Evaluating the signal power of predictor variables for recorded COVID deaths

Jamie Mariani, CFA, December-2021

***Abstract*—Pre meaningful data on the Omicron variant, reported COVID cases (smoothed) appear to be the key predictor of reported COVID deaths (smoothed) in the UK, US and World. Other meaningful health related independent variables include the number of hospital patients, the positive rate, total vaccinations and the prevalence of male smokers in the population. Prior to the first lockdown, the general public's risk appetite collapsed, observable in negative year over year growth in seated dining reservations. The strong negative correlation evident between seated dining growth in the UK and COVID cases in the UK suggests the general public were already taking significant steps to minimize their risk of infection pre government imposed lockdown. In the sample period there is no meaningful correlation apparent between stock market valuations and COVID cases. This likely reflects other variables, e.g. monetary and fiscal policy, beyond the scope of this study.**

I. Introduction

The COVID pandemic is the crisis of our time. Using a diverse range of data sources, this report investigates interconnections between deaths caused by the virus (the dependent variable) and a variety of predictor (independent) variables, including the number of reported cases.

The analysis starts looking for patterns evident within government provided health data across three distinct regions, the UK, the US and the World. Having determined the key predictor variables of reported deaths in the COVID health data, the focus turns to the full service dining sector. This industry has been hugely disrupted by the spread of the virus and associated government led lockdowns.

Time series demand for seated dining vs reported COVID cases is evaluated, in particular the predictive power of seated dining demand vs. reported COVID cases. A key question is whether wild fluctuations in COVID cases can be linked or even predicted by certain factors, in particular dining activity in full service restaurants ('The Wisdom of Crowds') [1]?

Finally this report looks at public equity valuations in the UK, US and World (in aggregate) during the COVID pandemic and the relationship with COVID cases reported. This recognises equity markets perform vital functions in market based economies; providing long-term capital to businesses, short-term liquidity where required and attractive long-term returns for asset owners.

Better understanding of these correlations could 1) provide constructive input into future public policy decisions, 2) offer the basis for an investment strategy for those seeking profit and 3) highlight a new anomaly in the 'Efficient Market Hypothesis' for academics (the EMH is a central theoretical plank of finance teaching) [2].

II. Analytical questions

Based on government reported health data, this report seeks to answer what predicts daily covid deaths? The signal power of a wide range of predictor variables is considered including seated restaurant daily demand data in the face of the COVID-19 pandemic, and daily public equity market valuations.

Three distinct geographic regions, the US, the UK and the Globe are considered, to ensure a range of different 'lens' are employed.

A further question this analysis raises is whether mandated government intervention (in the form of lockdowns) was required? Or whether the 'wisdom of crowds', as evidenced by seated diner daily demand, already suggested a significant change in social behaviour was unfolding, in direct response to reported COVID cases and deaths [1].

III. Data materials

For this project, three distinct datasets were used, from a UK, US and Global perspective:-

- Daily corona virus cases (infections) & deaths (US, UK, World), 18/2/2020 – 22/11/2021

  From the time-series data provided by 'Our World In Data' data [3]. Key variables tracked across geographies include total cases, new cases, total deaths, new deaths, cases & deaths per million, the reproduction rate, hospital patients, vaccinations, demographics and prevalence of diabetes and smoking in the population. This provides a rich data-set to consider the interconnections between deaths caused by the virus and a variety of 'health' predictor variables

- Daily seated diner data (same regions and dates)

  From the time-series 'State of the Industry' data from Open Table [4]. This tracks daily demand for full seated dining by region year-over-year for 2020 over 2019, and 2021 also over 2019 (given the base number in 2020 was often 0). It provides a fascinating insight into consumer behaviour and appetite for risk by region pre and post lockdowns. It allows us to investigate the relationship between COVID deaths and consumer demand and consider the signal power 'crowd' demand offers as a predictor of COVID deaths

- Daily equity market valuations for the S&P500 (US), FTSE-100 (UK) and MSCI ACWI (World)

  Public equity market valuations are from Bloomberg. Tracking daily changes in valuation, on a constant currency basis (US$'s), provide an insight into how the world's investors perceived the risk from COVID. The S&P 500 is a broad US equity market index. The FTSE100 contains the largest 100 companies in the UK. The MSCI All-Country-World-Index comprises near 3,000 companies across 50 countries. Market valuation data can be evaluated relative to daily corona virus cases and deaths.

