Effect of Daylength on the Prevalence of COVID19 in Europe

Cathy Wyse – 18210208 – Data Programming with Python Project – December 2020

Abstract¶

Most infectious diseases show seasonal patterns of infection, generally causing more disease in winter when daylength is short. The factors generating these patterns are unclear, and it is not yet known if COVID19 will be a seasonal disease. Seasonality of disease outbreaks are modulated by complex host, environment and pathogen related factors, and in this study we investigate the associations between some of the environmental factors (daylength, temperature and latitude) on the incidence of COVID19 in European countries. Data on COVID19 cases, tests and deaths in Europe over the course of the pandemic in 2020 were matched with data on daylength and outdoor temperature for each week and location. Linear regression analysis was applied to assess the association between COVID19 infection and daylength, while adjusting for potential confounding factors (population, tests, outdoor temperature and economic status). There was some evidence for a negative linear relationship between COVID19 case numbers and daylength, but it was not possible to eliminate strong collinearity with outdoor temperature. The results of this study support an association between COVID19 and daylength, that could be mediated partly through the effects of outdoor temperature.

1 Introduction

On Christmas Eve 2019, a bronchoalveolar lavage sample was collected from a patient in China with unidentified pneumonia which was later found to contain the first genomic evidence of a novel coronavirus (SARS-CoV-2). Further patients with viral pneumonia presented to hospitals in Wuhan over the following days, and on New Year's Eve, the WHO office in China notified its regional point of contact of a case of "viral pneumonia". As 2020 dawned, the world was on the brink of a global pandemic that would infect at least 50 million people across the world and kill over 2 million by November 2020 (WHO). The pandemic initiated an unprecedented global quarantine as authorities struggled to contain the transmission of the disease caused by SARS-CoV-2, COVID19.

Strategies based on testing and isolation have proven to be the only effective interim intervention as the world awaits a vaccine, and in the absence of any effective pharmacological treatment for COVID19. Prevention of infection and spread of COVID19 does not appear to be driven by the financial capability to respond to the pandemic, with countries with high disposable income per capita, the US and UK, also suffering the highest disease burden. Increasing human population, travel and migration mean that COVID19 is likely to be the first of a succession of 21st century pandemics. Understanding the complex interaction between host, pathogen and environment that determines susceptibility to infectious disease is now one of the most important challenges for medicine and science.

Almost all infectious respiratory viruses, including established coronaviruses, cause winter-time infection, (Dowell, 2004), and despite popular belief, temperature is an unlikely cause of this increased prevalence (Eccles, 2015). For example, viral disease is seasonal in tropical regions where temperature and humidity are constant (Bloom-Feshbach 2013; Tamerius 2011) and outbreaks of influenza occur annually and simultaneously at latitudes that are oceans apart (Lofgren 2007) regardless of variations in local climatic conditions and human behaviour.

There is increasing evidence that the capacity of the immune system to defend against infection has an endogenous seasonality with reduced function in winter, and that this might drive wintertime disease outbreaks (Dowell 2001; Stevenson, 2005). In support of this, respiratory viruses can be isolated from human volunteers at all times of the year, even though the diseases they cause occur in winter (Lee 2012). Seasonality is one of the most predictable aspects of viral infection, and one of the least understood. Studies of the host and environmental factors that generate seasonality might help curtail transmission of SARS-CoV2 and prevent escalation of the inevitable future outbreaks of novel viral infections to pandemics. It is not yet known if COVID19 is a seasonal disease, but this is highly likely as all existing circulating coronaviruses cause seasonal disease (Li, 2020).

1.1 Objective

The objective of this study is to investigate if the prevalence of COVID19 infection during the 2020 pandemic was associated with daylength, independent of confounding factors such as temperature, testing, population and economic status.

2. Methods

2.1 Data Import and Processing

The data on EU incidence of COVID-19 were checked to confirm csv format, and then processed to generate a single dataset relating cases, tests and deaths per week to changes in outdoor temperature and latitude in European Countries affected by COVID19. Missing data were detected by searching for a panel of possibilities including 'NA', 'NULL', punctuation mark, and re-coded as missing values.

Unemployment and income status were used as proxy measures of the economic status of each country. Data on latitude, daylength and outdoor temperature were acquired from information provided by the Global Monitoring Division of the US Government National Oceanic and Atmospheric Administration. The date recorded for case numbers and deaths, and for environmental data was standardised across all the datasets to average numbers per week. The unit of time used in all further analysis was week number during 2020, ranging from 1-42. The variable “income group” and “country code” were recoded as categorical variables, with 5 and 42 levels, respectively.

The codes used to indicate country were standardised across datasets to three-digit (iso3 format. At this point, it was possible to merge all data with reference to (i) country name and (ii) week number to yield a final table of COVID19 cases and deaths, by country, week and environmental factors (daylength, latitude)

All analyses were implemented using Python 3.8 programming language, using *pandas, Matplotlib, seaborn, linearmodels, and statsmodels* libraries.

