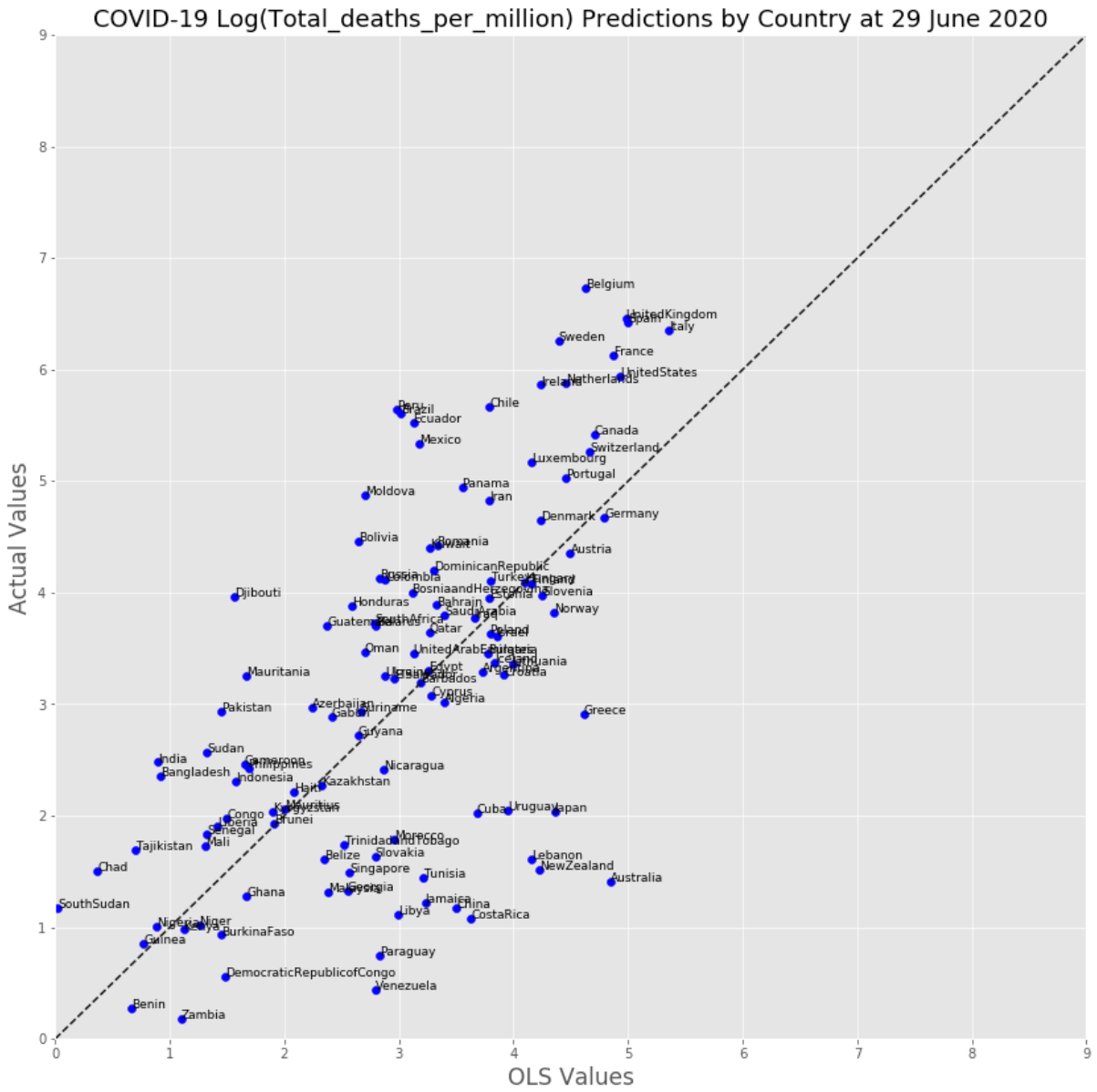
PANDEMICS - DATA & ANALYTICS

Exploring inter-country coronavirus mortality

Deaths Per Million ~ Age + Comorbidity + Obesity



Trevor Nell, Ian McGorian, Nick Hudson

Working paper version 1.2. Original released 7 July 2020. This version released 9 July 2020.

Contents

1. Abstract

One of the most interesting features of the COVID-19 outbreak is the stark difference between mortality experience in different countries. No simple and plausible explanations that we are aware of have been advanced. Though various hypotheses have been put forward, some more hopeful than others, many display an element of confirmation bias in attempting to locate all differences in non-pharmaceutical intervention approaches.

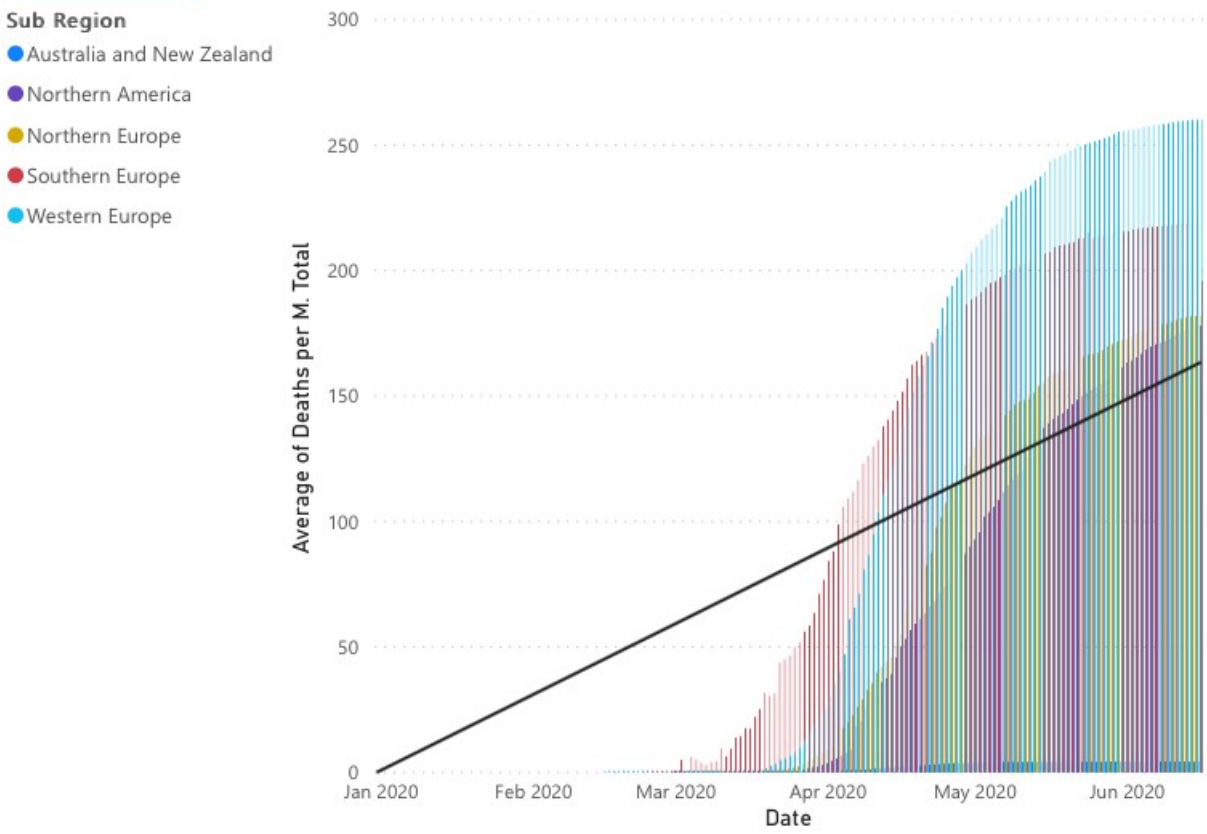
For each country put forward as an example, usually in some pairwise comparison and with an attendant single cause explanation, there are a host of countries that fail the expectation. We set out to model the disease with every expectation of failure. In choosing variables it was obvious from the outset that there would be contradictory outcomes in the real world. But there were certain variables that appeared to be reliable markers as they had surfaced in much of the media and pre-print papers. These included age, co-morbidity prevalence and the seemingly light population mortality rates in poorer countries than that in richer countries. Even the worst among developing nations—a clutch of countries in equatorial Latin America—have seen lighter overall population mortality than the developed world. Our aim therefore was not to develop the final answer, rather to seek common cause variables that would go some way to providing an explanation and stimulating discussion. There are some very obvious outliers in this theory, not the least of these being Japan.

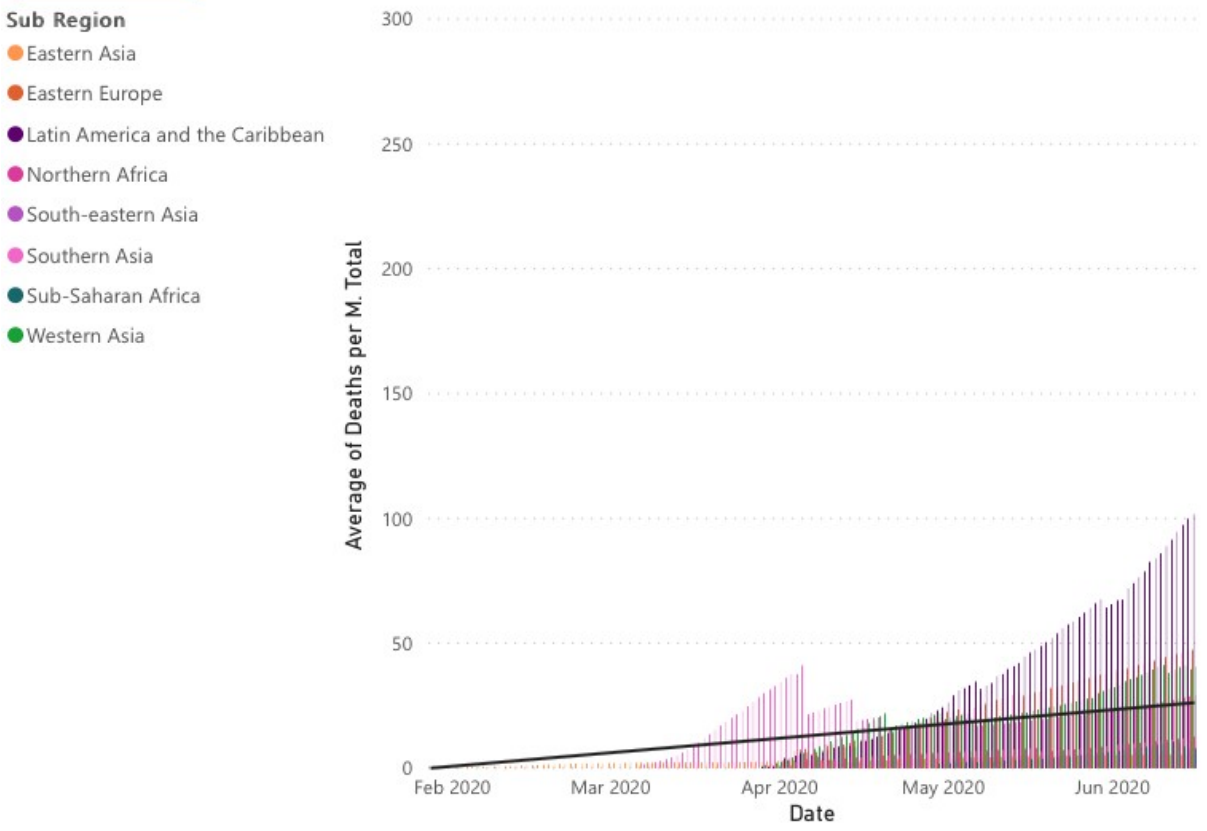
We test and find wanting the popular notions that lockdowns with their attendant social distancing and various other NPIs confer protection. Health care quality also fails to display any statistical benefit despite the intuitive appeal it has. Similarly, neither dread respiratory disease (*ie* TB), nor HIV prevalence, have proven the red flags posited by the medical fraternity. Of course, we would have been remiss had we not tested for other plausible concerns such as smoking, cholesterol, child mortality rates, altitude and so on.

While our results explain roughly half of the inter-country variability, they appear to be far more robust than the current explanations in circulation. We are hopeful that other researchers will identify factors that can improve our model.

2. Introduction

One of the most interesting features of the COVID-19 outbreak is the stark difference between the experience of developed and developing countries, plotted to scale below:





Our starting point is the "Panda Hypothesis" proposed in early May, that inter-country differences were driven by age, comorbidity burden and the hygiene hypothesis. We explore the drivers behind inter-country experience, explaining half of the variance between countries using the following factors:

1. the portion of the population that is older than 70, to capture age structure;

2. an index of several frequently observed co-morbidities associated with bad COVID-19 case outcomes, conflated with obesity prevalence, which we expected to be positively correlated with death rates per million;

3. an index proxying for the hygiene hypothesis, which we expected to be negatively correlated with death rates; and

4. healthcare spend per capita, which we expected to be positively correlated with death rates.

We show that the model can be marginally improved by replacing the hygiene score with obesity rates.

To our knowledge, there have been very few attempts to explore inter-country mortality rate variation seriously.

- Dr Chris Hope of the Judge Business School proposed an inverse correlation for European nations between COVID-19 mortality and previous season influenza mortality[1].
- In a separate paper, "Understanding Regional Differences in COVID-19 Mortality Rates", our colleagues Anthonyrajah and Lacerda explored the effects of average age, hospital bed availability and mobility.
- Dashboards have shown median age of population to be correlated with mortality.
- Knitzen and Ozaltun[2] found within-state relevance in the United States for public transport utilization, share of people not working, higher home values, higher summer temperatures, and lower winter temperatures, all associated with higher death rates, but not share of elderly, race, pollution, obesity, ICU beds per capita, or poverty rates.

[1] https://www.jbs.cam.ac.uk/fileadmin/user_upload/research/workingpapers/wp2003.pdf
[2] https://www.nber.org/papers/w27391

- Desmet and Wacziarg[3] wrote on spatial variation, but used absolute number of deaths as a response variable. We do not understand the meaning of this.

Our work is therefore to be regarded as preliminary and we invite comment, criticism and references to other papers or teams that may be working on this problem. We have several extensions in mind and fully expect to build on this preliminary work.

2.1. Intuitions behind the model

In countries where seroprevalence testing has been conducted, varying seroprevalence rates have been observed [4] [5] [6] [7] [8]. These have seemingly countered the prevailing position of epidemiologists positing high rates of susceptibility and predictions that "herd immunity" would only be achieved at 60 to 70% seroprevalence. Such predictions have been further countered by the lack of perceptible change in death rates upon the lifting of lockdowns, voluntary social distancing being eschewed at protest events and simple "disobedience" behaviours in the leisure context. Despite these falsifications, the proponents for lockdown measures remain adamant in core beliefs regarding the correct response measures. We have not been surprised by these phenomena, because we have long been persuaded by the idea that a significant proportion of all populations deal with the virus at the level of the cellular, or innate, immune systems [9] [10]. In addition to the very large asymptomatic group, who test positive for coronavirus but never notice any symptoms, and the small symptomatic group, who fall ill and a small portion of whom die, there is another group involved. These "bouncers" have healthy immune systems and a large repertoire of Tc and Th cells. When infected, probably by low viral doses, their cellular immune systems deal with the virus effectively. The T-cell response is effective and a B-cell antibody response does not follow. They generate a level of cellular, but not humoral immunity. They are very unlikely to test positive on virology studies and never will on serology studies. They are unlikely to spread the disease by the very nature of their immune system response.

[3] https://www.nber.org/papers/w27329.pdf
[4] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31304-0/fulltext
[5] https://preprints.scielo.org/index.php/scielo/preprint/download/404/497/494
[6] https://www.medrxiv.org/content/10.1101/2020.06.08.20125179v1
[7] https://www.researchsquare.com/article/rs-25862/v1
[8] https://english.elpais.com/society/2020-06-05/spains-macro-study-show-just-52-of-population-has-contractedthe-coronavirus.html
[9] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4125530/
[10] https://www.cell.com/cell/fulltext/S0092-8674(20)30610-3

Whilst there may be sceptics, general practitioners, paediatricians and physicians we have spoken to have no problem with the concept, which is based on an understanding of the immune system that is decades old. We expect the proportion of "bouncers" in a population to reflect the general burden of communicable disease in that population. This we posit as the hygiene hypothesis. In common parlance we speak of "Delhi belly"—travellers to low hygiene countries often get upset stomachs for a few days, whereas travellers from low hygiene countries do not. The idea is that this disease burden acts to eliminate people with weak Tc and Th repertoires and to exercise the repertoires of survivors, causing greater T-cell expression and delayed immunosenescence.