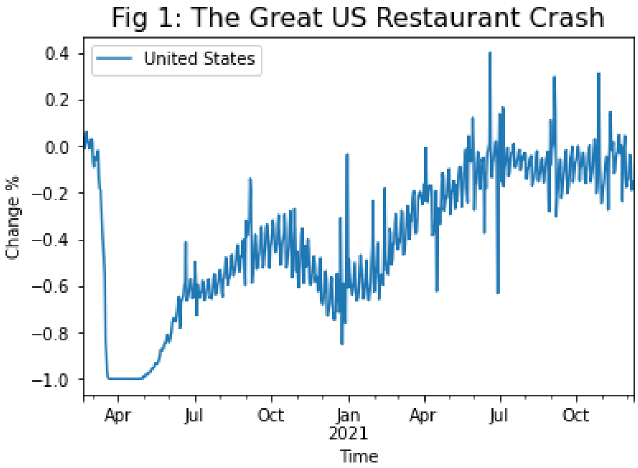
IV. Analysis

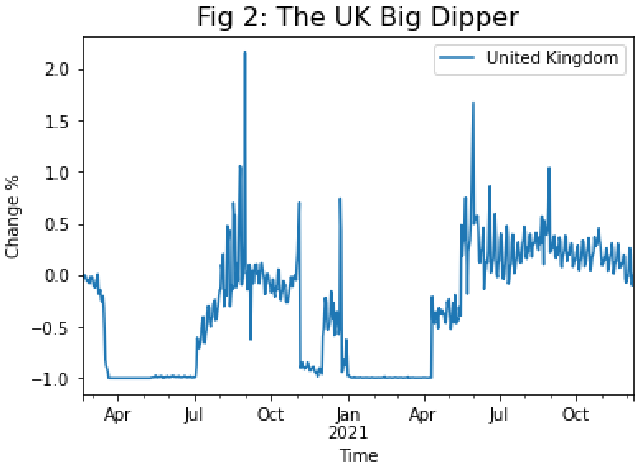
A. *Data*

The Open Table 'State of the Industry' set was the most challenging. It contained a significant number of countries and regions (cities) beyond the intended scope of this study.

The first step was to determine the number of redundant rows and drop these. The second step was to delete the first column ('Type'), which was also superfluous. The third step was to transpose the data (at source, countries/regions in the Open Table data are presented in columns, the percentage change year-over-year (YoY) in seated diners in rows).

When cleaned, a precipitous decline (year over year) in seated diners in the US (Fig. 1) and UK (Fig.2) is evident, reaching a bottom in March-2020 ('lockdown'). UK seated dining behaviour is noticeably more volatile than the US.





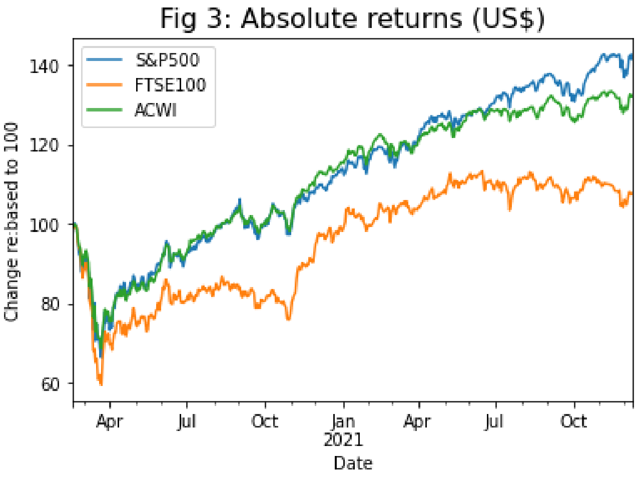
The relative surge in Seated Diner Data growth in summer 2020 in the UK (vs. the US) is likely explained by i) the lifting of lockdown restrictions (June-20) and ii) the infamous 'Eat out to help out' strategy of subsidised dining launched by UK Chancellor Rishi Sunak in August-2020.