2.2 Outcome Variable

It was elected to confine the analyses to times of evidence of ongoing exposure to COVID19, to minimise the modulation of the outcome variable by non-environmental factors linked to disease exposure. Ongoing exposure was defined as >1 death per week. The effects of environmental factors such as daylength are more easily assessed by investigating periods of active infection. In other words, susceptibility to infection can only be assessed while the disease is spreading. There are no data available on social and travel restrictions in this study, so it was not possible to derive a more accurate measure of the possibility of exposure to SARS-CoV-2.

2.2 Linear Modelling

Linear regression models were applied to investigate the association between predictor variables and the study outcome (COVID19 cases). Ordinary least squares linear regression models were applied as follows:

$$y_i = \alpha_i + \beta x_i + \epsilon_i$$

where y_i is an outcome variable, and β and ϵ_i are regression coefficients and error terms, respectively. Mixed models that incorporated time and country as random effects were also applied:

$$y_{it} = \alpha_i + \gamma_t + \beta' x_{it} + \epsilon_{it}$$

where α_i is the country effect and γ_i is the time effect, β contains the variables of interest, temperature, etc, α_i are country-specific components and ϵ_{it} are errors that are independent of α_i and the covariates x_{it}. This model allowed for the autocorrelation caused by inclusion of the week number variable in the model (time entity) and for the non-independence induced by including data from multiple countries (entity).

Linearity, equal variance and normality of residuals were investigated to ensure that the data did not violate the assumptions of linear regression. Data are presented as mean and sd, or median and interquartile range, as appropriate. Statistical significance was accepted at $p<0.05$.

3. Results

3.1 Exploratory Data Analysis

Descriptive data for COVID19 incidence/week over the course of 2020 (42 weeks in total), are shown in Table 1, below. Data are presented as median and interquartile range, due to the skewed distribution of the case, deaths, and tests/week variables. Three countries presented data for all 42 weeks of the study, Italy, Finland and the Czech Republic where the virus first spread into Europe from China. There was considerable variation in the number of cases per week in European countries, even considering population differences. The highest median weekly death rate was recorded in the UK (56.3 [16.3,287.6], median [IQR]), as was the highest median cases identified in a week, 1622.9 [945.0,4227.2]. Data describing the daily photoperiod (daylength, number of hours of daylight/24h) and on outdoor temperature in the EU countries are shown in Table 2. Europe spans most of the populated areas of the Northern hemisphere, and this is reflected by latitudes ranging from 60^{o}N (Iceland) to 35^{o}N (Cyprus). Mean temperature over the study period ranged from 4^{o}C (Finland) to 23^{o}C in Cyprus.

The number of hours of daylight per day (daylength) and temperature vary as a function of latitude due to the effect of the curvature of the Earth on its proximity to the Sun. The continuous variables in this study were plotted on scatterplots and on a heatmap-correlation matrix, to investigate their sampling distribution and interactions (Figure 1-2). Daylength was normally distributed, but temperature showed a negative skew; evidence of some divergence of these two correlated variables. Latitude shows a positive skew, reflecting the Northerly location of most of the European countries. The data for Cases, Deaths and Tests are not normally distributed. The correlation coefficient matrix and scatterplots show clear evidence for linear relationships between temperature and daylength (r = 0.58), between daylength and latitude (r=0.20), and temperature and latitude (r = -0.50). The relationships between Cases, Tests and Deaths and the other variables are difficult to interpret due to their skewed distribution, but their correlation coefficients are evidence for close associations between these parameters (Figure 2)

Table 1 COVID19 cases, tests and deaths in European countries

	AUT	BEL	BGR	CYP	CZE	DEU	DNK	ESP	EST	FIN	FRA	GBR	GRC	HRV	HUN
Weeks (n)	26	33	18	31	42	30	36	37	36	42	33	28	31	37	32
Cases, median [Q1,Q3]	101.7 [45.8, 267.9]	451.4 [167.1,858.9]	145.9 [128.9,215.0]	7.0 [2.4,13.9]	98.1 [9.5,232.3]	1024.3 [512.0,1791.5]	74.1 [32.0,140.2]	1261.3 [339.7,6177.3]	8.2 [2.1,24.5]	20.6 [3.1,58.4]	1125.0 [540.4,3597.9]	1622.9 [945.0,4227.2]	42.7 [18.4,203.3]	45.6 [2.7,82.6]	34.1 [13.3,78.3]
Deaths, median [Q1,Q3]	1.6 [0.6,3.0]	6.6 [3.1,33.0]	6.1 [4.1,7.4]	0.0 [0.0,0.1]	1.1 [0.0,3.1]	10.2 [5.1,47.2]	0.5 [0.3,2.4]	24.7 [1.7,157.4]	0.0 [0.0,0.1]	0.1 [0.0,0.8]	36.6 [13.7,98.4]	56.3 [16.3,287.6]	1.6 [0.7,3.0]	0.9 [0.0,1.9]	1.6 [0.7,7.2]
Tests, median [Q1,Q3]	54241.5 [41335.5,76953.0]	93868.0 [75991.0,145319.0]	28138.5 [22719.2,32508.5]	14137.0 [8770.0,17728.5]	32274.0 [273.0,45312.2]	450039.5 [362108.8,853033.8]	92387.5 [23387.0,141175.2]	235870.0 [159411.0,391216.0]	6451.5 [2741.8,8381.5]	16946.0 [982.5,33462.5]	243813.0 [140316.0,551668.0]	813554.5 [590499.2,1215772.2]	32223.0 [7557.0,74488.0]	8950.0 [2511.0,11346.0]	18565.0 [13417.0,26807.2]