Separately, we observe that the presence of comorbidities has been widely found to be causal in case outcomes. Obesity is not consistently recorded as a comorbidity, but is thought to have a causative role in development of chronic comorbidities. The word "chronic" does a lot of work here. For example, young diabetics do not show up prolifically among deaths, whereas older ones do, suggesting that sustained presence of a comorbidity causes damage that places sufferers at risk to death from COVID-19. We are sympathetic to the idea that the relevant damage is vascular.

It has been observed that certain countries devote a large portion of their healthcare expenditure to the last few years of their populations' lives[11] [12] [13] [14] [15]. We theorize that, by spending a lot keeping frail people alive, such countries set up large numbers of people to being "carried out" by disease outbreaks. The observation that this years' higher death tolls follow a mild influenza season in such countries last year is concordant with this notion.

Both the hygiene and the healthcare expenditure factors can be colloquially expressed as the idea that, in developing countries, the susceptible people are already dead.

2.2. How explanatory do we expect our model to be?

Idiosyncratic elements of reporting methods and policy responses are known to be important. In particular, among developed world countries, the handling of both with respect to nursing homes has been pivotal:

[11] https://www.ncbi.nlm.nih.gov/books/NBK476262/
[12] https://healtheconomicsreview.biomedcentral.com/articles/10.1186/s13561-019-0224-z
[13] https://publichealthmatters.blog.gov.uk/2019/01/29/ageing-and-health-expenditure/
[14] https://www.medrxiv.org/content/10.1101/2020.05.11.20098442v2
[15] https://www.rand.org/pubs/research briefs/RB9146-1.html

- Most deaths (more than 80% in Canada[16]) have occurred in nursing homes.

- While Sweden attributes all nursing home deaths associated with a positive COVID-19 test in the month prior to death to COVID-19, even if the actual cause of death is independent of clinical expression of COVID-19 symptoms, Finland initially reported only hospital deaths, ignoring deaths in nursing homes entirely[17].

- A significant portion of COVID-19 deaths in the United States seem to have resulted from a policy implemented by six states that forced nursing homes to accept infected hospital cases[18].

Such idiosyncrasies are not currently amenable to statistical analysis, leaving us with a sense that our model will inevitably leave much inter-country variance unexplained.

A danger with regression analyses is that they turn into data mining exercises, consistent with the empiricist approach that characterises much of the bad science around COVID-19. We have tried to adhere to a Popperian approach of deploying regression as a means to falsify our original hypothesis, contained in "The Panda Hypothesis". Evidence that the hypothesis predates our statistical investigation is scattered throughout our articles, interviews and correspondence.

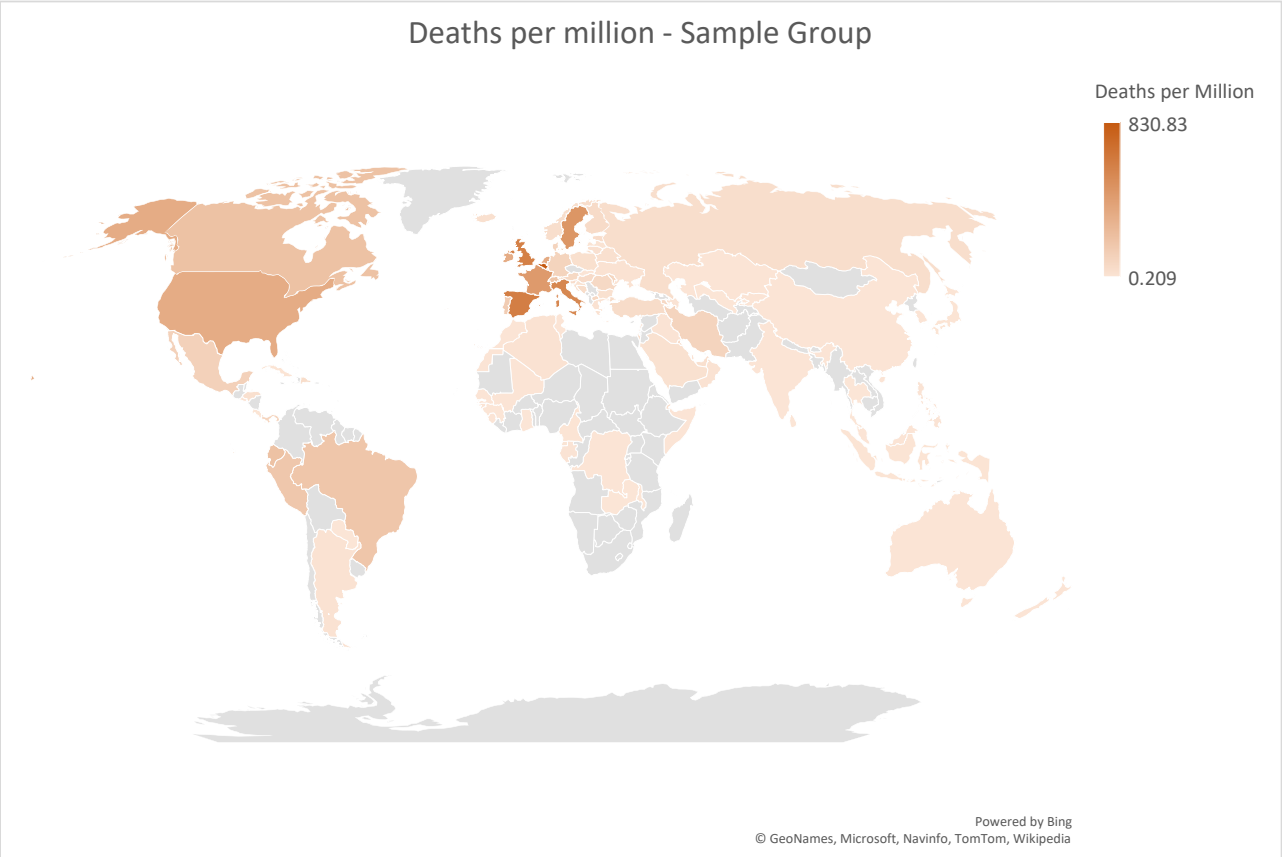
[16] https://www.wsj.com/articles/coronavirus-lays-bare-poor-conditions-in-canadas-nursing-homes-11592996400
[17] https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html

[18] https://www.politico.com/news/2020/06/04/states-nursing-homes-coronavirus-302134

3. Method

3.1. Dependent variable

Our model is calibrated to cumulative deaths per million. Geographic clustering is a notable feature of the response variable:



A future extension of this work would be to "complete the curve" by fitting Gompertz curves to the empirical distributions as proposed by Levitt [19] and successfully deployed by our colleagues Anthonyrajah and Lacerda in "Understanding Regional Differences in COVID-19 Mortality Rates"[20], who found good fits for this method. This extension is considered supplementary to the current paper. It is not expected that results presented here would be impacted significantly.

3.2. Independent variable: Age

Having tested more subtle approaches, we favour parsimony, utilizing the percentage of people over the age of 70.

[19] https://www.medrxiv.org/content/10.1101/2020.06.26.20140814v1
[20] https://www.pandata.org.za/wp-content/uploads/2020/06/covid-19paper.pdf

3.1. Independent variable: Hygiene hypothesis

We construct an index of communicable disease burden by summing the population fatality rates for hepatitis, tuberculosis, diarrheal diseases and HIV/AIDS.

3.2. Independent variable: Comorbidity

We construct an index of comorbidity by simply summing the population prevalence rates for diabetes, dementia, cardiovascular diseases, lower respiratory infections, respiratory diseases and kidney diseases..

3.3. Independent variable: Duration of epidemic

We tested two alternative designs. The first considered only 100 post-peak countries, defined as countries that have passed the peak for confirmed cases, confirmed by calculated active cases being lower than resolved cases. Resolved cases were defined as any active cases that are 20 days or older, the premise being that the individual has either recovered or died. This serves to address obvious and wholly understandable inaccuracies in reporting of recoveries. This approach permitted out-of-sample testing in the sense that it could produce post-peak estimates for pre-peak countries. A plot of their predicted values using this model against their cumulative deaths to date produced pleasing correlations, suggesting that our efforts had not been an exercise in data-mining.

The second design added a time variable (days since cumulative deaths pass 0.1 per million of population). The model produced quite similar results and is the version presented here, simply because it covers more of the world map. This is a crude adjustment that we will remove as more countries' epidemics mature, or improve upon as described in section 3.1.

3.4. Variables tested and omitted

We detail the effects of including obesity prevalence and healthcare expenditure in the model. Both are significant, but causally related to the above variables.

It is noteworthy that neither average nor maximal lockdown stringency[21] are at all correlated with the residuals or the response variable. Lockdowns do not appear to reduce deaths or flatten epidemic curves in any way. We suspect that the survival of the perception that they do is a result of the belief

[21] https://raw.githubusercontent.com/OxCGRT/covid-policy-tracker/master/data/OxCGRT_latest.csv

that initial epidemiological modelling, which failed to take into account, inter alia, the factors presented in this article, was accurate. This creates a sense that the difference between modelled and actual experience is owing to lockdown, when instead it is owing to model error. The picture presented is consistent with the oft-cited observation that the reproduction rate decays quite linearly, producing Gompertz-family fits and that this linear decay is interrupted neither by imposition of lockdown measures nor by their termination. The burden of proof for lockdown efficacy must surely reside with the proponents for this intervention. We find no evidence that is supportive and have yet to see evidence in support of this previously untested intervention.

We tested whether capturing fine-grain age structure proportions would increase the power of the regression. We did this by multiplying the infection fatality rates inferred from a Swedish randomized seroprevalence study[22] by the age distribution for each country and arbitrarily dividing this by the same result for Sweden. The resultant scalar is best thought of as a relative exposure to risk measure, capturing the intense variance in hazard ratios by age. This did not enhance performance by enough to justify the extra complexity.

We also tested incidence of female mortality in the age group 15 to 29 and infant mortality as hygiene proxies, with no improvement in the model.

None of cigarette consumption, daily cholesterol intake, latitude or UV index displayed statistical significance.

In a separate paper, "Understanding Regional Differences in COVID-19 Mortality Rates", our colleagues Anthonyrajah and Lacerda did not establish significance for average age, hospital bed availability or mobility.

Several intra-country studies[23] [24] [25] have noted that the O blood group is beneficial or that the A blood group is harmful in the context of COVID-19 mortality. We have detected a very strong effect using international blood group prevalence data, mildly at odds with this finding, as well as a strong effect for Rhesus factor prevalence. We are doing further research into the potential causal

[22] https://medicalxpress.com/news/2020-05-stockholm-virus-antibodies-sweden.html
[23] https://www.nejm.org/doi/full/10.1056/NEJMoa2020283?query=featured_coronavirus

[24] https://gulfnews.com/world/23andme-provides-more-evidence-that-blood-type-plays-role-in-covid-19-virus-1.1591635107796
[25] https://www.bloomberg.com/news/articles/2020-06-08/23andme-provides-more-evidence-that-blood-type-plays-role-in-virus

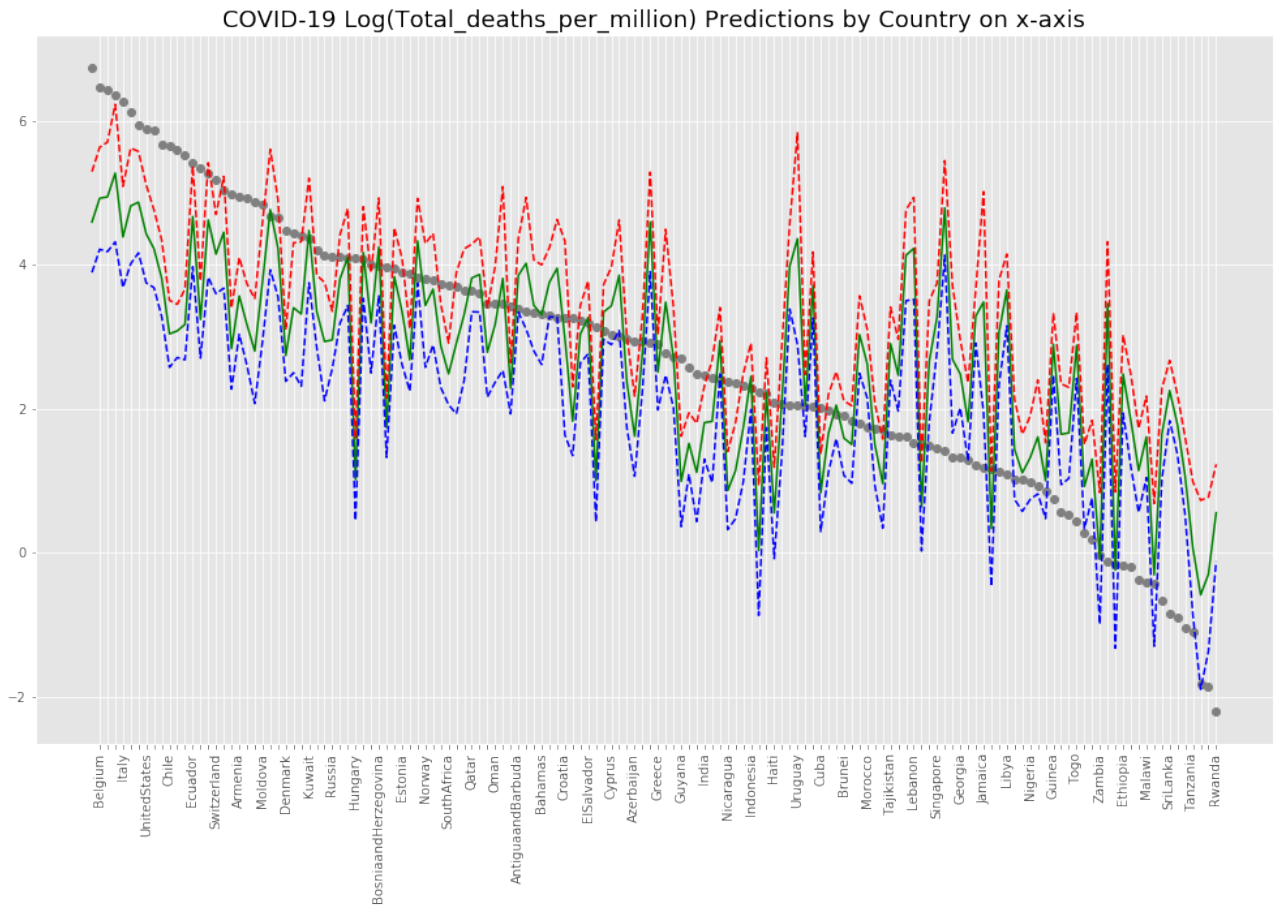
mechanisms and, until this is complete, will not release our findings. We also note collinearity with the related variables and the hygiene factor, which may have implications for our model structure.

3.5. Variables not yet explored

Factors that we have not yet tested include alternative methods of assessing the impact of density, altitude, diet, prior season influenza severity (noted to be discriminant in the Nordic countries), influenza timing in different locales and vitamin D deficiency. We are still working on compiling datasets.