Subsequently, the UK was forced to lock down again, whereas seated dining was still possible in the US and across the rest of the world (as tracked by Open Table).

The COVID data from OWID and the Index valuations from Bloomberg required less manipulation. For the COVID data the primary change was removing redundant data (post feature selection) and replacing the 'NaN' (Not a Number) readings from the dataframe with a zero, using a fillna (0) command; both essential for regression analysis.

For the Index data from Bloomberg, the primary change made was re-naming the column headings from region (e.g. UK) to the local public equity index (e.g. FTSE100).

Fig. 3 shows the valuation (in US$'s) of three major equity indices; the S&P500, the FTSE100 and finally the MSCI ACWI during the sample period. A stark visual contrast to the seated dining data (Fig's 1 & 2) is already apparent.

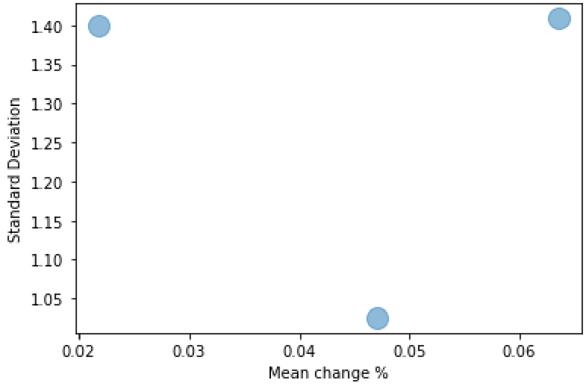


From Fig. 3 it is observable that the US market (the S&P500) provided the best absolute return over the sample period, at over >40%. It was closely followed by the world market (ACWI), with a return just over ~30%. The UK (the FTSE100) was a relative laggard, returning just ~8%.

This raises an interesting question about 'home bias'. Home bias refers to the tendency of equity investors to favour investing in domestic stocks over investing in the stocks of foreign companies. The benefits of international diversification appear clear during the COVID crisis.

This becomes more pronounced when the absolute return is considered vs. the standard deviation in returns (Fig. 4). As a measure of volatility, the standard deviation can be viewed as a proxy for risk borne by investors owning each index. During the sample period the standard deviation for the US and UK indices was similar, the world was lower.

Fig 4: Mean daily return yoy vs. standard deviation



In financial markets, risk and return are normally positively correlated [5]. For any additional increment of 'risk' an investor is prepared to take, there should be a corresponding increment of 'reward'.

Given that the mean return was significantly higher for the US market than the UK market, it is somewhat surprising investors in the UK market had to bear a similar level of volatility to those invested in the US.

B. Construction of models

Before slicing and dicing the COVID data-set further, the study considered the correlation between the available variables, to aid feature selection in the subsequent analysis.