	IRL	ISL	ITA	LTU	LUX	LVA	MLT	NLD	POL	PRT	ROU	SVK	SVN	SWE
Weeks (n)	36	32	42	31	33	32	32	31	31	32	31	32	32	37
Cases, median [Q1,Q3]	77.3 [15.6,239.2]	4.7 [0.7,10.6]	476.6 [187.1,1602.8]	13.1 [5.1,35.0]	38.1 [7.6,77.3]	7.1 [2.7,12.1]	8.0 [2.2,26.2]	435.1 [167.4,928.3]	372.3 [299.7,600.3]	313.2 [214.0,535.1]	346.9 [258.9,1194.4]	22.1 [6.6,55.1]	18.2 [6.4,33.5]	293.3 [172.7,552.9]
Deaths, median [Q1,Q3]	1.3 [0.2,5.4]	0.0 [0.0,0.0]	16.6 [5.6,125.8]	0.3 [0.0,0.7]	0.1 [0.0,0.7]	0.1 [0.0,0.3]	0.0 [0.0,0.1]	6.6 [2.0,27.8]	11.6 [8.4,17.9]	5.9 [3.1,12.8]	22.1 [14.1,39.4]	0.0 [0.0,0.3]	0.3 [0.1,0.6]	4.0 [1.4,38.4]
Tests, median [Q1,Q3]	34452.0 [18513.2,52749.8]	7809.5 [3092.8,13220.0]	323096.0 [40425.0,409384.5]	20941.0 [11137.5,35276.0]	28763.0 [7121.0,40638.0]	10882.5 [8905.5,12568.2]	7324.0 [5850.0,13200.2]	65541.0 [33279.0,141713.5]	143345.0 [85851.0,157244.5]	94829.5 [84752.0,98557.5]	68223.0 [51538.0,134895.5]	13772.5 [9010.5,22820.0]	6882.0 [5538.8,7846.8]	36466.0 [12349.0,69529.0]

Table 2 Demographic and environmental factors in European countries

	AUT	BEL	BGR	CYP	CZE	DEU	DNK	ESP	EST	FIN	FRA	GBR	GRC	HRV	HUN
Weeks (n)	26	33	18	31	42	30	36	37	36	42	33	28	31	37	32
Population	8858775	11455519	7000039	875899	10649800	83019213	5806081	46937060	1324820	5517919	67012883	66647112	10724599	4076246	9772756
Latitude	47.5	50.5	42.7	35.1	49.8	51.2	56.3	40.5	58.6	61.9	46.2	55.4	39.1	45.1	47.2
Unemployment %	4.7	5.6	4.3	7.3	1.9	3	4.9	14	5.1	6.6	8.4	3.9	17.2	6.9	3.4
Temperature, mean (SD)	13.6 (3.5)	13.8 (4.5)	20.2 (3.5)	23.3 (5.0)	9.9 (7.6)	14.0 (4.5)	10.6 (6.3)	16.3 (5.8)	9.3 (7.7)	4.2 (9.6)	14.8 (4.5)	12.3 (2.6)	20.2 (5.2)	14.7 (6.6)	16.1 (5.3)

	IRL	ISL	ITA	LTU	LUX	LVA	MLT	NLD	POL	PRT	ROU	SVK	SVN	SWE
Weeks (n)	36	32	42	31	33	32	32	31	31	32	31	32	32	37
Population, mean (SD)	4904240	356991	60359546	2794184	613894	1919968	493559	17282163	37972812	10276617	19414458	5450421	2080908	10230185
Latitude, mean (SD)	53.4	65	41.9	55.2	49.8	56.9	35.9	52.1	51.9	39.4	45.9	48.7	46.1	60.1
Unemployment, mean (SD)	4.9	2.8	9.9	6.4	5.4	6.5	3.5	3.2	3.5	6.3	4	5.6	4.2	6.5
Temperature, mean (SD)	11.1 (3.4)		15.3 (6.5)	11.2 (7.0)		11.5 (6.1)		13.7 (3.9)	13.4 (5.4)	13.4 (5.3)	15.5 (5.4)	12.7 (5.2)	14.8 (5.0)	5.7 (8.1)