3.6. Observations regarding actual versus predicted values

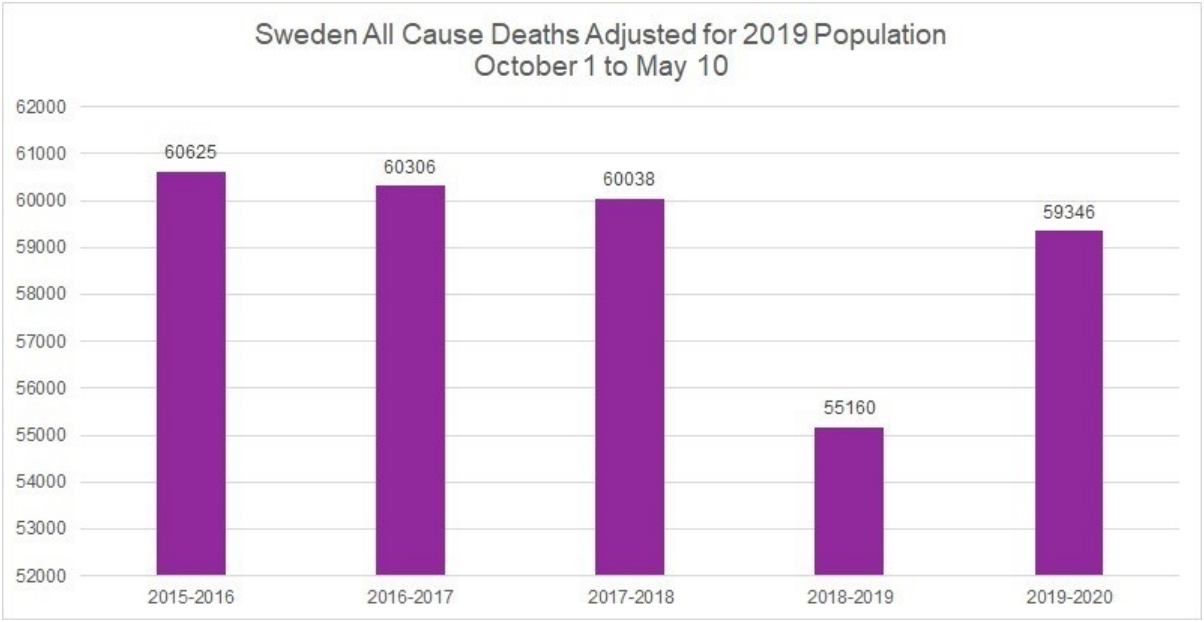
With only half of variance explained, many countries fall outside the 95% confidence interval of the regression model:



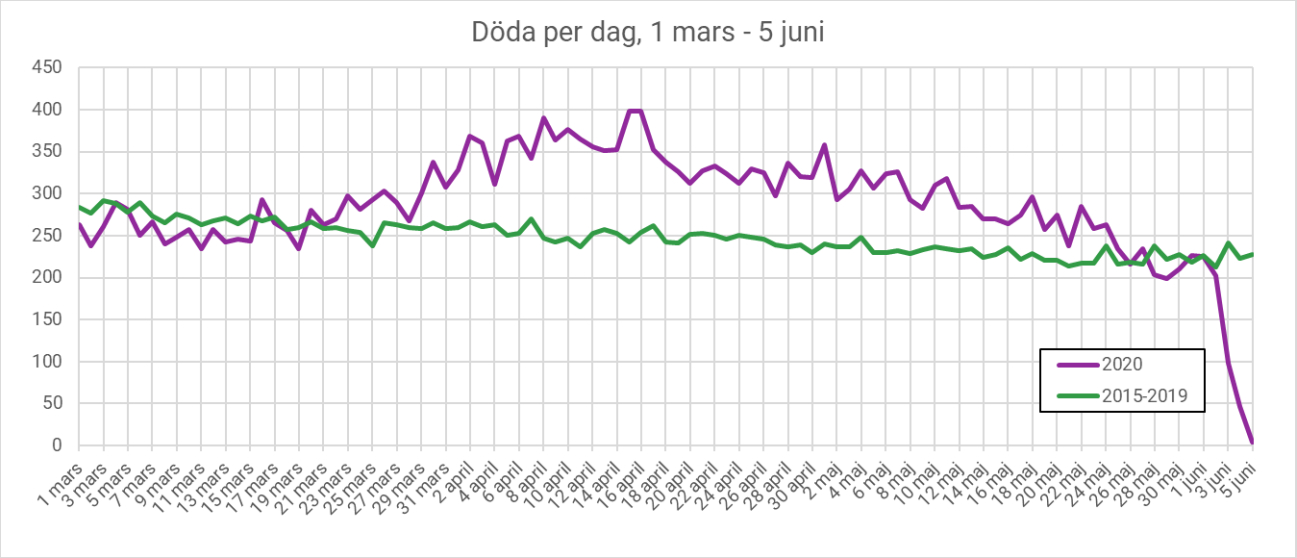
We have chosen not to release a table of predicted values for fear that these would be interpreted as guides for policymakers. It would be incorrect to abstract values from the above—our purpose is to go some way towards explaining inter-country differences, not to forecast epidemics.

3.7. Potential improvement

The vexing issue of inconsistency among nations in their approach to defining COVID-19 deaths is hard to resolve. Many have suggested assessing total excess mortality relative to the prior year and this has gained significant media attention. However, Sweden's cumulative mortality is well below that of its five-year average. We note that this may be in part explained by in-country reporting lags.



This is not to suggest that Covid-19 hasn't caused a deviation in deaths during the course of the outbreak.



https://www.scb.se/en/About-us/news-and-press-releases/statistics-sweden-to-publish-preliminarystatistics-on-deaths-in-sweden/

Also, as we pointed out in our previous paper ("Quantifying years of life lost to lockdown"[26]), we expect much greater mortality to result from lockdown than from COVID-19. It has been suggested that such mortality is already evident. Only a third of the excess deaths seen in the community in England and Wales can be explained by COVID-19[27].

Potential causes include reluctance or reduced ability to seek medical attention, suicide, and increased alcoholism and drug use. In developing nations, one might add interrupted vaccination programs, ARV treatments and malnutrition.

[26] https://www.pandata.org.za/wp-content/uploads/2020/06/PANDA-Research-Report-Quantifying-Years-of-Lost-Life-PDF_.pdf
[27] https://www.bmj.com/content/369/bmj.m1931

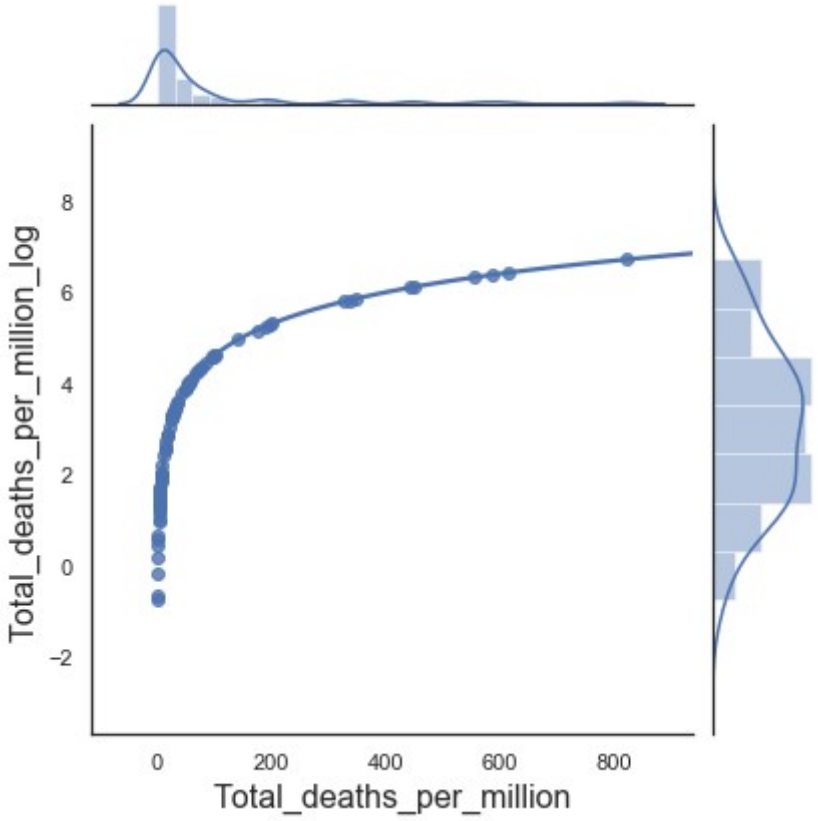
4. Results

Unprecedented data has flowed in connection with this epidemic. At the time of writing, some 100 countries have inflected, meaning that their peak daily deaths have nearly been achieved. It is recognised that the disease has followed a quite consistent pattern for gaining traction in different parts of the world at different time periods.

The population fatality rate, used as the response variable in this paper, is total deaths per million (DPM), calculated as follows:

DPM = total cumulative deaths at time t / population of country in millions

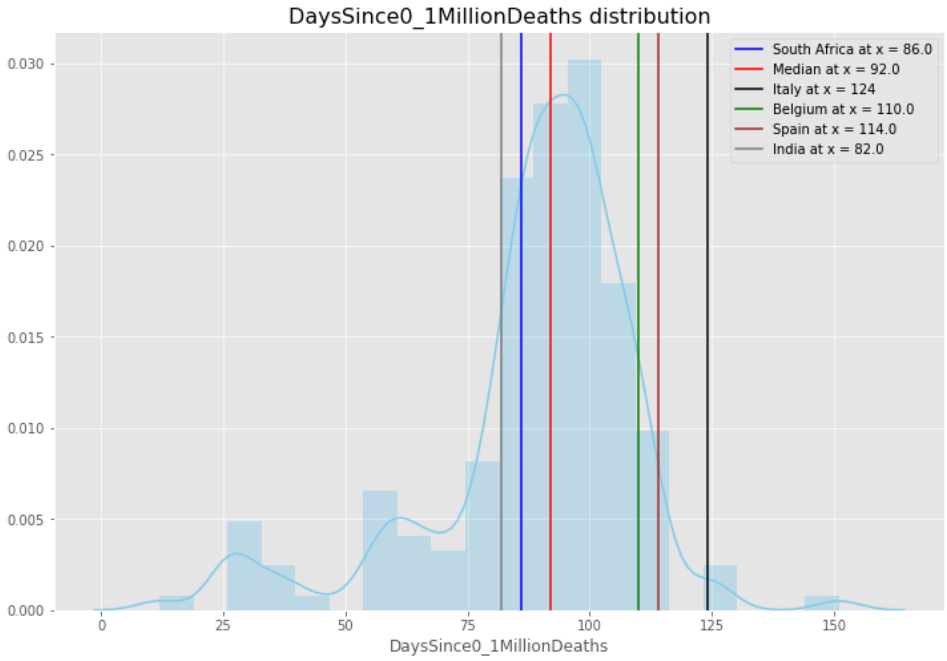
The distribution of the response variable is extremely skew, so we applied a log transformation to make the distribution more normal or, equivalently, to make non-linear relationships with the independent variables more linear:



4.1. Days Since 0.1 Million Deaths

We initially started regressing using only 100 or so “post-peak” countries, but found that adding “pre-peak” countries did not alter the results materially, provided that we added a compensating

variable—days since 0.1 per million deaths. This simply allows us to paint a picture for a greater portion of the planet, including 146 countries. In a month or two this variable will be dropped. This also means that, as far as this model is concerned, predicted values for pre-peak countries cannot be assessed as approximating full-course values for their epidemics.



A univariate model captures 33% of the variance:

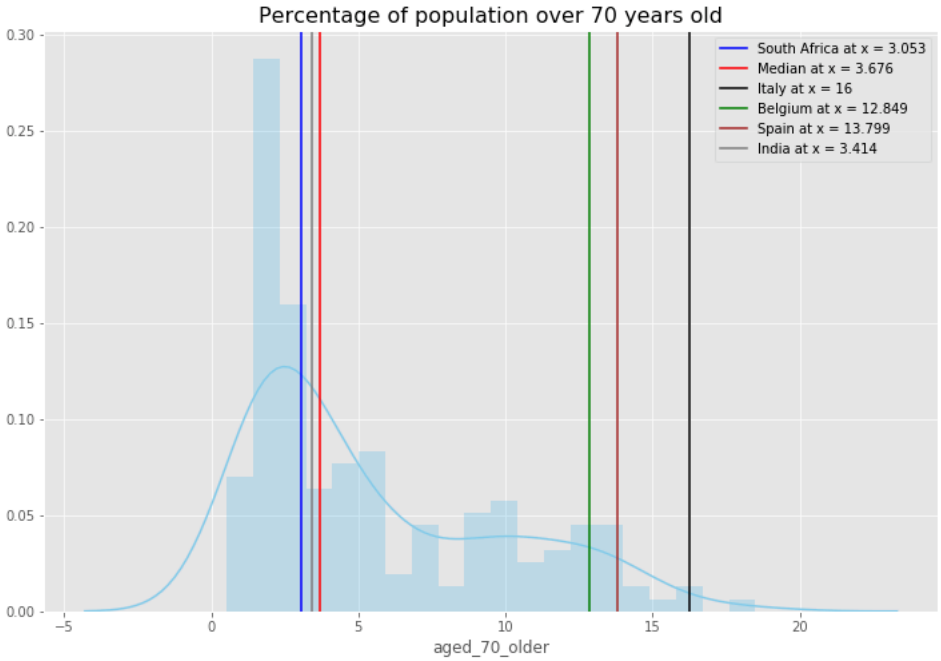
OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.329
Model:	OLS	**Adj. R-squared:**	0.325
Method:	Least Squares	**F-statistic:**	70.69
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	3.70e-14
Time:	07:35:23	**Log-Likelihood:**	-274.45
No. Observations:	146	**AIC:**	552.9
Df Residuals:	144	**BIC:**	558.9
Df Model:	1		
Covariance Type:	nonrobust		

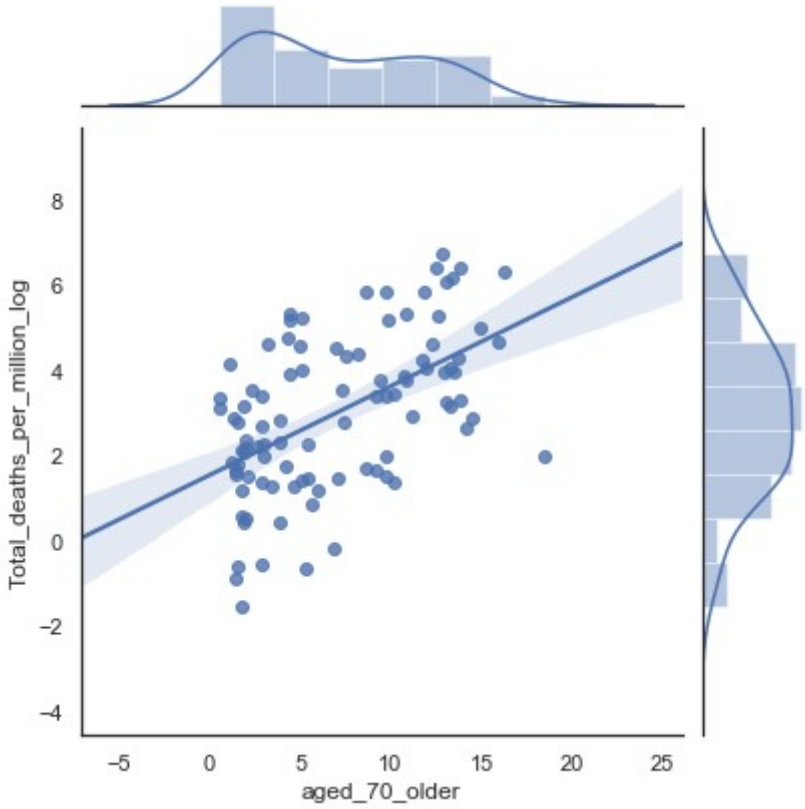
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-2.2425	0.604	-3.714	0.000	-3.436	-1.049
DaysSince0_1MillionDeaths	0.0552	0.007	8.408	0.000	0.042	0.068

4.2. Independent variable: Population over 70

The percentage of the population aged 70 years and older plots a histogram as follows:



The relationship between log (total deaths per million) and 70+ percentage is linear, with correlation 0.52:



In a univariate regression model, the 70+ population variable explains 27% of the variance:

OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.271
Model:	OLS	**Adj. R-squared:**	0.266
Method:	Least Squares	**F-statistic:**	53.42
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	1.71e-11
Time:	10:28:07	**Log-Likelihood:**	-280.58
No. Observations:	146	**AIC:**	565.2
Df Residuals:	144	**BIC:**	571.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.3837	0.228	6.070	0.000	0.933	1.834
aged_70_older	0.2287	0.031	7.309	0.000	0.167	0.291

Adding it to the model improves the R squared significantly:

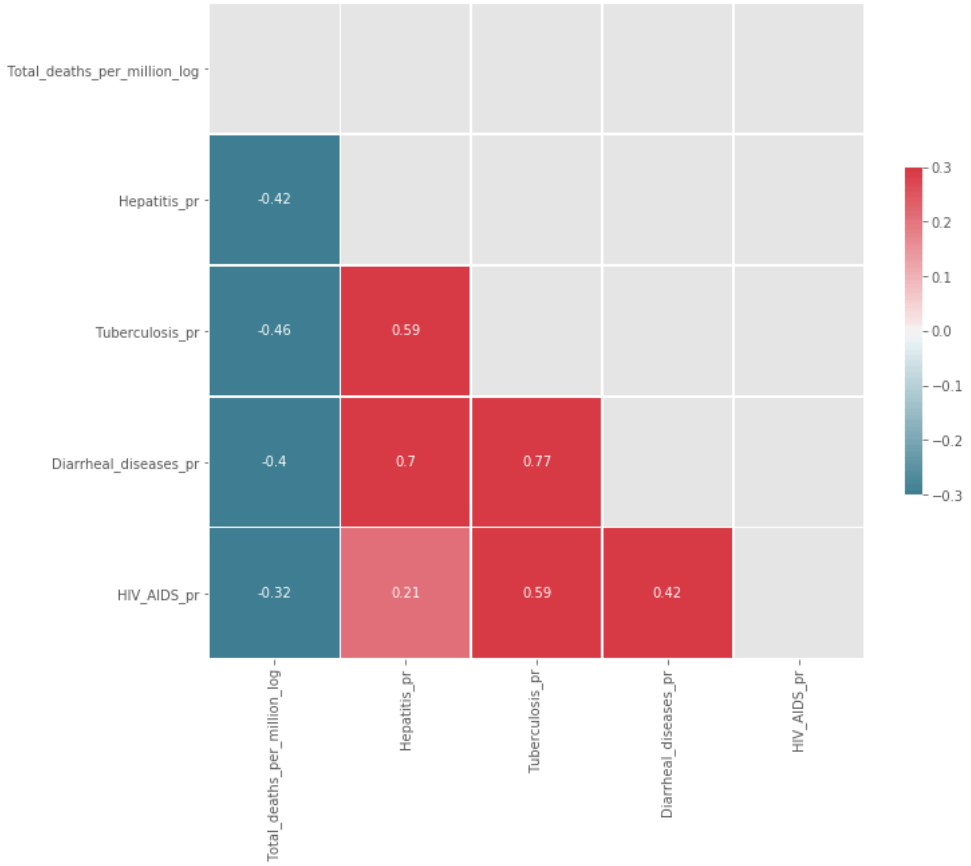
OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.403
Model:	OLS	**Adj. R-squared:**	0.394
Method:	Least Squares	**F-statistic:**	48.17
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	1.02e-16
Time:	10:28:37	**Log-Likelihood:**	-266.01
No. Observations:	146	**AIC:**	538.0
Df Residuals:	143	**BIC:**	547.0
Df Model:	2		
Covariance Type:	nonrobust		