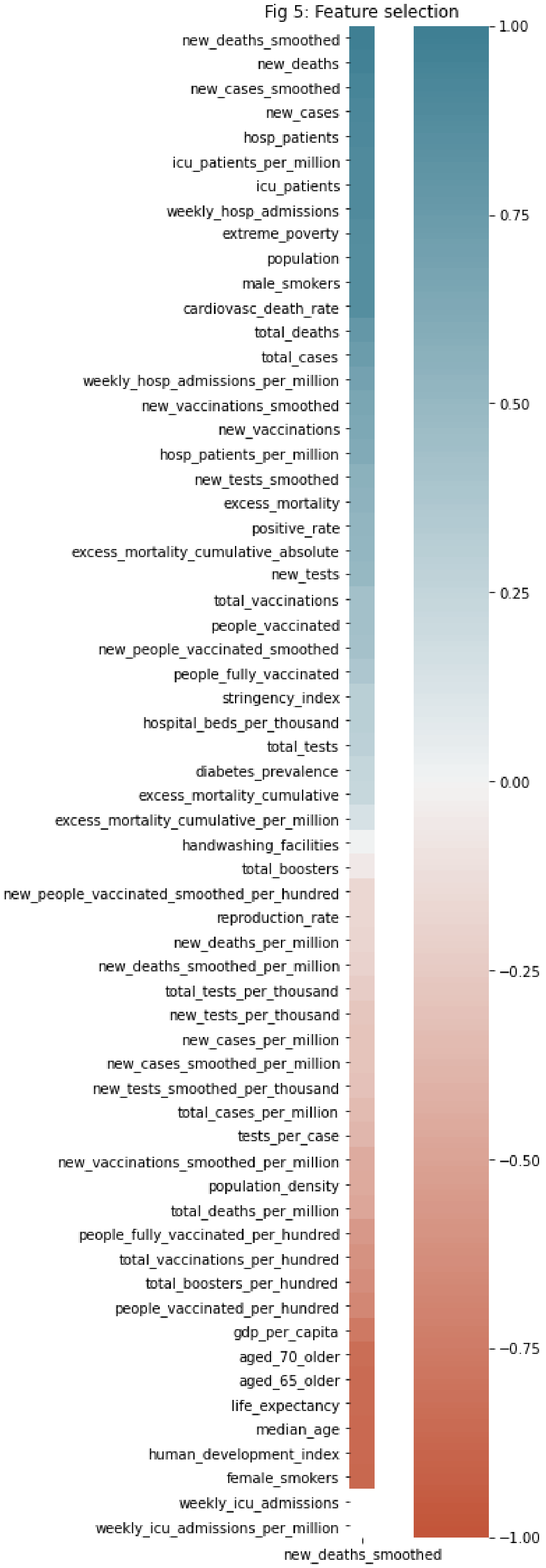
As a starting point, the approach adopted by John Burn-Murdoch, the FT's data scientist, was used to aid feature selection [6]. Initially, 'new cases per million' were considered as the key predictor variable. It was with some surprise that it didn't appear to have a particularly strong correlation with any of the other COVID variables.

The correlation to 'total boosters per hundred' (r = 0.69) was a little shocking, potentially suggesting that as more of the population received booster vaccinations, confidence in meeting together in high risk situations increased, resulting in substantial new case growth.

Through a process of trial and error, it was found that 'New deaths smoothed' (for a 7 day rolling average) had high correlations with a much wider range of metric (Fig.5). The positive correlations to new cases smoothed (r = 0.94), hospital patients (r = 0.91) and ICU patients (r = 0.88) were particularly noteworthy, as were the negative correlations to median age (r = -0.86), amongst other variables.

From COVID deaths, new cases detected are unmistakably a key driver. Hospital admissions and ICU patients are also manifestly important to policy makers, health professionals and the general public (e.g. the UK government’s 'Protect the NHS' slogan). Arguably, hospital usage data indicates how harmful the virus is at any given stage. By definition ICU patients are in poor health.

Fig 5: Feature selection



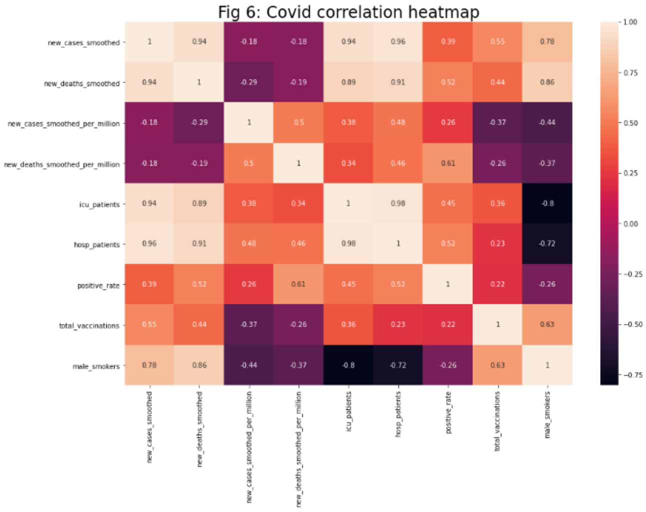
This prompted a more detailed review of hospital usage data. What was subsequently surprising was the weak correlation evident between 'icu patients per million' and 'new cases per million' (r = 0.46) for the three regions.