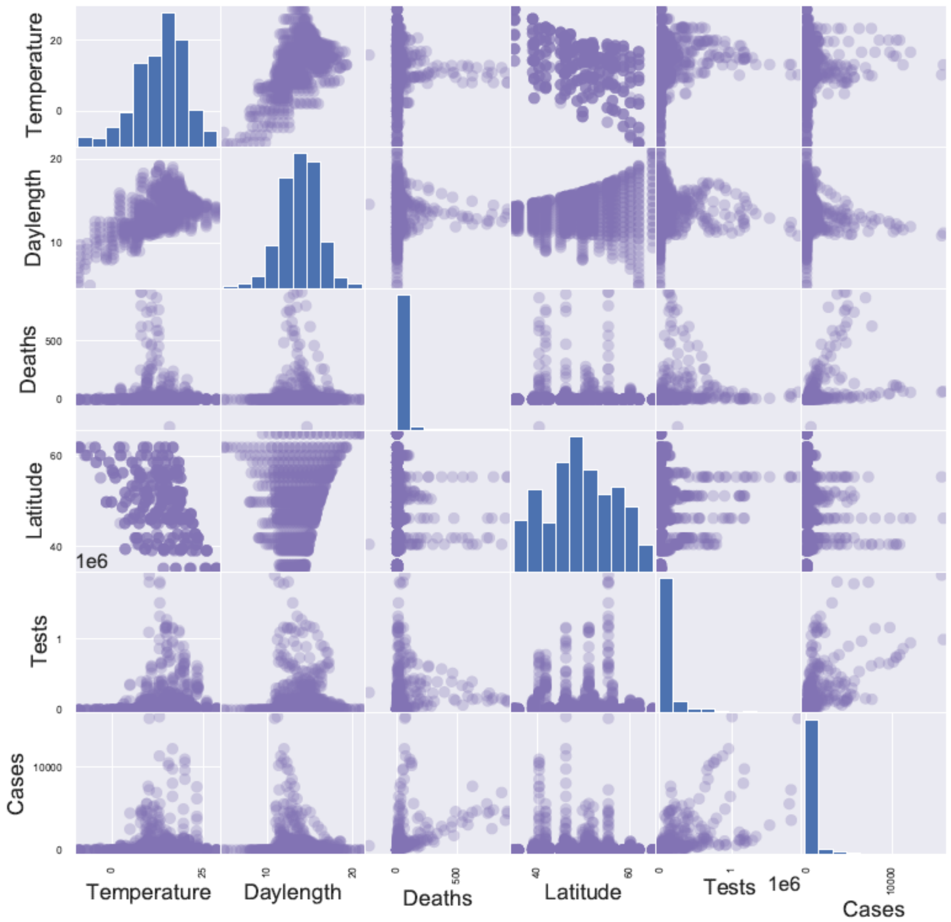


Figure 1 Histograms and scatterplots illustrating the sampling distributions and associations among the continuous variables in this study