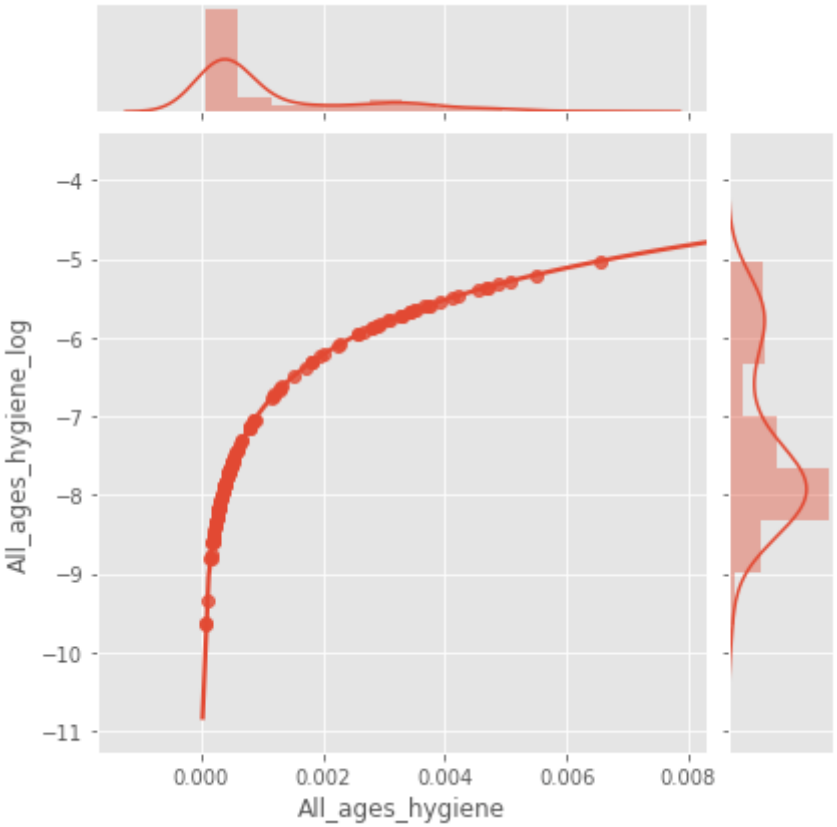
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.6992	0.586	-2.898	0.004	-2.858	-0.540
DaysSince0_1MillionDeaths	0.0403	0.007	5.619	0.000	0.026	0.054
aged_70_older	0.1371	0.033	4.187	0.000	0.072	0.202

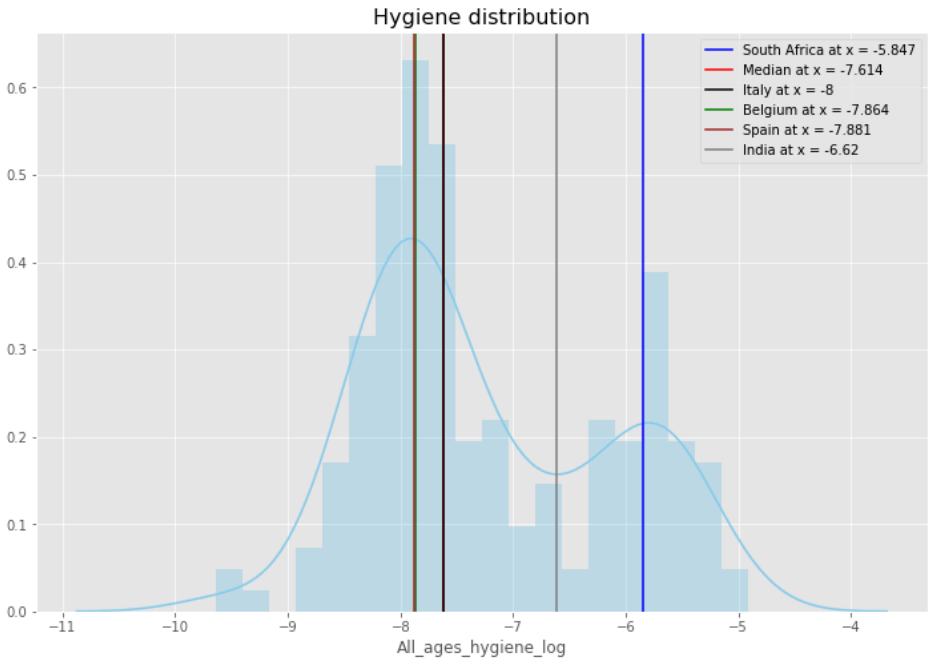
4.3. Independent variable: Hygiene

Under the hygiene hypothesis, we assess whether diseases common in low hygiene countries confer a protective benefit in terms of COVID-19 mortality. We proxy a hygiene factor by considering the burden of hepatitis, tuberculosis, diarrheal diseases and HIV/AIDS. Do countries with high prevalence of these enjoy protection via a cross-vaccination or vaccination-like effect against the COVID-19 disease, or alternatively, via removing people with weak innate immune systems from the population? The idea is that exposures to infectious agents deliver a broader innate immune response relative to populations where such diseases are rare. Here is a correlation plot depicting the relationship between the response variable and the hygiene disease fatality rates:



The correlations between each disease and the response variable are directionally the same and of comparable strength. Based on the above, we built an index from the hygiene diseases, by summing the fatality rates, which we then log-transform:





A univariate regression model explains around 25% of the variance:

OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.247
Model:	OLS	**Adj. R-squared:**	0.241
Method:	Least Squares	**F-statistic:**	47.11
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	1.85e-10
Time:	07:38:54	**Log-Likelihood:**	-282.95
No. Observations:	146	**AIC:**	569.9
Df Residuals:	144	**BIC:**	575.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.8124	0.961	-3.968	0.000	-5.711	-1.913
All_ages_hygiene_log	-0.8923	0.130	-6.864	0.000	-1.149	-0.635

The factor contributes marginally to the cumulative case regarding this hypothesis:

OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.420
Model:	OLS	**Adj. R-squared:**	0.407
Method:	Least Squares	**F-statistic:**	34.24
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	1.03e-16
Time:	07:37:20	**Log-Likelihood:**	-263.88
No. Observations:	146	**AIC:**	535.8
Df Residuals:	142	**BIC:**	547.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.1923	0.930	-3.432	0.001	-5.031	-1.353
DaysSince0_1MillionDeaths	0.0339	0.008	4.392	0.000	0.019	0.049
All_ages_hygiene_log	-0.3008	0.147	-2.053	0.042	-0.590	-0.011
aged_70_older	0.1134	0.034	3.296	0.001	0.045	0.181

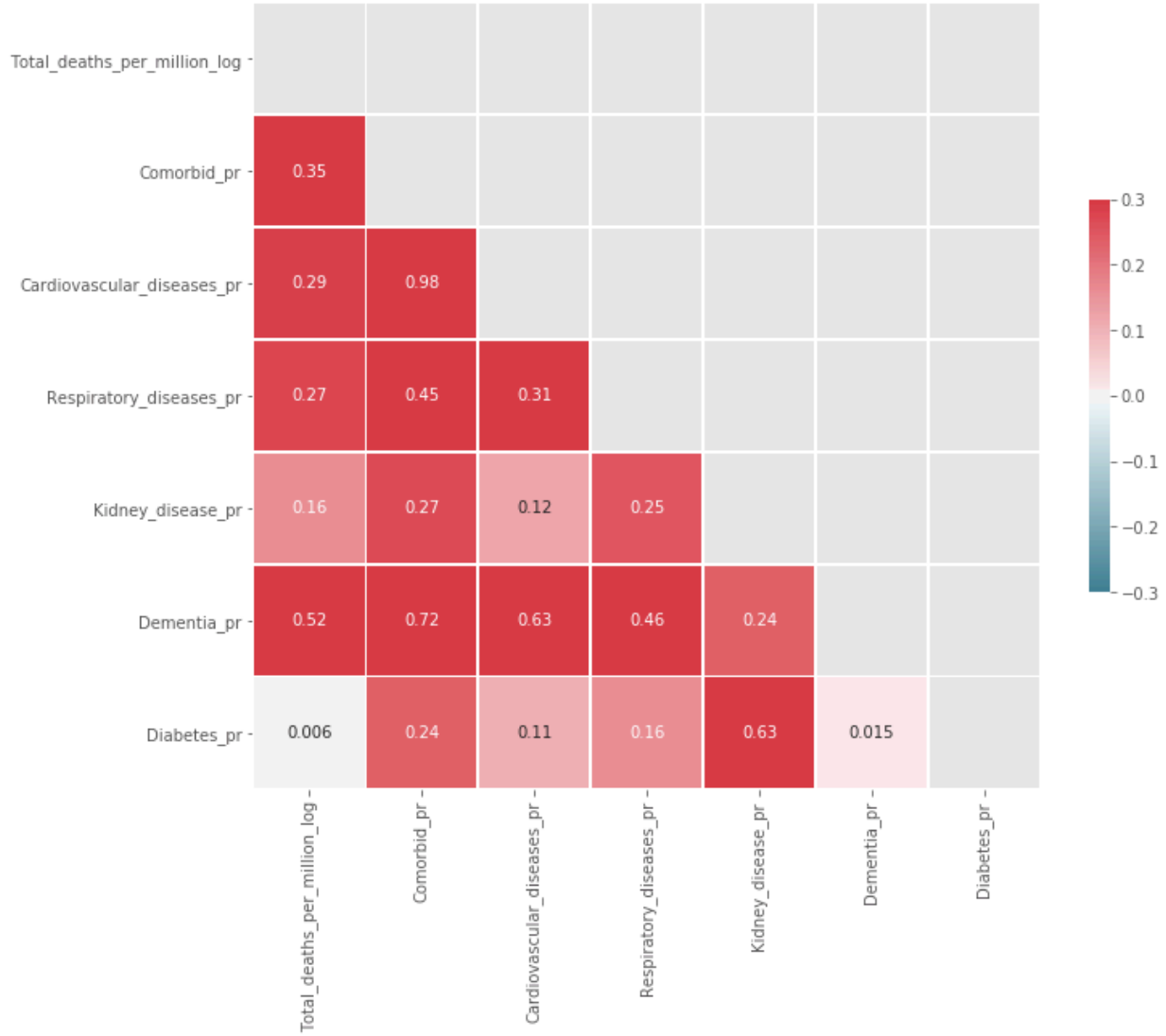
The variance inflation factors being well below 2 assure us there is little multicollinearity present:

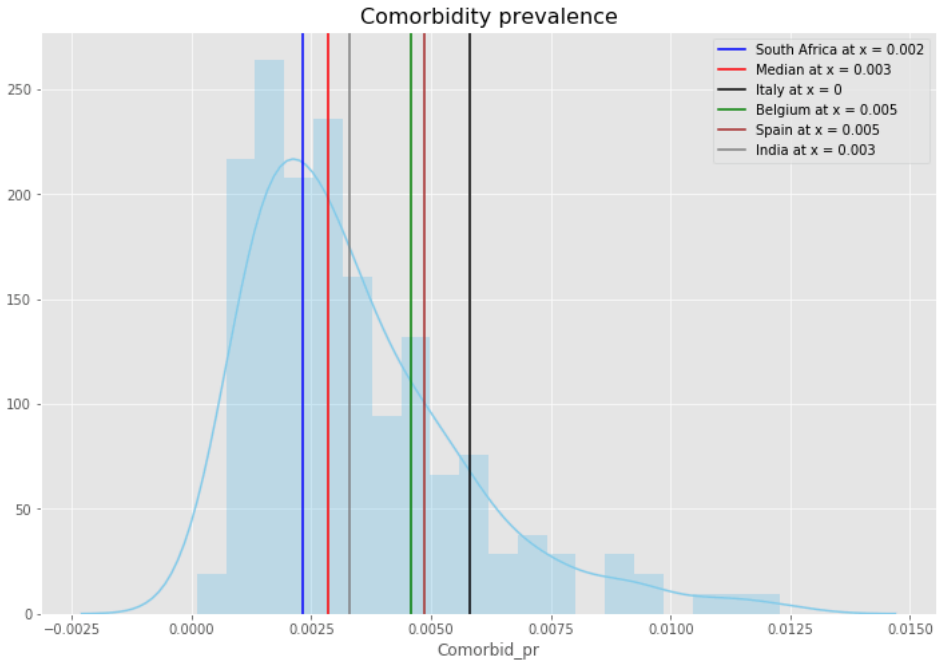
	variables	VIF
0	Intercept	56.502422
1	DaysSince0_1MillionDeaths	1.579938
2	All_ages_hygiene_log	1.626487
3	aged_70_older	1.498390

4.4. Comorbidity

The comorbidity variable is designed to assess whether burden of diabetes, dementia, cardiovascular diseases, lower respiratory infections, respiratory diseases and kidney diseases, which have been observed to contribute to poor COVID-19 case outcomes, has a positive relationship with population fatality. Do countries with high levels of these diseases suffer worse COVID-19 mortality?