The COVID data suggests the number of icu patients per million is much more closely correlated with the general number of hospital patients (r = 0.95) & weekly hospital admissions (r = 0.92). 'Extreme poverty' (r = 0.61), 'diabetes prevalence' (r = 0.61) and 'gdp_per_capita' (r = 0.61) also show strong positive correlation to icu patients per million.

In other words, for a given number of hospital admissions in any period in the sample, health professionals could be allocating available ICU capacity (beds), regardless of COVID dynamics; build it and they will come.

That said, the icu patients per million variable still shows a high level of correlation with 'new deaths smoothed' (r = 0.89), and new cases smoothed (r = 0.88). Following the removal of redundant columns, a correlation matrix (Fig. 6) was produced to inform final feature selection.

Fig 6: Covid correlation heatmap

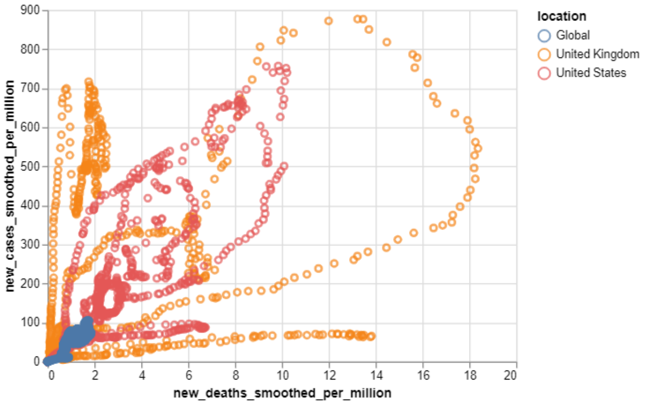


As a result, this study analysed the relationship between new covid cases and new covid deaths (both smoothed) across the three regions (UK, US and World).

As predicted by the prior study of correlations, the 'Daily COVID cases vs. deaths' scatterplot (Fig. 7) shows a strong positive correlation between new cases and new deaths. The per million data is used to normalise for population size.

It's clear from the per million data that the UK and US have had a rough time with new COVID cases and death vs. the world. We can also see that in total reported cases per million (158k in the UK, 150k in the US vs. 34k for the World) and total reported deaths per million (2.2k in UK, 2.4k in US v. only 0.7k for the World).

Fig 7: Daily COVID cases vs. deaths



Part of this maybe due to demographics and the underreporting of cases and deaths in less developed economies. The latter could possibly be a result of (i) lack of resources, or (ii) political malfeasance (e.g. in China), a discussion of which is beyond this study's scope.

To further understand the COVID data, a simple linear regression was constructed with new cases and new deaths.

This study also considered whether the explanatory power of the simple linear regression could be improved with additional independent variables (a multiple regression). The multiple regression model incorporated five independent variables; i) new cases smoothed, ii) hosp patients, iii) positive rate, iv) total vaccinations and v) male smokers. Findings are presented in V.

Having inspected each of the three data sets separately, the next modelling stage was to connect them using multiple regression, seeking out variables predictive of new cases.

Initially this study addressed whether the dining data offered incremental insight, in particular in the early stages of the COVID pandemic period (before the first lockdown), as a predictor variable for new cases?