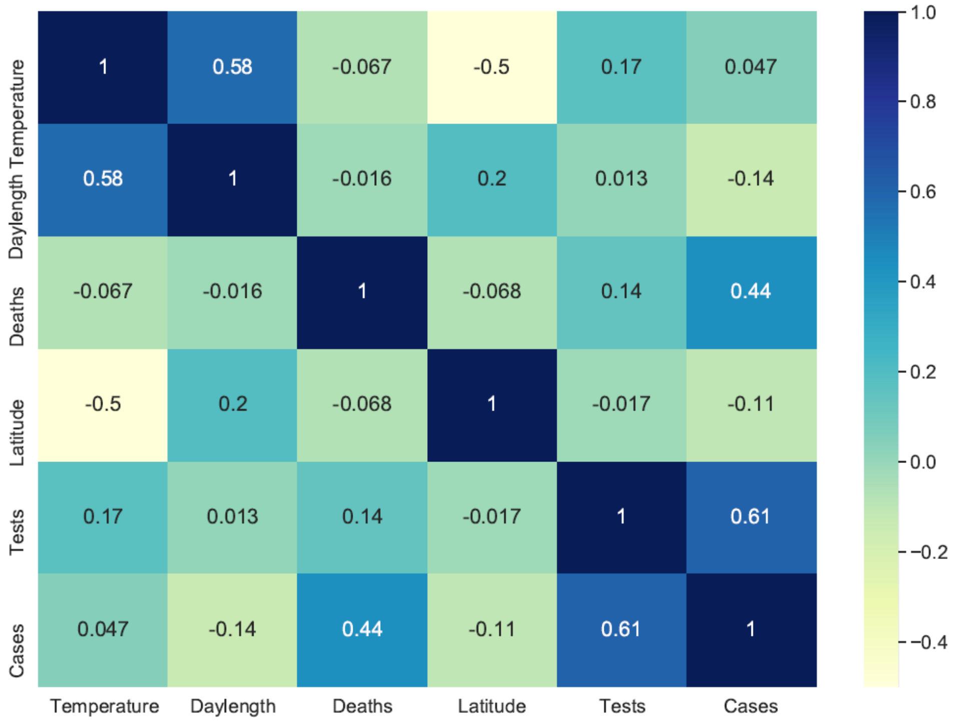


Figure 2 Heatmap matrix showing correlation coefficient (r) for the associations among the continuous variables in this study.

The skewed distribution of the Cases, Tests and Deaths variables were further examined using histogram plots of the raw and log-transformed data. With the exception of the Deaths data, the skewed log-normal distributions followed a more bell-shaped distribution following log(1+x) transformation (Figure 2). As described in the methods section, the analysis was confined to weeks that included more than one death.

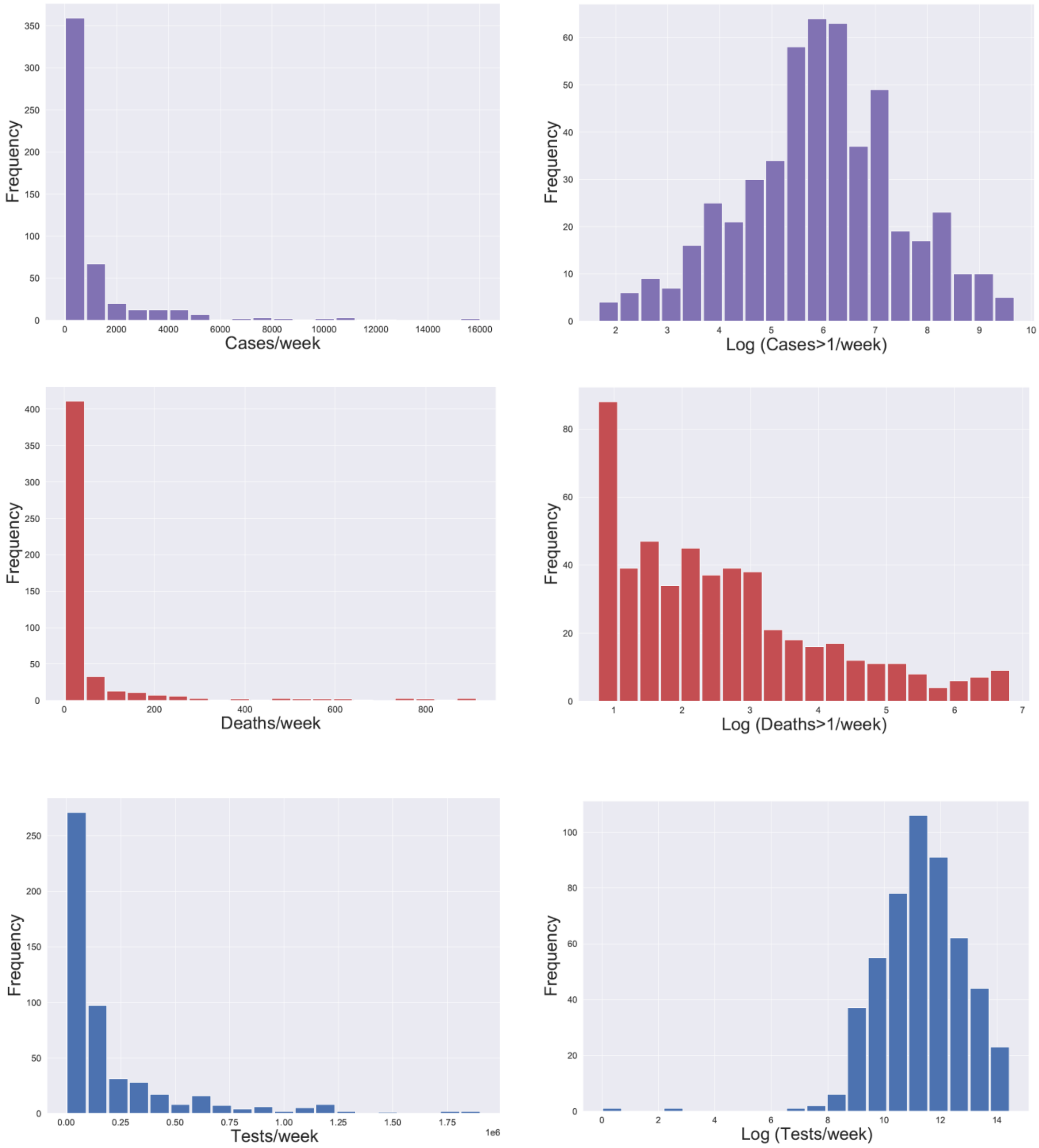


Figure 2: Histograms showing the sampling distribution of raw and log transformed data for Cases, Deaths and Tests

The complex relationships between latitude, temperature and daylength that are addressed in this study are illustrated in Figure 3.

