicaragua	0
Niger	0

Nigeria	0
Norway	12
Oman	51
Pakistan	0
Panama	0
Papua New Guinea	0
Paraguay	0
Peru	0
Philippines	54
Poland	167
Portugal	0
Qatar	26
Romania	30
Russia	0
Rwanda	0
Saint Kitts and Nevis	0
Saint Lucia	0
Saint Vincent and the Grenadines	0
Samoa	0
San Marino	0
Saudi Arabia	81
Senegal	0
Serbia	0
Seychelles	2
Sierra Leone	0
Singapore	66
Slovakia	1
Slovenia	47
Solomon Islands	0
Somalia	0
South Africa	16
South Korea	0
South Sudan	0
Spain	198
Sri Lanka	0
Sudan	0
Suriname	0
Sweden	66
Switzerland	753
Syria	0
Tajikistan	1
Tanzania	0

Thailand	452
Togo	1
Trinidad and Tobago	0
Tunisia	0
Turkey	0
Uganda	0
United Arab Emirates	14
United Kingdom	409
United States	7337
Uruguay	0
Uzbekistan	0
Vanuatu	0
Venezuela	0
Vietnam	0
Yemen	0
Zambia	0
Zimbabwe	0
Kosovo	0
Taiwan	110

Appendix 6 – Restricted model for f-test

Source	SS	df	MS	Number of obs	=	94
				F(2, 91)	=	21.70
Model	60991315.3	2	30495657.7	Prob > F	=	0.0000
Residual	127872693	91	1405194.43	R-squared	=	0.3229
				Adj R-squared	=	0.3081
Total	188864009	93	2030795.79	Root MSE	=	1185.4

Incident_rate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fiveg	.3467704	.1606414	2.16	0.034	.027676	.6658649
tests_per1000	2.04281	.344798	5.92	0.000	1.357911	2.727709
_cons	881.2378	151.046	5.83	0.000	581.2036	1181.272

Appendix 7 – Restricted model for f-test with dropped variables

Source	SS	df	MS	Number of obs	=	82
				F(2, 79)	=	20.74
Model	50359815.7	2	25179907.8	Prob > F	=	0.0000
Residual	95933582.7	79	1214349.15	R-squared	=	0.3442
				Adj R-squared	=	0.3276
Total	146293398	81	1806091.34	Root MSE	=	1102

Incident_rate	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fiveg	.3327301	.1505235	2.21	0.030	.0331205	.6323397
tests_per1000	2.196671	.3917979	5.61	0.000	1.416817	2.976526
_cons	836.9478	148.7718	5.63	0.000	540.825	1133.071

Appendix 8 – F-test

```
( 1)  pop_density = 0
( 2)  gdp_per_capita = 0
( 3)  Eduaction = 0
( 4)  GINI = 0
( 5)  Corruption = 0
( 6)  Stignecy_index = 0
( 7)  Median_age = 0

       F(  7,     72) =    3.07
            Prob > F =    0.0069
```

Appendix 9 – correlation all variables model.e3

	Incide~e	gdp_pe~a	pop_de~y	Eduact~n	GINI	tes~1000	Corrup~n	fiveg	Stigne~x	Median~e
Incident_r~e	1.0000									
gdp_per_ca~a	0.4928	1.0000								
pop_density	-0.0589	-0.0009	1.0000							
Eduaction	0.5384	0.6659	-0.1245	1.0000						
GINI	-0.2127	-0.4378	-0.0977	-0.4332	1.0000					
tests_p~1000	0.5511	0.7569	-0.0213	0.5027	-0.3537	1.0000				
Corruption	0.3808	0.8115	-0.0473	0.6938	-0.3866	0.6084	1.0000			
fiveg	0.2886	0.3035	-0.0613	0.1897	-0.0060	0.1632	0.2088	1.0000		
Stignecy_i~x	0.3630	0.1785	0.1069	0.3388	-0.1963	0.1713	0.0904	0.1659	1.0000	
Median_age	0.5114	0.5726	-0.0487	0.8426	-0.5569	0.4432	0.6404	0.1470	0.2792	1.0000

Appendix 10 – Breusch-Pegan test on lin-lin model

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of Incident_rate

         chi2(1)      =     4.12
         Prob > chi2  =   0.0424
```

Appendix 11 - Breusch-Pegan test on ln-log model

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lnIncident_rate

         chi2(1)      =     0.79
         Prob > chi2  =   0.3753
```

NTNU

Norwegian University of
Science and Technology