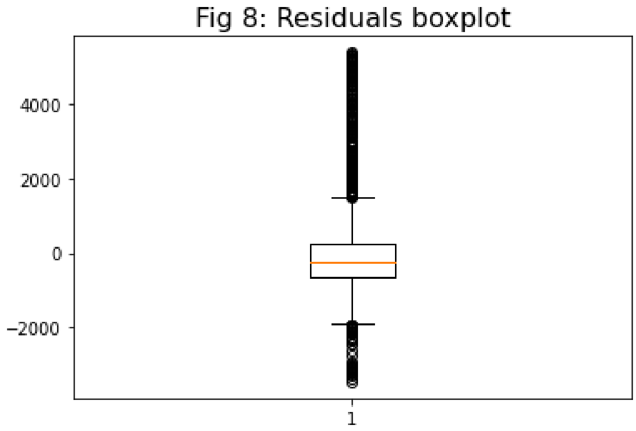
Finally, from a capitalistic lens, this study considered the relationship between new cases and stock market valuations. Put another way, could stock market valuations be used as a predictor variable for new cases?

C. *Validation of results*

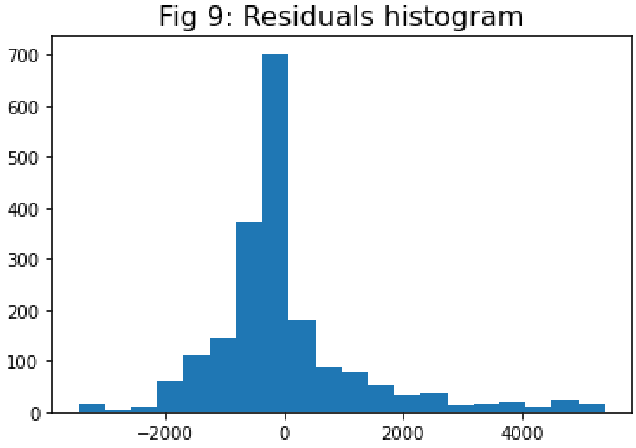
The impression gained from the 'Daily COVID cases vs deaths' scatterplot (Fig. 7) is backed up by the calculation of Pearson's correlation (r= 0.94), with a p-value close to 0.0. Spearman's correlation (at r = 0.92, p-value of 0.0) was slightly lower, suggestive of a linear relationship between the two variables.

The p-values are the probabilities that correlation of the population is not correlated (i.e. with a null-hypothesis that the population are not correlated). With a p value of 0.0 in both calculations, the null hypothesis is rejected, the relationship is statistically significant.

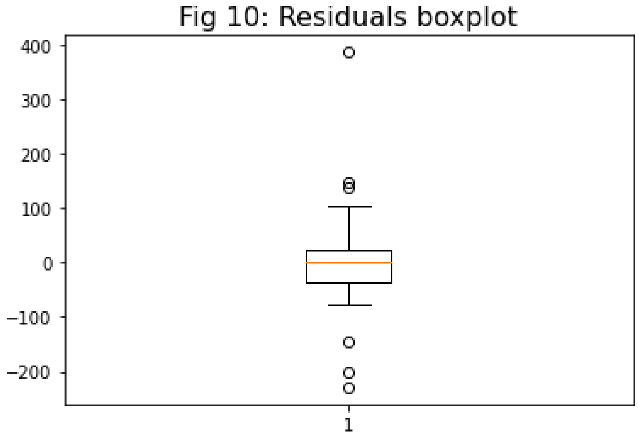
For the simple linear regression of new cases smoothed and new deaths smoothed, this study calculated the residuals (observed vs predicted values). The boxplot (Fig. 8) from the linear regression of new cases and new deaths shows a large number of outliers, which is somewhat unexpected given the r^2.