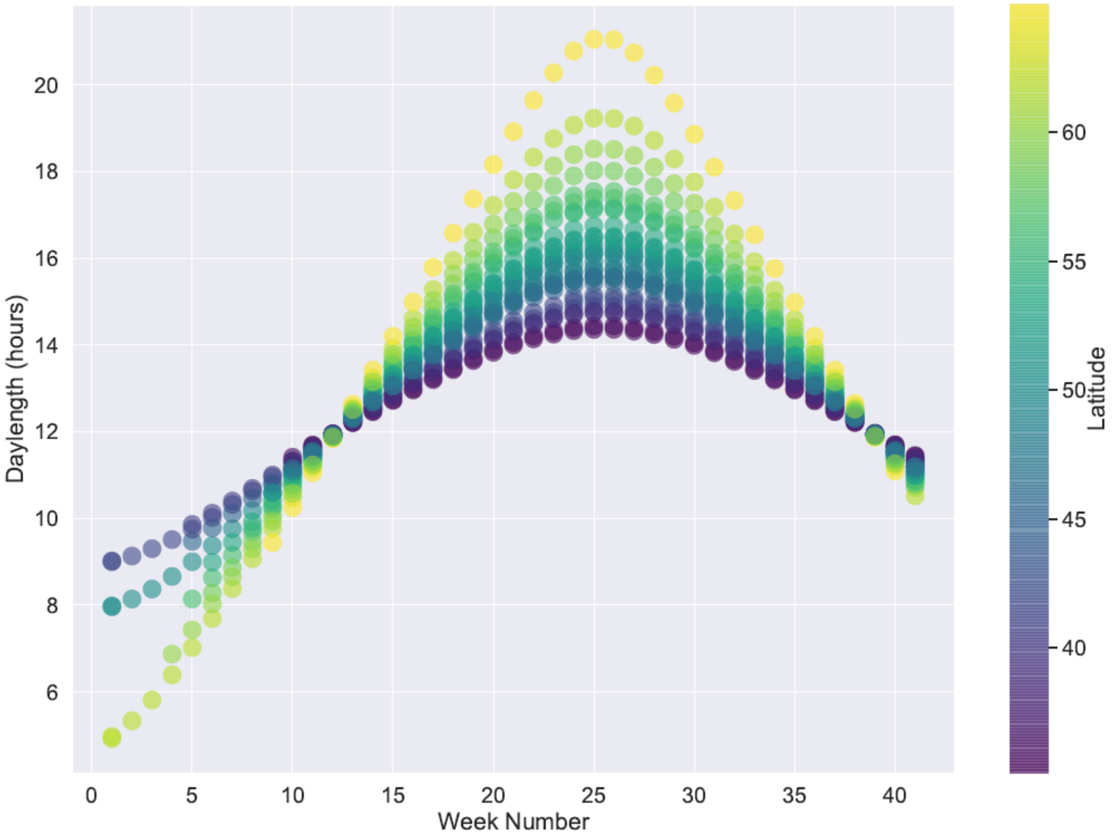


Figure 3 The relationship between daylength and latitude over the course of the 42 weeks of 2020 in this study. The rate of change of daylength is faster at the poles due to the curvature of the Earth, which explains the increasing amplitude of the latitudinal clines plotted here.