The following correlation chart shows the response variable, the individual diseases and the index "Comorbid_pr", which is a sum of the prevalence rates for these diseases:





This index accounts for 12% of the variance, which we found surprisingly low:

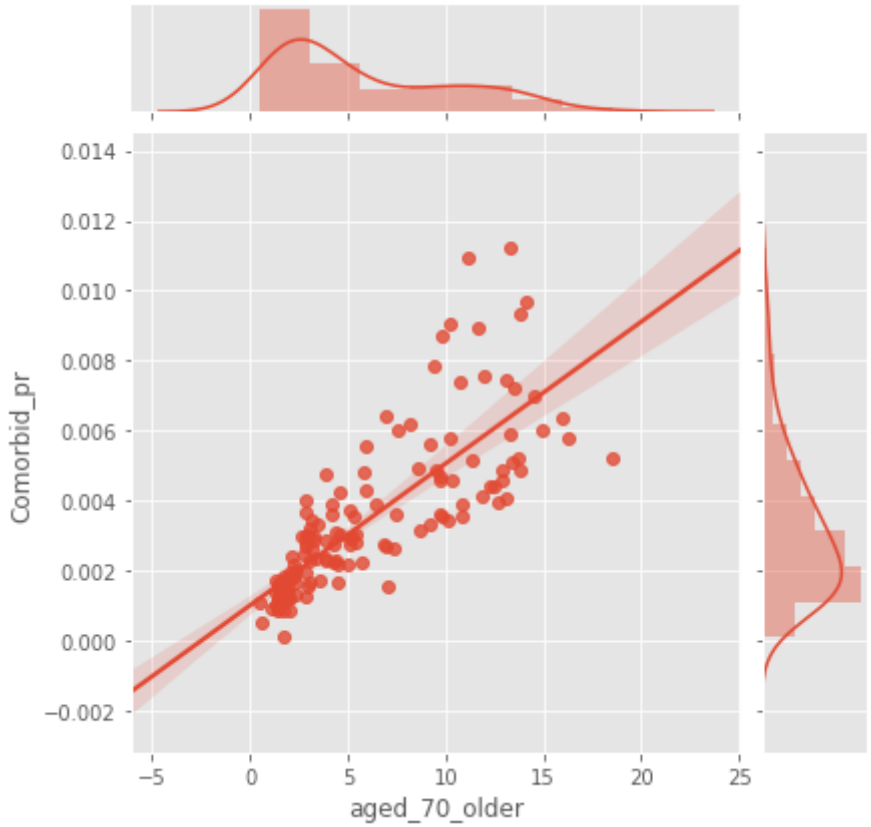
OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.123
Model:	OLS	**Adj. R-squared:**	0.117
Method:	Least Squares	**F-statistic:**	20.27
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	1.38e-05
Time:	07:38:36	**Log-Likelihood:**	-293.99
No. Observations:	146	**AIC:**	592.0
Df Residuals:	144	**BIC:**	598.0
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.6847	0.273	6.160	0.000	1.144	2.225
Comorbid_pr	304.9522	67.728	4.503	0.000	171.083	438.822

Modest multicollinearity between the comorbidity index and age is unsurprising, given that the strongest indicator was dementia, a disease of the elderly:

	variables	VIF
0	Intercept	56.924674
1	DaysSince0_1MillionDeaths	1.592641
2	All_ages_hygiene_log	1.634341
3	aged_70_older	3.310979
4	Comorbid_pr	2.794887



When we add the index into the multivariate model, it consequently adds little explanatory power, and its coefficient inherits a counter-intuitive sign:

OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.428
Model:	OLS	**Adj. R-squared:**	0.412
Method:	Least Squares	**F-statistic:**	26.41
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	2.37e-16
Time:	07:49:02	**Log-Likelihood:**	-262.79
No. Observations:	146	**AIC:**	535.6
Df Residuals:	141	**BIC:**	550.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-3.0761	0.930	-3.307	0.001	-4.915	-1.237
DaysSince0_1MillionDeaths	0.0329	0.008	4.262	0.000	0.018	0.048
All_ages_hygiene_log	-0.3155	0.146	-2.156	0.033	-0.605	-0.026
aged_70_older	0.1680	0.051	3.299	0.001	0.067	0.269
Comorbid_pr	-134.0449	92.409	-1.451	0.149	-316.732	48.642

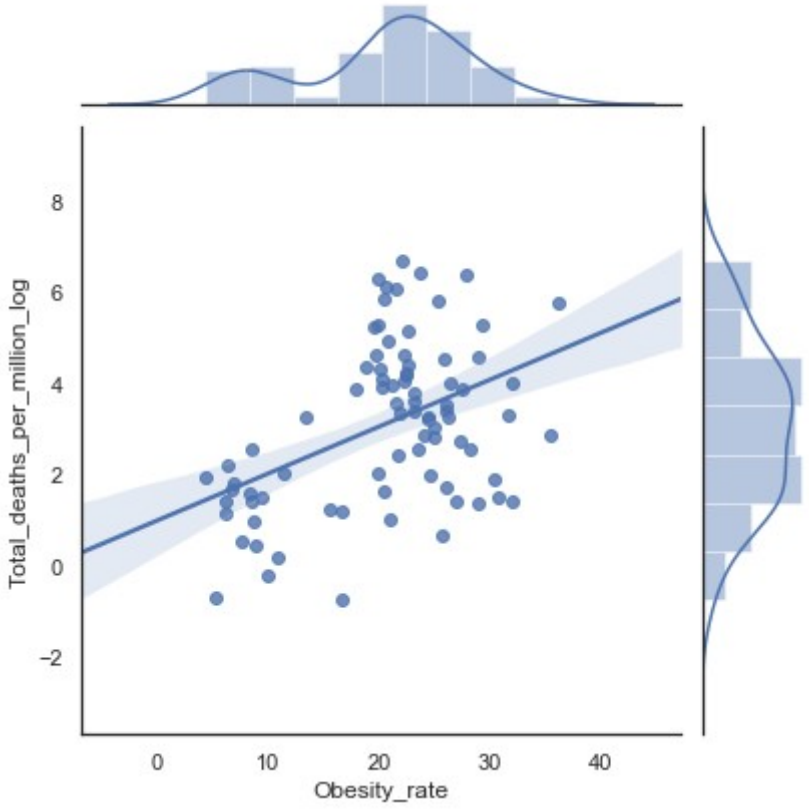
At this stage the model embeds three hypotheses:

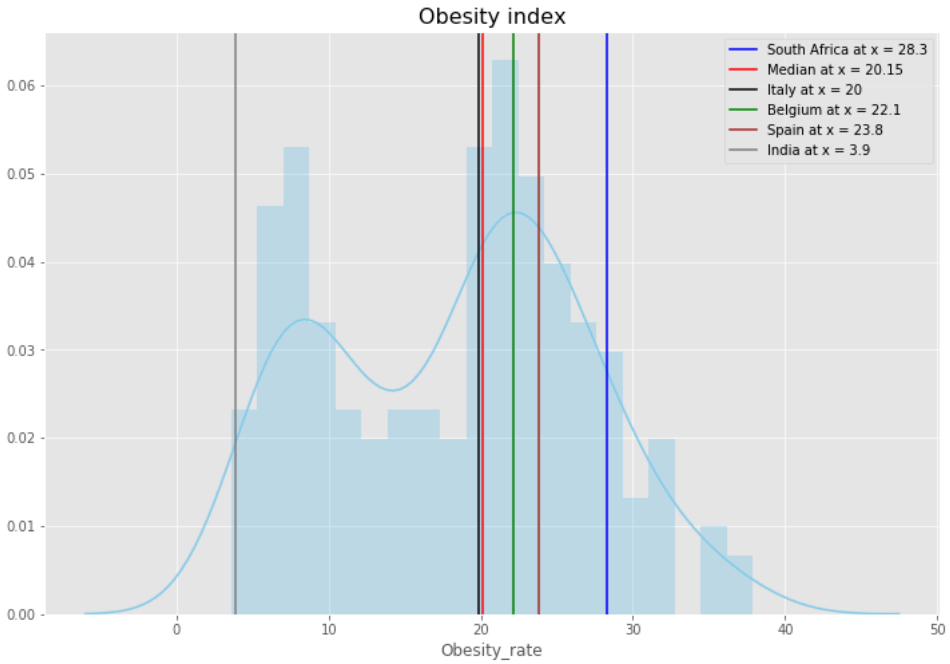
1. The age hypothesis
2. The hygiene hypothesis
3. The comorbidity hypothesis

It has been demonstrated that each of the variables exhibit some explanatory power when it comes to accounting for fatality rate differences, but that with an unsatisfactory multilinear model when they are all combined.

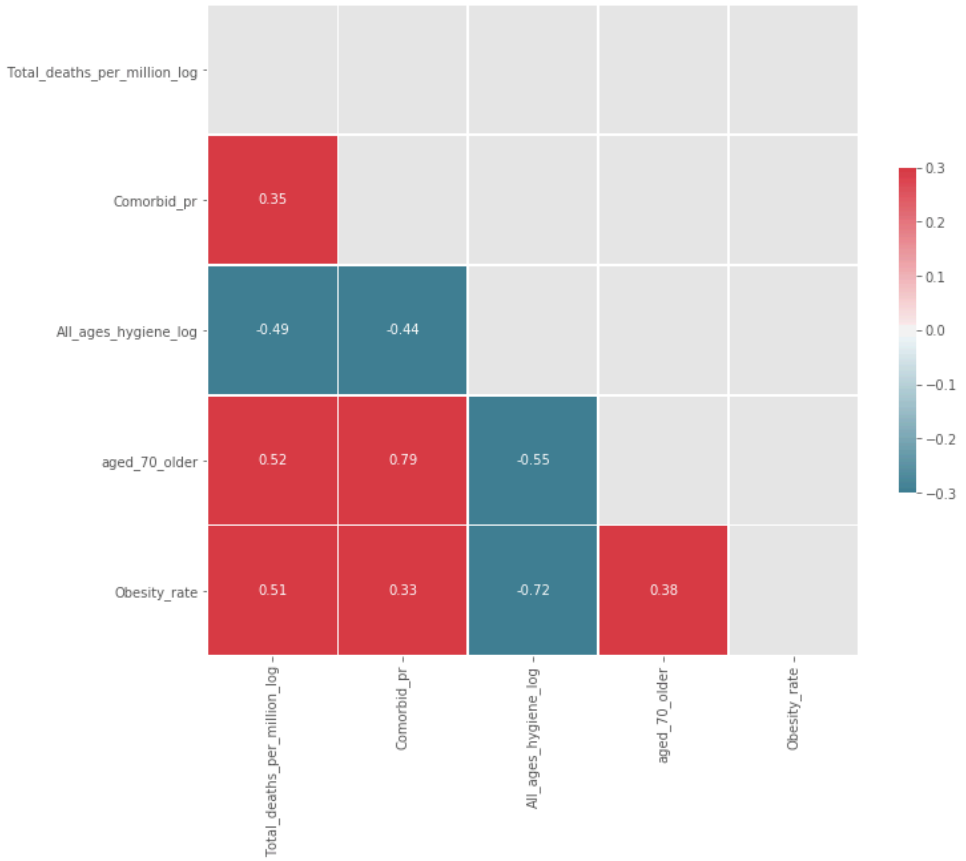
4.5. Comorbidity hypothesis extended—Obesity

Obesity was observed as a triage factor during the peak of the New York City outbreak. Country adult obesity rates, obtained from the WHO, are correlated with the response variable as follows:


Total_deaths_per_million_log
Obesity_rate
8
6
4
2
0
−2
0
10
20
30
40


Obesity index
South Africa at x = 28.3
Median at x = 20.15
Italy at x = 20
Belgium at x = 22.1
Spain at x = 23.8
India at x = 3.9
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
Obesity_rate

Obesity exhibits a 51% correlation with the response variable and explains 27% of the variance:



OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.272
Model:	OLS	**Adj. R-squared:**	0.267
Method:	Least Squares	**F-statistic:**	53.69
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	1.54e-11
Time:	07:55:25	**Log-Likelihood:**	-280.48
No. Observations:	146	**AIC:**	565.0
Df Residuals:	144	**BIC:**	570.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.5504	0.325	1.691	0.093	-0.093	1.194
Obesity_rate	0.1149	0.016	7.327	0.000	0.084	0.146

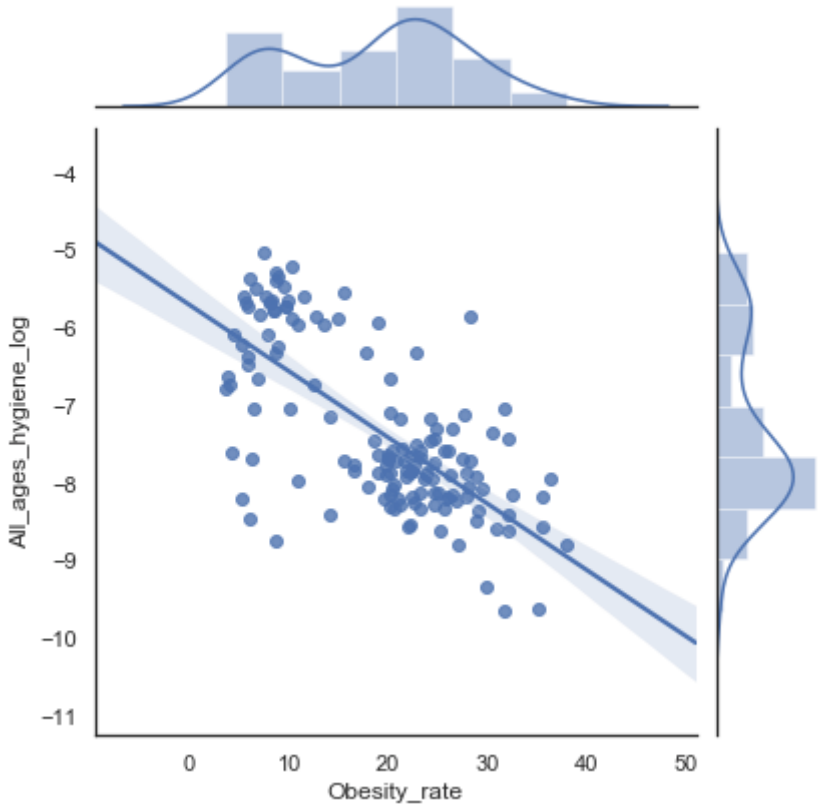
Adding obesity to the model lifts the variance explained from 42% to 47%:

OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.472
Model:	OLS	**Adj. R-squared:**	0.453
Method:	Least Squares	**F-statistic:**	25.01
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	5.93e-18
Time:	07:56:02	**Log-Likelihood:**	-257.02
No. Observations:	146	**AIC:**	526.0
Df Residuals:	140	**BIC:**	543.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.3424	1.032	-1.300	0.196	-3.383	0.698
DaysSince0_1MillionDeaths	0.0280	0.008	3.692	0.000	0.013	0.043
All_ages_hygiene_log	0.0376	0.175	0.214	0.830	-0.309	0.384
aged_70_older	0.1859	0.049	3.761	0.000	0.088	0.284
Comorbid_pr	-153.3351	89.321	-1.717	0.088	-329.928	23.258
Obesity_rate	0.0664	0.020	3.396	0.001	0.028	0.105

However, introducing obesity also renders hygiene insignificant. This is intuitive, as both are mediated by poverty. The two are negatively correlated (70%):



Obesity has, of course, been observed as a triage factor and comorbidity, so we tried replacing the hygiene factor with obesity, maintained the explanatory power, explaining 47% of variance:

OLS Regression Results

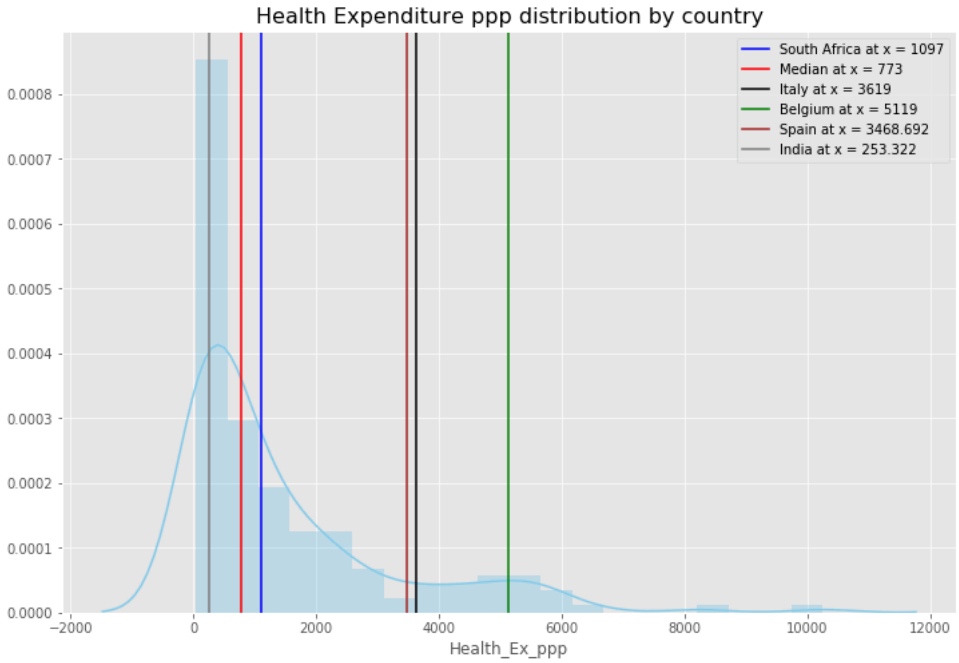
Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.472
Model:	OLS	**Adj. R-squared:**	0.457
Method:	Least Squares	**F-statistic:**	31.46
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	1.01e-18
Time:	07:56:36	**Log-Likelihood:**	-257.04
No. Observations:	146	**AIC:**	524.1
Df Residuals:	141	**BIC:**	539.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.5268	0.569	-2.682	0.008	-2.652	-0.401
DaysSince0_1MillionDeaths	0.0277	0.007	3.740	0.000	0.013	0.042
aged_70_older	0.1837	0.048	3.810	0.000	0.088	0.279
Comorbid_pr	-153.6794	89.004	-1.727	0.086	-329.635	22.276
Obesity_rate	0.0639	0.016	4.074	0.000	0.033	0.095