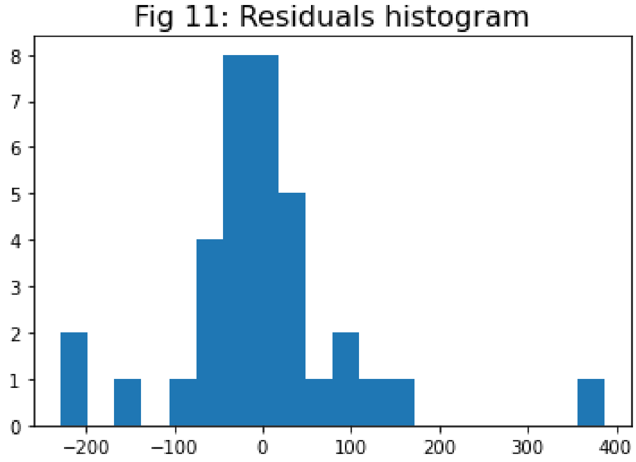
The histogram suggests the residuals are broadly normally distributed (Fig. 9). This is important. It underpins the validity of using this model.



For the small sample of UK dining data used as a predictor of new cases, the boxplot (Fig. 10) shows a small number of outliers, which is expected given the high r2.



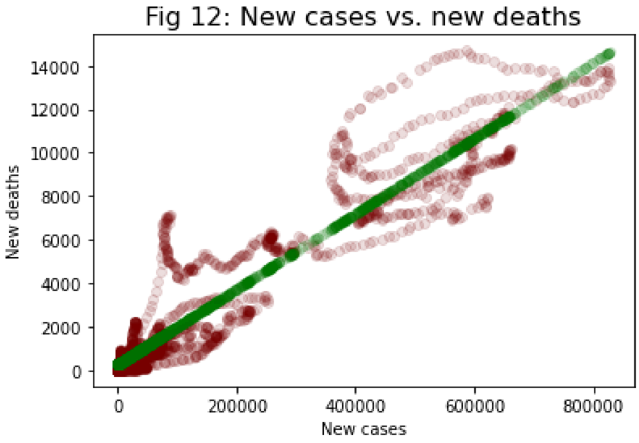
The histogram (Fig. 11) suggests the residuals approximate towards a normal distribution, with some gaps. This is likely due to the relatively small number of datapoints.



V. Findings, Reflections and Further Work

A. COVID models

The simple linear regression, using new cases as a predictor variable for new deaths produced an r^2 of 0.89 (Fig. 12). This is a very high level. For the sample period, just under ~90% of the variation in new deaths can be explained by new cases.



Given the findings of this simple linear regression, politicians, civil servants, health workers, the media and the general public appear justified in their focus on new case numbers as a key data point to monitor.

A known unknown is whether the pattern observed in the sample holds with new COVID variants (e.g. Omicron), given past studies of pandemic waves (pointing to increased contagion, with lower levels of virulence) [7].

In the multiple regression, the addition of independent variables increased the r^2 to ~0.94. Whilst that is an impressive level of variation explained, particularly given no pruning of outliers, the level of improvement from the linear regression was ultimately modest.

B. *Dining models*

The linear regression performed on seated dining data as a predictor of new cases was disappointing. In aggregate there appeared to be no meaningful pattern (r^2 near 0).

This prompted an important clarification; had lockdowns disrupted the relationship (e.g. caused a complete loss of seated diners, with cases still going up)?

To check, this study re-ran the seated dining linear regression, on a smaller sub-set of the data focused on the UK from the beginning of the time series in February to the 23rd March, when the first lock down in the UK was announced by the government (Fig. 13).

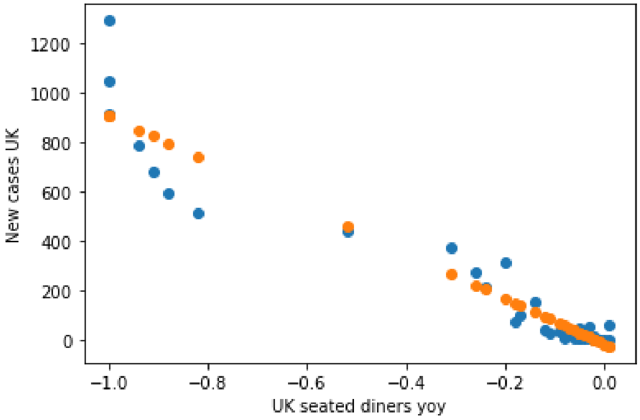


Fig 13: UK diners yoy vs. uk new cases

Intriguingly this new analysis suggested a very high negative correlation between year-over-year seated diner growth in the UK and new cases ($r = -0.95$, $r^2 = 0.91$) during this early stage of the pandemic.