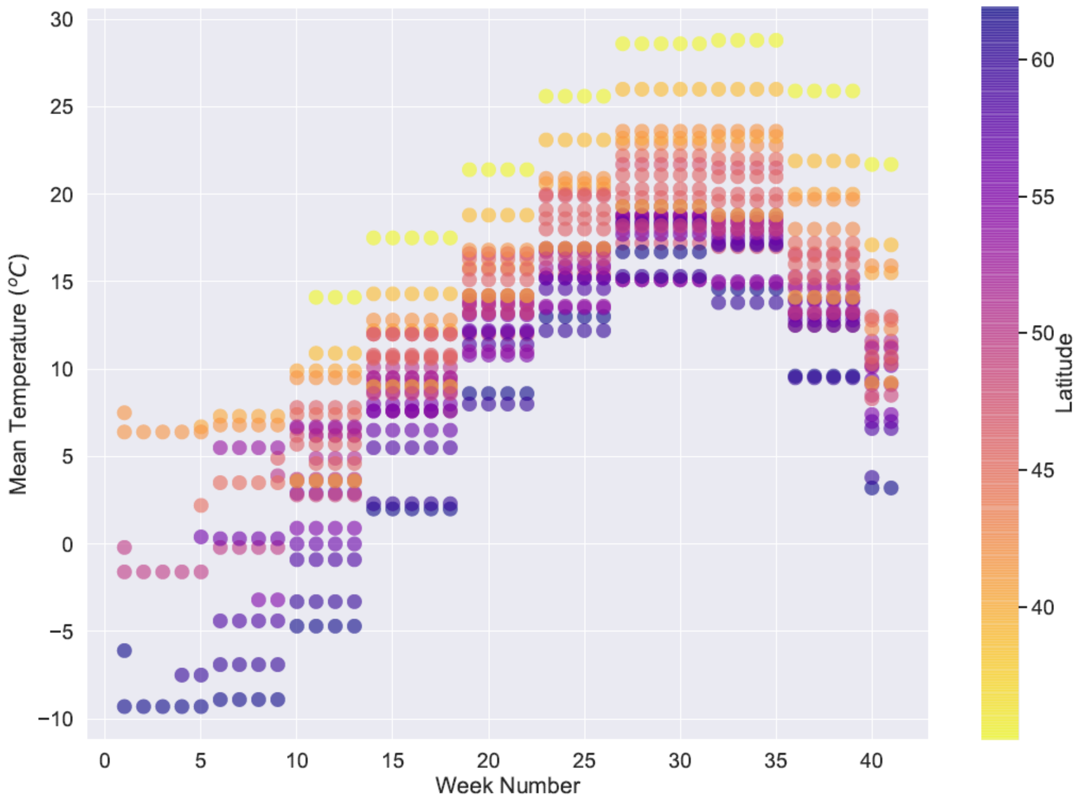


Figure 4 The relationship between temperature and latitude over the course of the 42 weeks of 2020 in this study. The pattern of change in temperature over the course of this study is also related to latitude and follows a similar but not identical pattern to that of daylength in Figure 3.

3.2 Associations between COVID19 and Temperature, Latitude and Daylength

The data for Cases was taken forward for investigation of associations between COVID19 and the environmental parameters. The association was examined graphically by plotting the log transformed data against Daylength and Temperature.

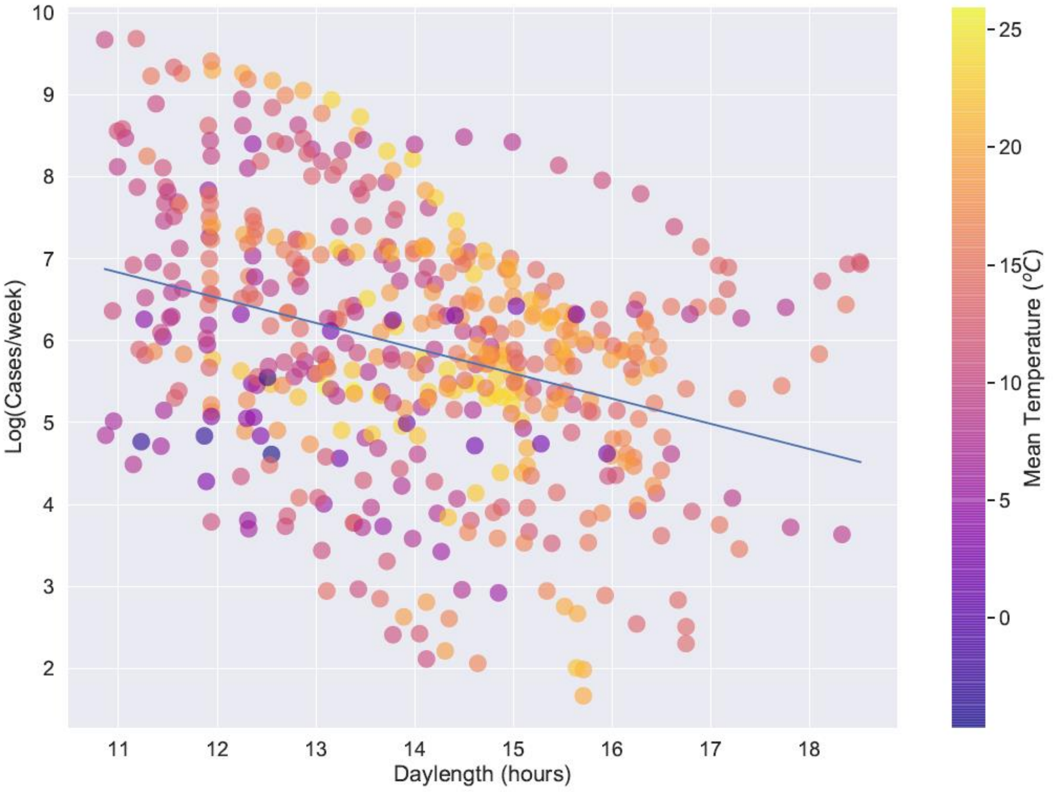


Figure 5 The relationship between Cases/week and Daylength shows evidence of a linear trend, but this does appear to be an effect shared with temperature, as seen from the colour gradient