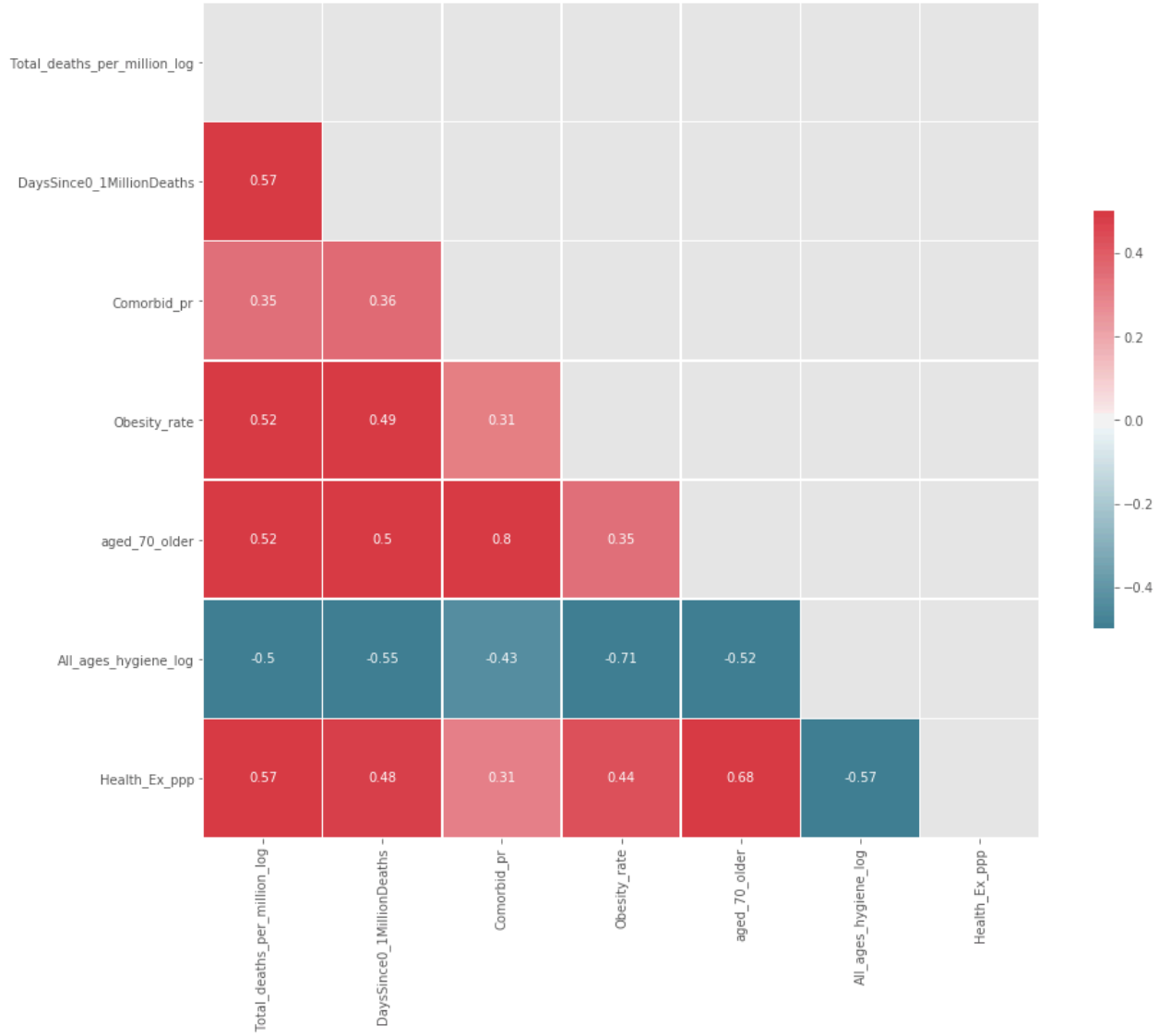
The current bundle of hypotheses is by no means exhaustive in terms of explanatory power, but nevertheless the model in its current form constitutes a useful heuristic device in partly explaining the drivers of Covid-19 fatality rates. In the next sections we continue to investigate other potentially useful factors such as healthcare expenditure, lockdown stringency and WHO healthcare rankings.

4.6. Healthcare expenditure

The healthcare expenditure hypothesis proposes that wealthier nations spend more on their population, particularly late in life, to prolong life. This ultimately creates a population significantly at risk to disease outbreaks, especially after a weak prior influenza season.



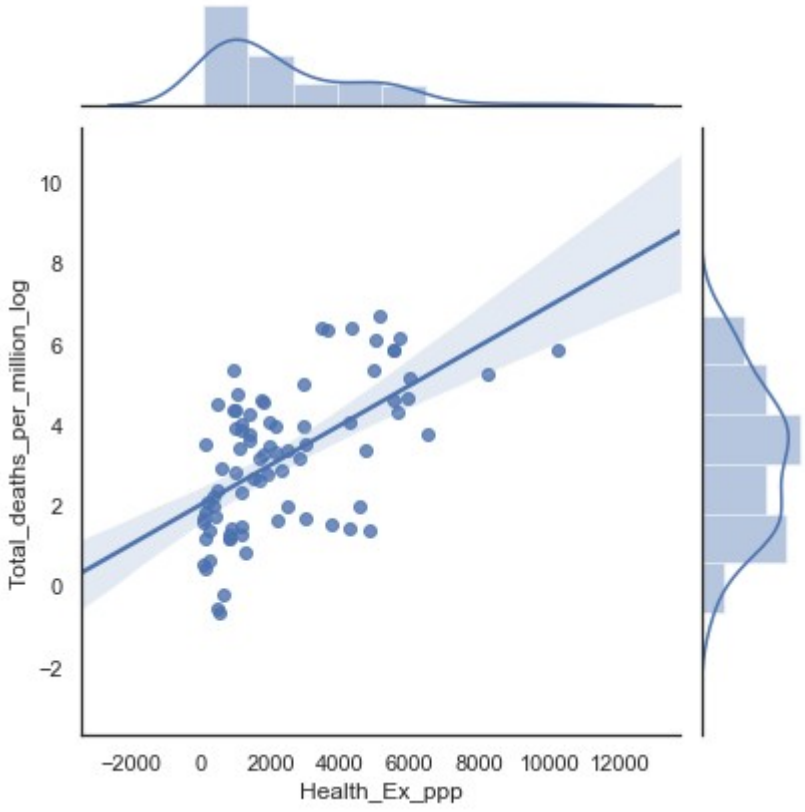
Healthcare expenditure is positively correlated (0.57) with the response variable and explains 32% of the variance:



OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.320
Model:	OLS	**Adj. R-squared:**	0.316
Method:	Least Squares	**F-statistic:**	67.91
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	9.61e-14
Time:	18:52:30	**Log-Likelihood:**	-275.41
No. Observations:	146	**AIC:**	554.8
Df Residuals:	144	**BIC:**	560.8
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.7653	0.176	10.048	0.000	1.418	2.112
Health_Ex_ppp	0.0006	7.12e-05	8.240	0.000	0.000	0.001



If we add the variable to the current model, we get the following results:

OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.486
Model:	OLS	**Adj. R-squared:**	0.468
Method:	Least Squares	**F-statistic:**	26.50
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	8.88e-19
Time:	18:59:17	**Log-Likelihood:**	-254.99
No. Observations:	146	**AIC:**	522.0
Df Residuals:	140	**BIC:**	539.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.3806	0.568	-2.430	0.016	-2.504	-0.257
DaysSince0_1MillionDeaths	0.0265	0.007	3.598	0.000	0.012	0.041
Obesity_rate	0.0527	0.017	3.193	0.002	0.020	0.085
Comorbid_pr	-25.1962	109.076	-0.231	0.818	-240.845	190.453
aged_70_older	0.0771	0.072	1.076	0.284	-0.065	0.219
Health_Ex_ppp	0.0002	0.000	1.997	0.048	2.22e-06	0.000

Healthcare expenditure becomes relevant and comorbidities and age are rendered insignificant. This is intuitive. The improvement to the model isn't significant, as can be seen when we drop the comorbidity index and age variable in favour of the healthcare expenditure variable:

OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.476
Model:	OLS	**Adj. R-squared:**	0.465
Method:	Least Squares	**F-statistic:**	42.98
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	7.98e-20
Time:	19:02:08	**Log-Likelihood:**	-256.45
No. Observations:	146	**AIC:**	520.9
Df Residuals:	142	**BIC:**	532.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.4338	0.566	-2.531	0.012	-2.553	-0.314
DaysSince0_1MillionDeaths	0.0296	0.007	4.151	0.000	0.016	0.044
Obesity_rate	0.0510	0.016	3.202	0.002	0.020	0.083
Health_Ex_ppp	0.0003	7.46e-05	4.381	0.000	0.000	0.000

5. The lockdown hypothesis

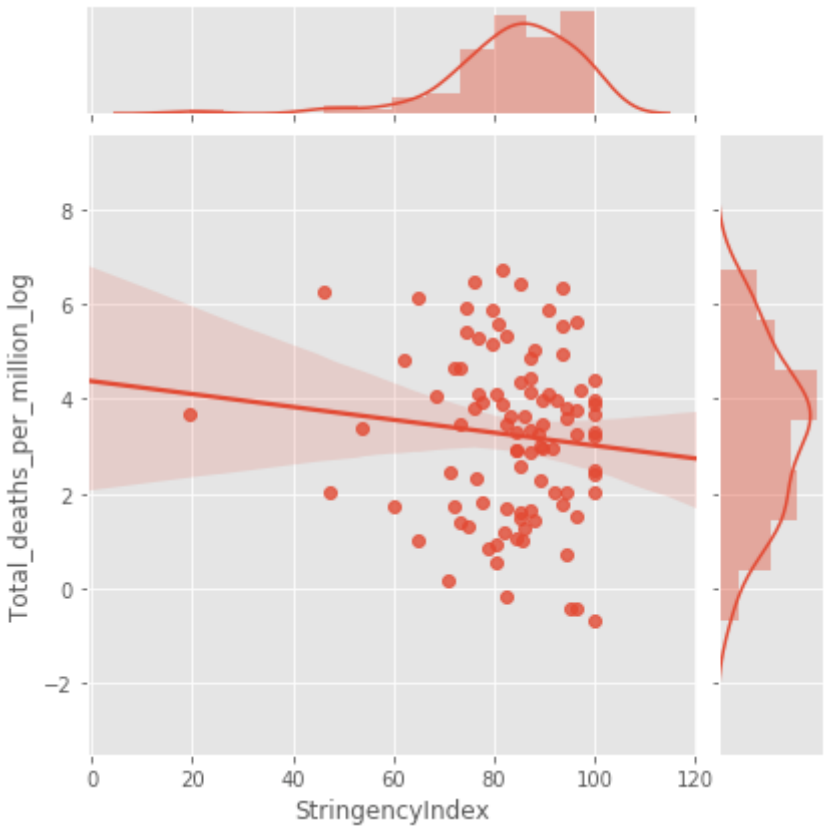
To evaluate the effectiveness of lockdowns, we sought to evaluate stringency on two levels:

- Overall impact on DPM
- Curve-flattening impact

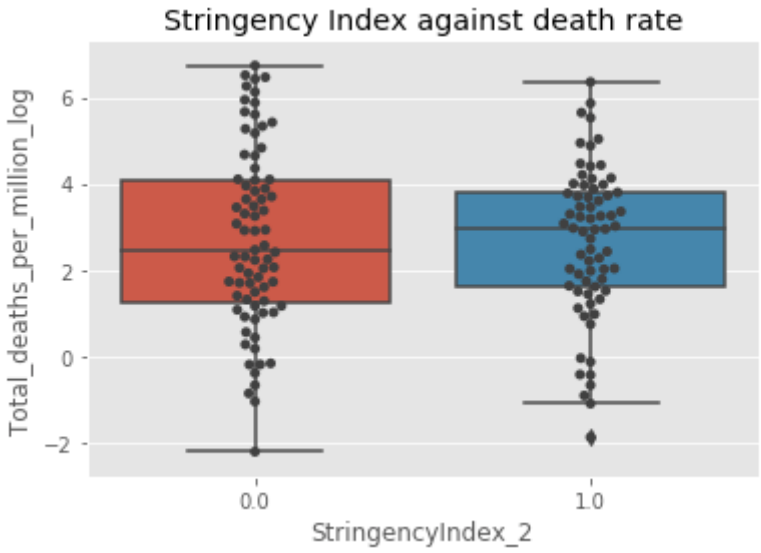
5.1. Overall impact on DPM

To evaluate the overall impact on DPM, we calculate a stringency index as the maximum per the Oxford Blavatnik School stringency model.

If we look at a scatter plot against the response variable, we see that there isn't much correlation:



Splitting the stringency index into those above the median and those below, we can confirm that there isn't much relationship between the response variable and the stringency index:



Running a univariate model on the stringency index expectedly explains none of the variance. (Note that the loss of a few cases is due to a lack of stringency data for some countries.)

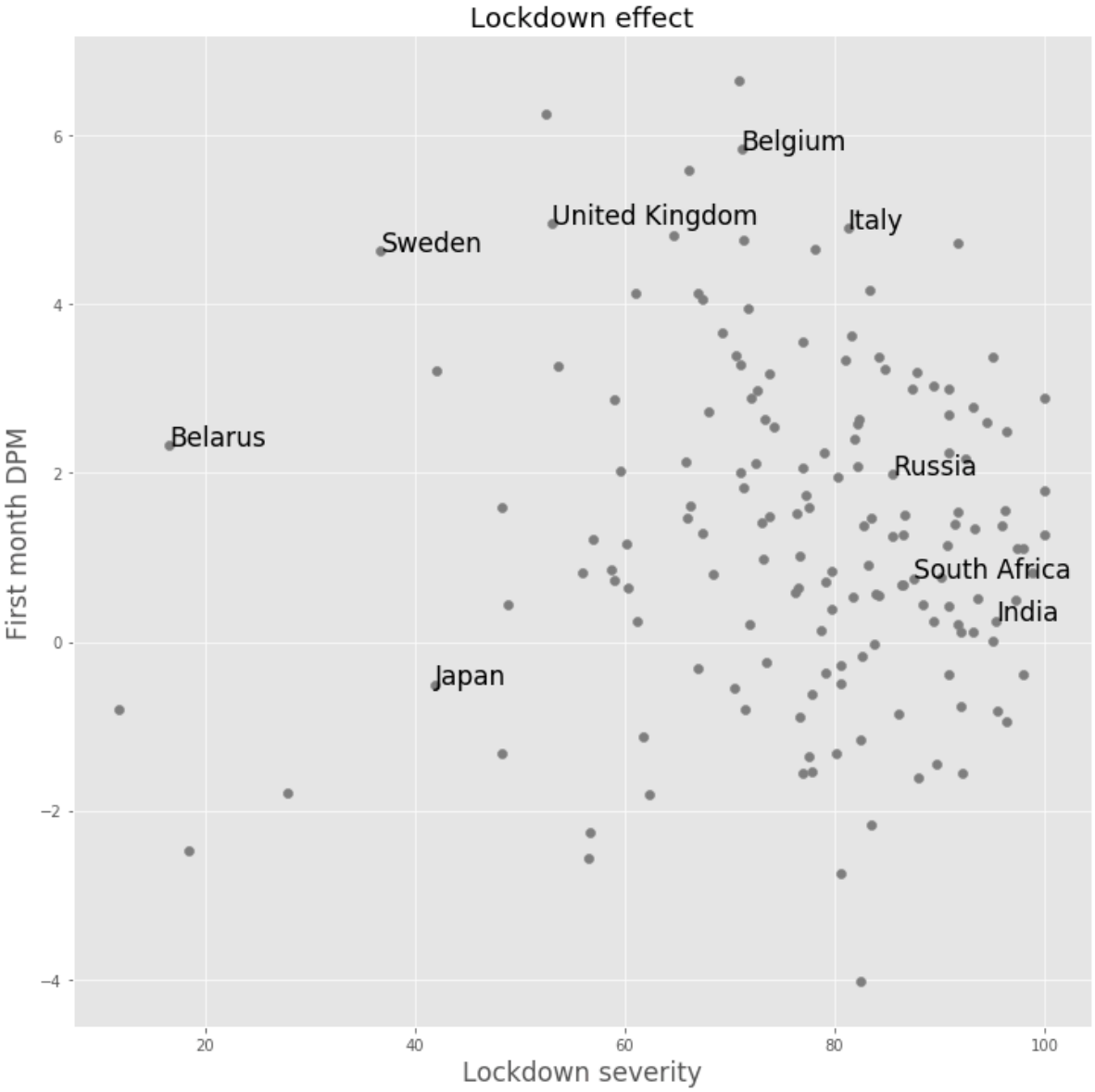
OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.002
Model:	OLS	**Adj. R-squared:**	-0.005
Method:	Least Squares	**F-statistic:**	0.2881
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	0.592
Time:	19:52:44	**Log-Likelihood:**	-276.06
No. Observations:	132	**AIC:**	556.1
Df Residuals:	130	**BIC:**	561.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.2472	1.007	3.224	0.002	1.255	5.240
StringencyIndex	-0.0064	0.012	-0.537	0.592	-0.030	0.017

5.2. Curve-flattening impact

We then created another view of the stringency hypothesis. We took the mean of the stringency index for all countries over the first 30 days after reaching 0.1 per million cumulative deaths. We then plotted this mean against the sum of the cumulative deaths per million for the same period:

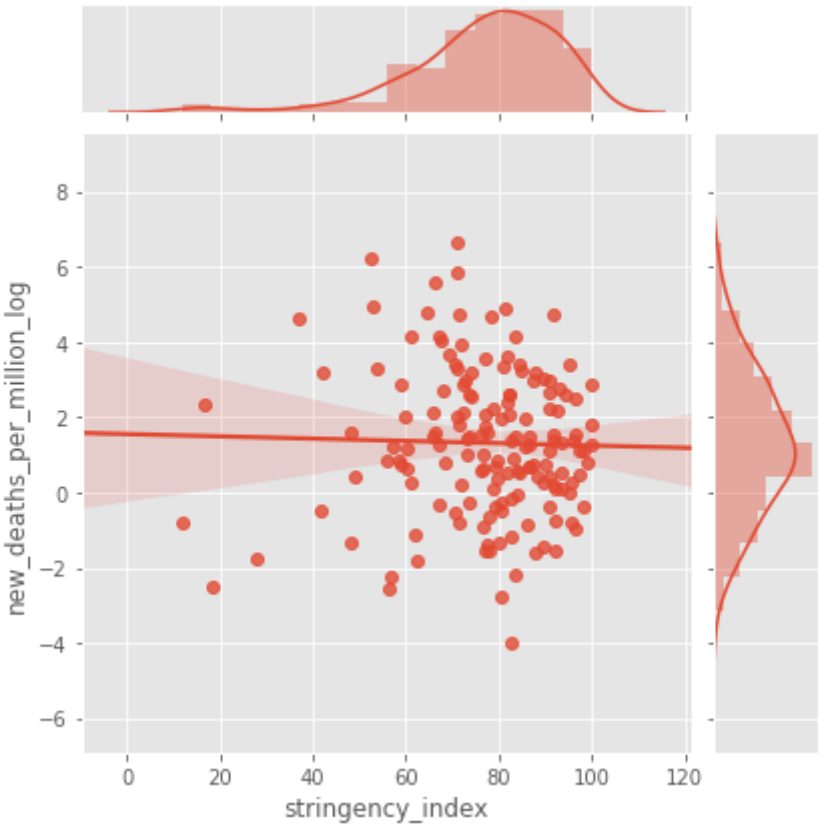


Again, there is no apparent relationship between the two variables, the regression is fruitless:

OLS Regression Results

Dep. Variable:	new_deaths_per_million_log	**R-squared:**	0.001
Model:	OLS	**Adj. R-squared:**	-0.006
Method:	Least Squares	**F-statistic:**	0.1005
Date:	Sun, 05 Jul 2020	**Prob (F-statistic):**	0.752
Time:	08:15:05	**Log-Likelihood:**	-327.37
No. Observations:	157	**AIC:**	658.7
Df Residuals:	155	**BIC:**	664.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.5543	0.752	2.066	0.040	0.068	3.040
stringency_index	-0.0030	0.010	-0.317	0.752	-0.022	0.016



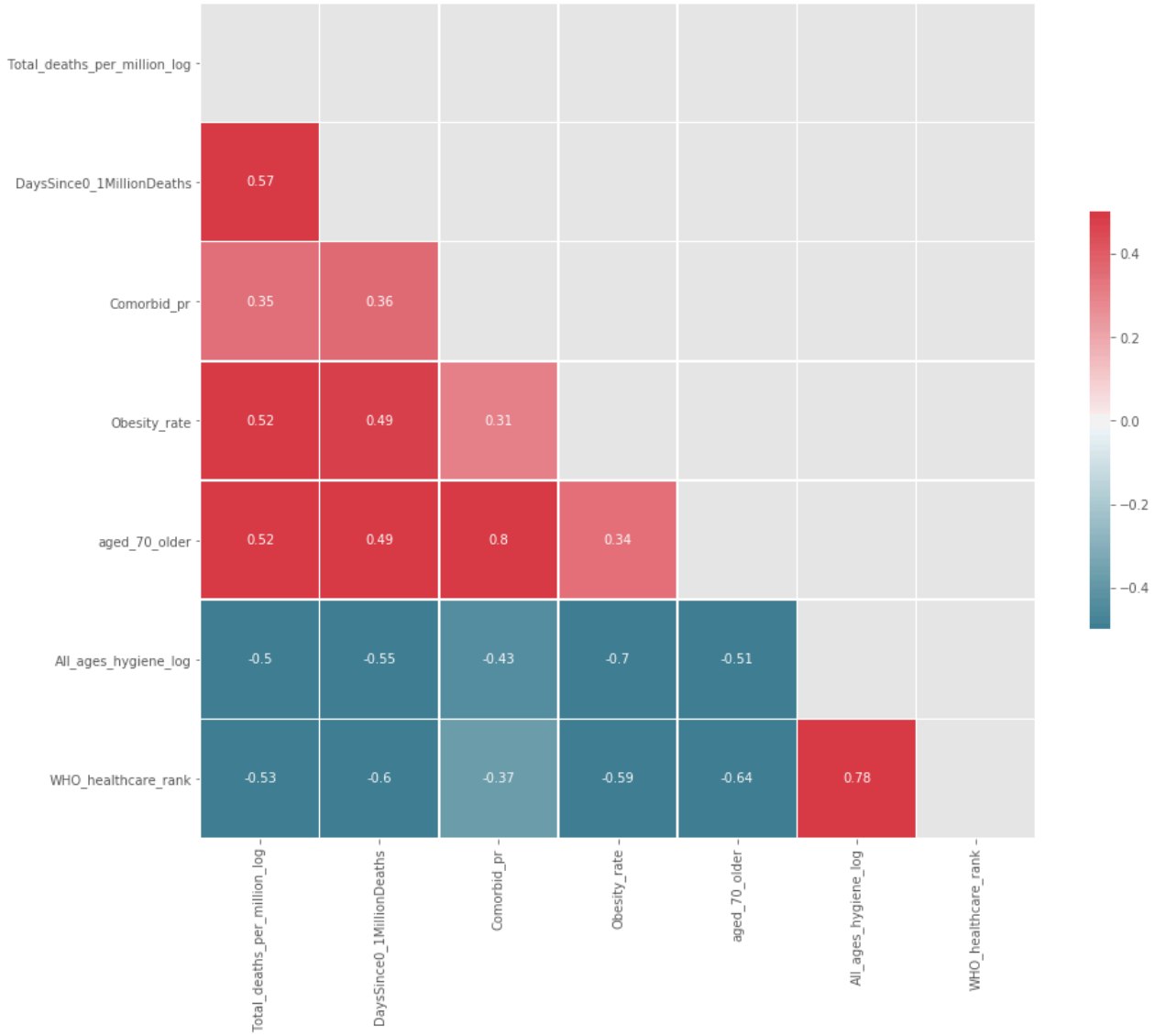
5.3. Stringency conclusion

Consistent with observations that imposition and lifting of lockdown has not been observed to effect the rate of decay of the country reproduction rates significantly, our analysis suggests there is no basis for expecting lockdown stringency to be an explanatory variable. We will continue to assess this as the few remaining pre-peak countries' epidemic curves mature over the next month or two.

In this regard we note that, for lockdowns to be expected to "flatten the curve" significantly enough to reduce the burden on healthcare systems, the impact on the response variable in 5.2 would have to be significant. We will investigate a sensible threshold, but our sense is that a correlation of less than 50% would be wholly inadequate.

6. WHO healthcare rankings

The WHO healthcare rankings[28] were entered into the model as a test for whether they had any predictive power regarding the response variable. What is immediately interesting is that the sign is counter-intuitive—the worse the rank, the lower the death rate:



This says little about the accuracy of the rankings. Quality of healthcare is likely a proxy for the hygiene effect, which we considered above.

[28] https://www.who.int/healthinfo/paper30.pdf

We include univariate and multivariate regressions for completeness:

OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.283
Model:	OLS	**Adj. R-squared:**	0.277
Method:	Least Squares	**F-statistic:**	56.31
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	6.05e-12
Time:	19:37:40	**Log-Likelihood:**	-277.89
No. Observations:	145	**AIC:**	559.8
Df Residuals:	143	**BIC:**	565.7
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.3796	0.261	16.786	0.000	3.864	4.895
WHO_healthcare_rank	-0.0185	0.002	-7.504	0.000	-0.023	-0.014

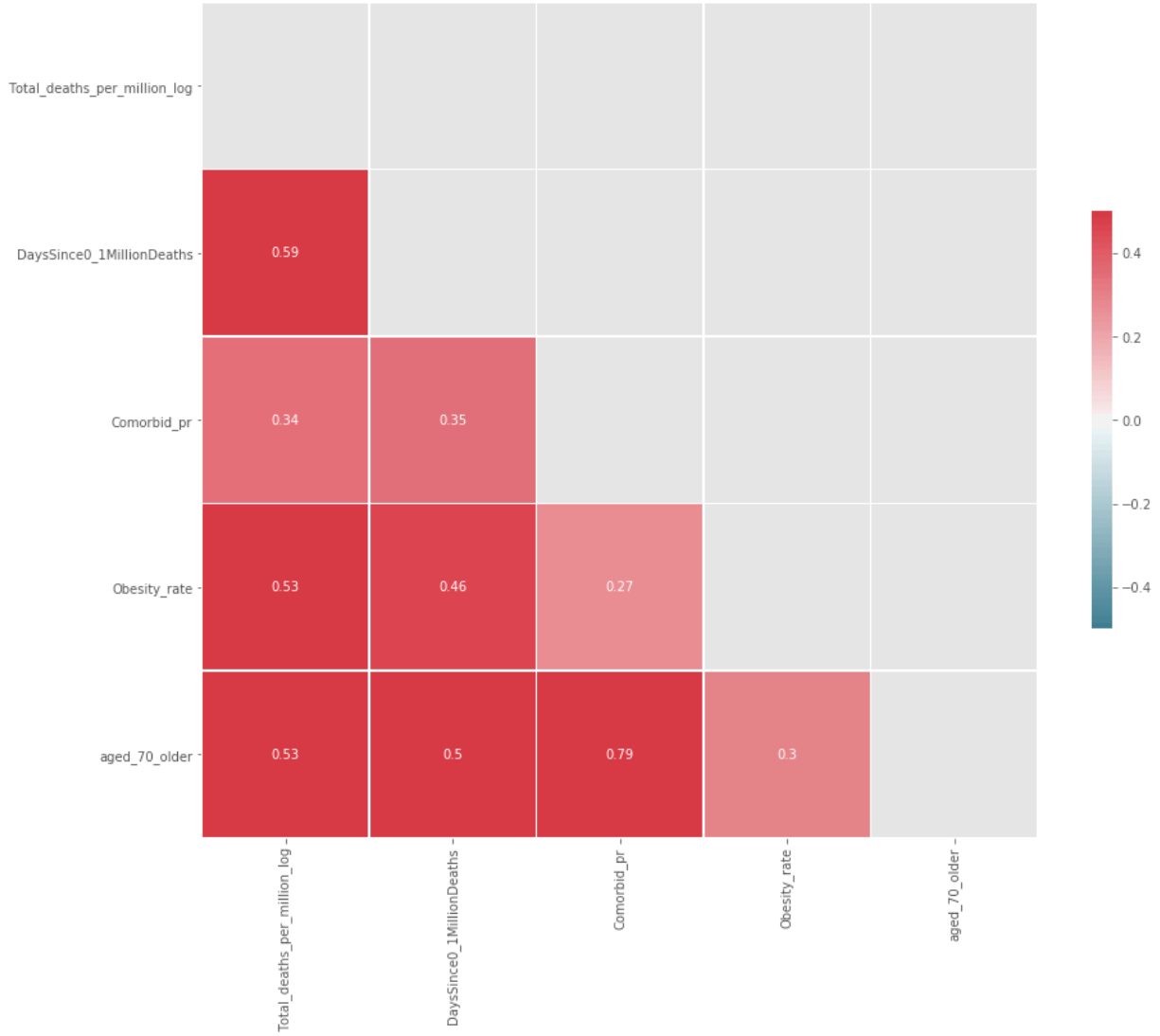
OLS Regression Results

Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.475
Model:	OLS	**Adj. R-squared:**	0.456
Method:	Least Squares	**F-statistic:**	25.10
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	5.68e-18
Time:	19:26:41	**Log-Likelihood:**	-255.32
No. Observations:	145	**AIC:**	522.6
Df Residuals:	139	**BIC:**	540.5
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.9021	0.942	-2.019	0.045	-3.765	-0.039
DaysSince0_1MillionDeaths	0.0288	0.008	3.739	0.000	0.014	0.044
Obesity_rate	0.0684	0.018	3.812	0.000	0.033	0.104
Comorbid_pr	-170.2208	96.213	-1.769	0.079	-360.450	20.008
aged_70_older	0.1998	0.059	3.386	0.001	0.083	0.316
WHO_healthcare_rank	0.0016	0.004	0.440	0.661	-0.006	0.009

7. Model summary

The current model is expressed in the following correlation plot:



Multicollinearity is not severe enough to warrant action given significance of the p-values:

	variables	VIF
0	Intercept	23.074623
1	DaysSince0_1MillionDeaths	1.582705
2	aged_70_older	3.211078
3	Comorbid_pr	2.805332
4	Obesity_rate	1.352444

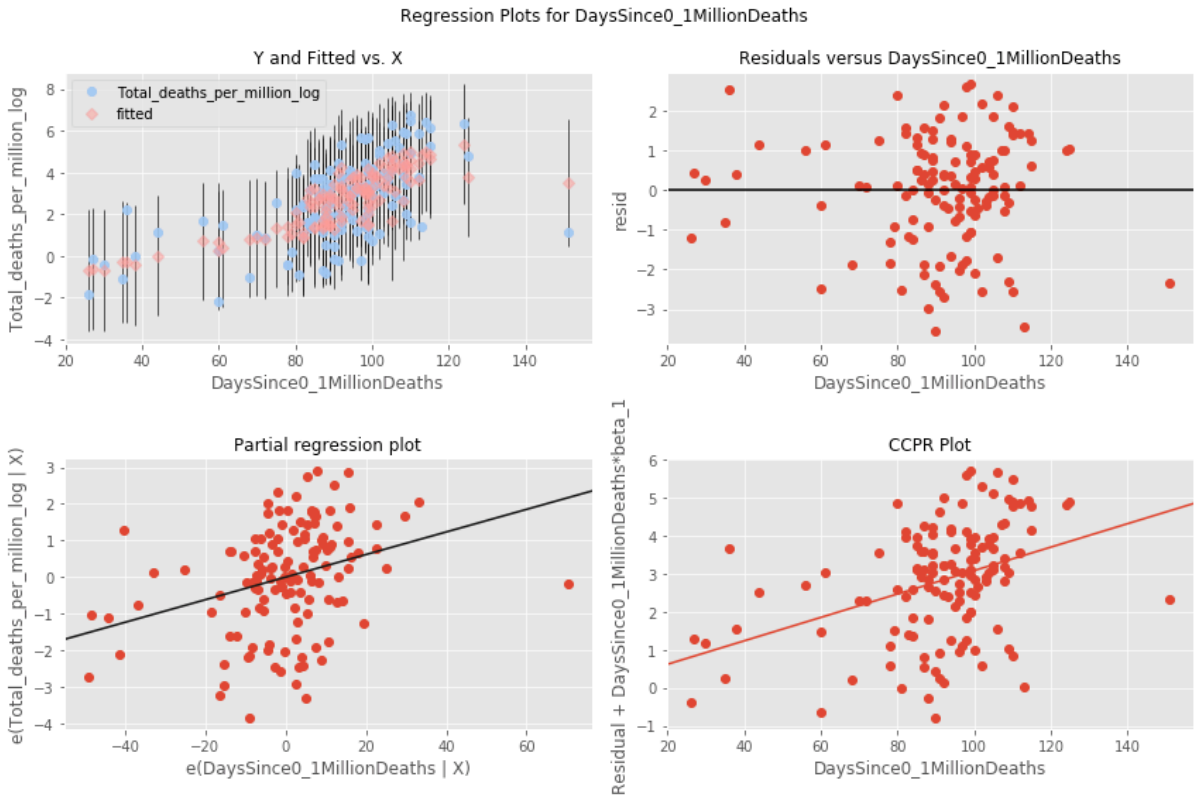
The final model accounts for 47% of the variance:

OLS Regression Results

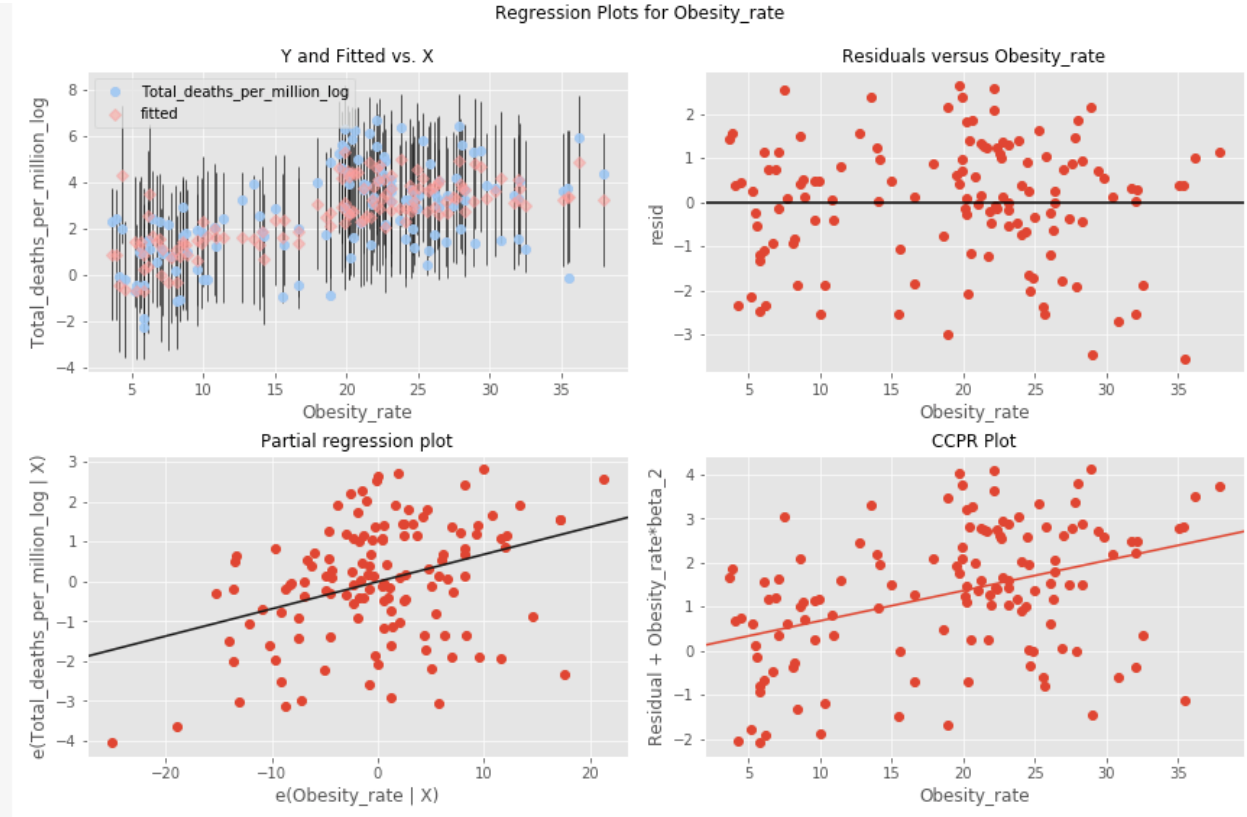
Dep. Variable:	Total_deaths_per_million_log	**R-squared:**	0.472
Model:	OLS	**Adj. R-squared:**	0.457
Method:	Least Squares	**F-statistic:**	31.46
Date:	Sat, 04 Jul 2020	**Prob (F-statistic):**	1.01e-18
Time:	07:56:36	**Log-Likelihood:**	-257.04
No. Observations:	146	**AIC:**	524.1
Df Residuals:	141	**BIC:**	539.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.5268	0.569	-2.682	0.008	-2.652	-0.401
DaysSince0_1MillionDeaths	0.0277	0.007	3.740	0.000	0.013	0.042
aged_70_older	0.1837	0.048	3.810	0.000	0.088	0.279
Comorbid_pr	-153.6794	89.004	-1.727	0.086	-329.635	22.276
Obesity_rate	0.0639	0.016	4.074	0.000	0.033	0.095

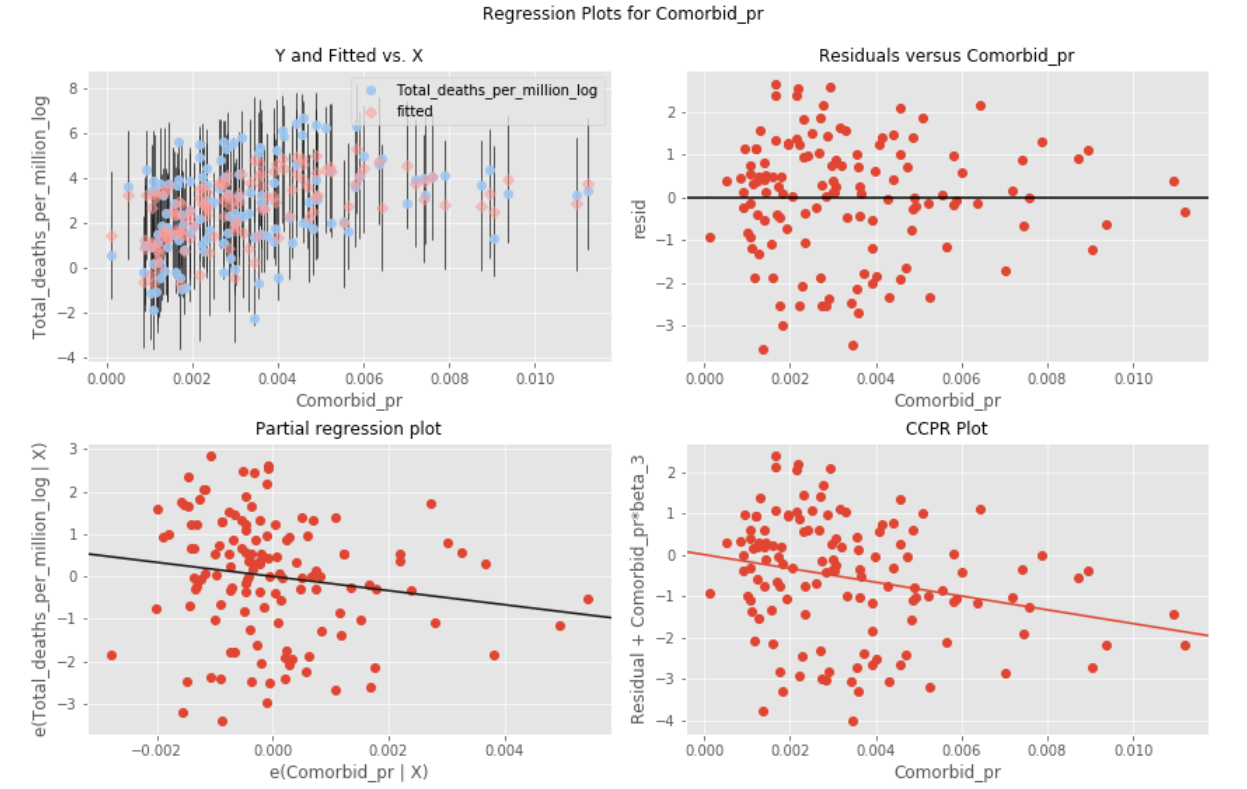
Days Since 0.1 Million Deaths variable:



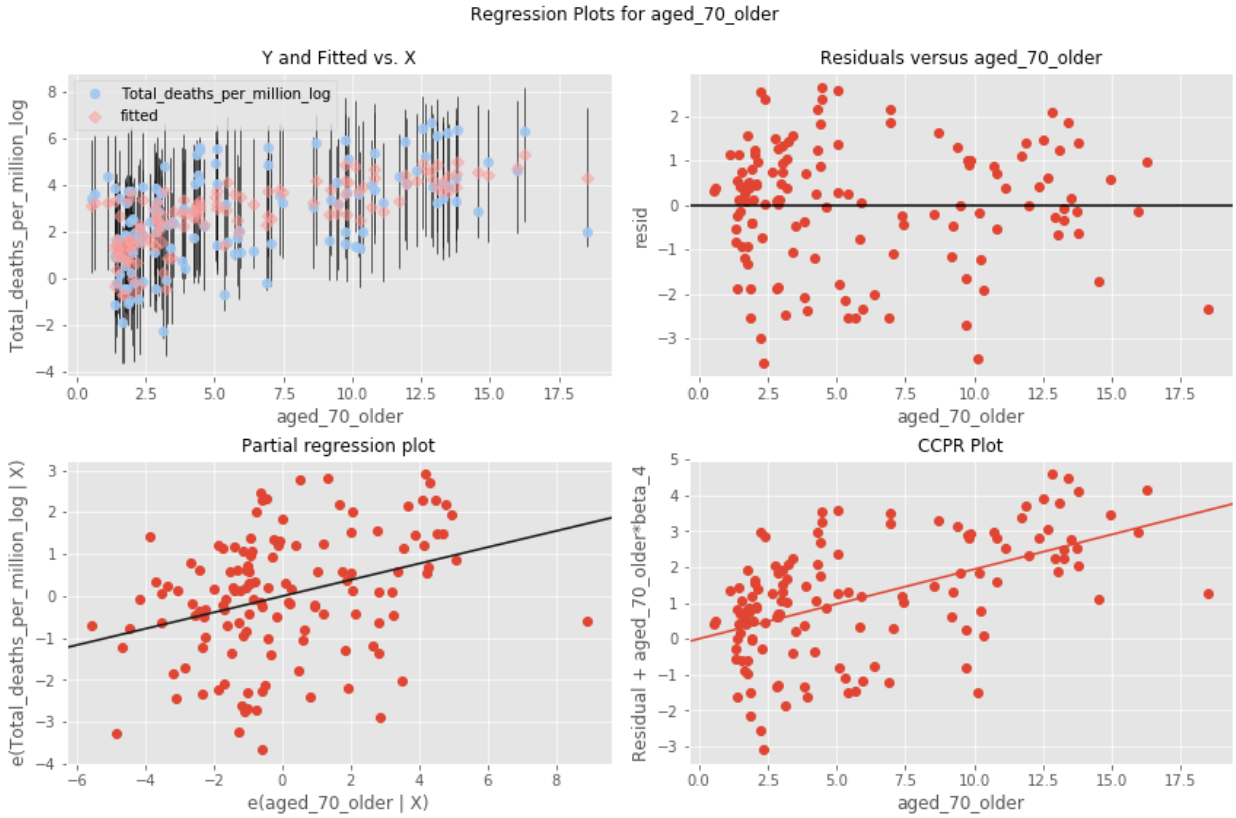
Obesity variable:



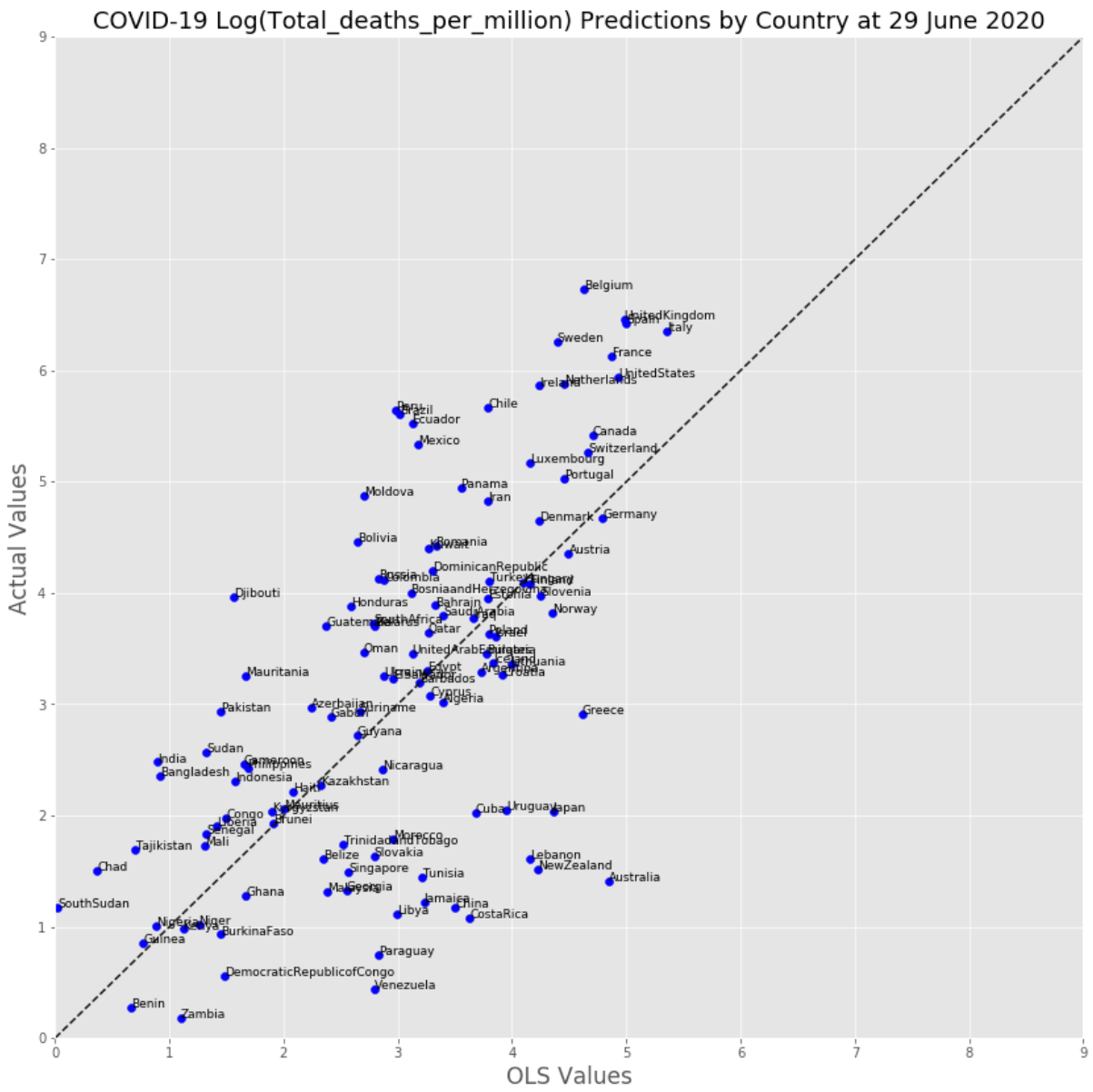
Comorbidity index:



70+ percentage variable:



Predictions (log-log) plot:



8. Conclusion

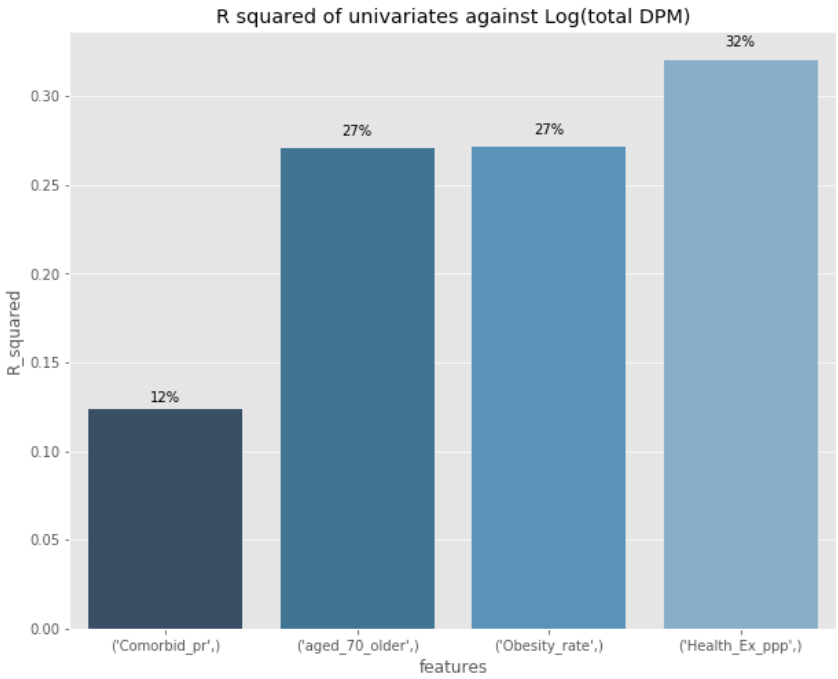
We investigated some preregistered hypotheses:

- age
- hygiene and
- comorbidity.

We demonstrated that each of them had some or other significant explanatory power in terms of the response variable, the logarithm of cumulative deaths per million, adding a time stabliser. At that point the comorbidity variable was rendered marginally insignificant on the forces of the other variables. Replacing hygiene with obesity, which enjoy a close inverse relationship, explained more of the variance and improved the comorbidity variable's significance.

Healthcare expenditure is strongly predictive on its own, but did not add to the explanatory power of the model.

Here is a bar chart showing the predictive power of the variables on their own:



We then examined lockdown stringency and WHO healthcare rankings, finding no relevance.

We will continue the search for knowledge, hoping to close the variance explanation gap even further.