In other words, absent the distortion presented by the lockdowns, there was a very strong relationship between these variables. This finding could suggest that fear of the virus began to grip the UK population in February and March 2020, resulting in a huge slump in demand for seated dining reservations, at the same time new cases (smoothed) in the UK were ramping.

Intuitively this feels rational, and is suggestive of the ‘wisdom of crowds’ [1]. Arguably it has implications for policy makers looking to control behaviour.

If dining data were to become available for a country like Sweden, where no government imposed lockdown was implemented, a natural evolution of this study would be to compare the findings with this initial study.

C. *Stock markets*

Unfortunately for policy makers, no significant patterns were evident looking across the three regions during the sample period. For example the correlation between new cases and stock market valuations ($r = 0.02$) was very close to zero. A 1-day lagged analysis was also performed, with no discernible improvement.

Other factors (central bank intervention via quantitative easing and near zero interest rates, fiscal policy by government, e.g. the US CARES Act, unleashing a US$2.2m stimulus) are possible reasons [8]. Put another way, the stock market's traditional role as a discounting mechanism, effectively pricing risk, could have been over-ridden by other stakeholders, most notably governments.

A possible follow-up to this study would be to review the change in market valuations for specific sectors in country as a more meaningful predictor variable of new cases, as opposed to looking at aggregate indices. For example, the 4 largest stocks in the S&P 500 are Apple, Microsoft, Amazon and Alphabet, all of whom have been beneficiaries of the COVID pandemic. Valuation changes in the US ‘lodging and travel’ sector may have been a more productive independent variable.

References

[1] J. Surowiecki, The wisdom of crowds: why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations, 1st ed. New York: Doubleday, 2004.

[2] B. G. Malkiel, ‘The Efficient Market Hypothesis and Its Critics’, Journal of Economic Perspectives, vol. 17, no. 1, pp. 59–82, Feb. 2003, doi: 10.1257/089533003321164958.

[3] D. B. Hannah Ritchie Edouard Mathieu, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald and M. Roser, ‘Coronavirus Pandemic (COVID-19)’, Our World in Data, 2020, [Online]. Available: https://ourworldindata.org/coronavirus#citation

[4] OpenTable, ‘The restaurant industry in recovery’. https://www.opentable.com/state-of-industry

[5] H. Markowitz, ‘PORTFOLIO SELECTION*’, The Journal of Finance, vol. 7, no. 1, pp. 77–91, Mar. 1952, doi: 10.1111/j.1540-6261.1952.tb01525.x.

[6] J. Forrest, ‘How John Burn-Murdoch’s Influential Dataviz Helped The World Understand Coronavirus’, Apr. 2021, [Online]. Available: https://medium.com/nightingale/how-john-burn-murdochs-influential-dataviz-helped-the-world-understand-coronavirus-6cb4a09795ae

[7] S. Alizon, A. Hurford, N. Mideo, and M. Van Baalen, ‘Virulence evolution and the trade-off hypothesis: history, current state of affairs and the future: Virulence evolution and trade-off hypothesis’, Journal of Evolutionary Biology, vol. 22, no. 2, pp. 245–259, Feb. 2009, doi: 10.1111/j.1420-9101.2008.01658.x.

[8] M. Joyce, D. Miles, A. Scott, and D. Vayanos, ‘Quantitative Easing and Unconventional Monetary Policy – an Introduction’, The Economic Journal, vol. 122, no. 564, pp. F271–F288, Nov. 2012, doi: 10.1111/j.1468-0297.2012.02551.x.