3.3 Multiple Linear Regression Modelling

The obvious multicollinearity between temperature, latitude and daylength precludes including all of these variables in a linear regression model. The degree of multicollinearity was investigated by calculating variance inflation factors and using these parameters to indicate which variables could be included. The VIF for predictor variables included in two models were compared, Model 1 containing daylength and Model 2 containing latitude. Model 2 shows the least evidence of collinearity (Table 3) VIF between 5 and 10 indicates strong multicollinearity, but does not preclude further linear regression modelling of these covariables (O'Brien, 2007). VIF > 10 means that estimates of linear regression parameters can not be interpreted.

Table 3 Variance Inflation Factors for the predictor variables

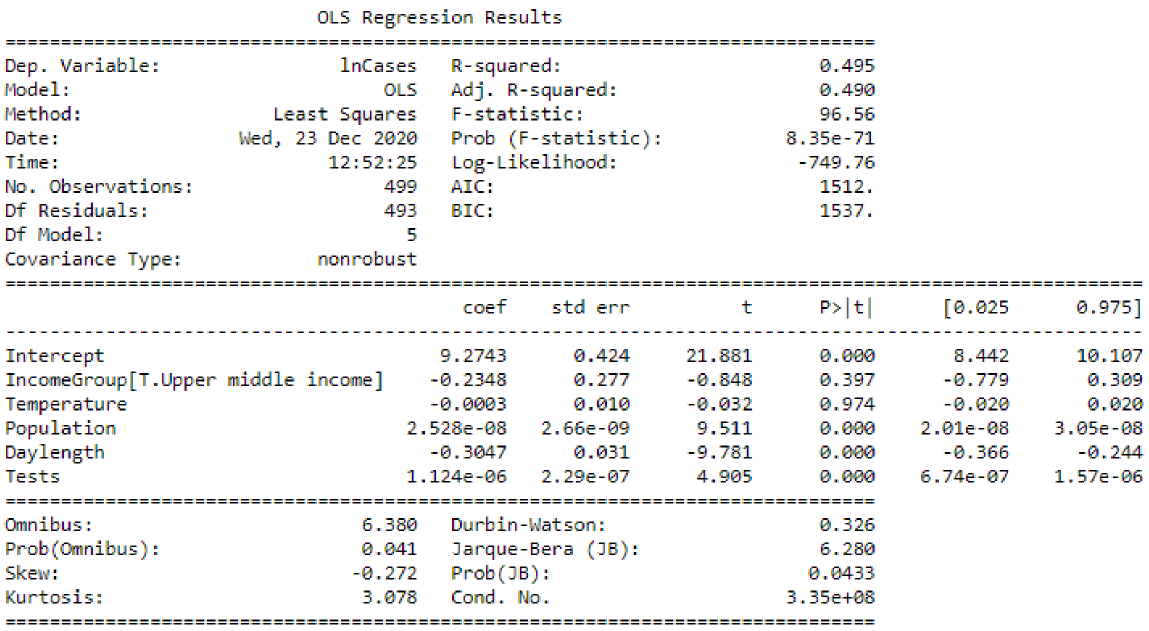
MODEL 1		**MODEL 2**	
Feature	*VIF*	*Feature*	*VIF*
Temperature	14.43	Temperature	9.78
Population	4.55	Population	4.39
Tests	3.38	Tests	2.94
Unemployment	3.99	Unemployment	3.98
Daylength	133.1	Daylength	9.57
Latitude	84.8		

Model 1 was defined as:

$$y_i = \beta_1 Temperature + \beta_2 Population + \beta_3 Tests + \beta_4 Unemployment + \beta_5 Daylength + \epsilon_i$$

Where y_i are log(1-cases/week), ϵ_i is an error term, and β_1- β_4 are slope coefficients

Table 4 Regression analysis of the relationship between cases/week and the predictor variables in this study



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                lnCases   R-squared:                       0.495
Model:                            OLS   Adj. R-squared:                  0.490
Method:                 Least Squares   F-statistic:                     96.56
Date:                Wed, 23 Dec 2020   Prob (F-statistic):           8.35e-71
Time:                        12:52:25   Log-Likelihood:                -749.76
No. Observations:                 499   AIC:                             1512.
Df Residuals:                     493   BIC:                             1537.
Df Model:                           5
Covariance Type:            nonrobust
========================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------------
Intercept                               9.2743      0.424     21.881      0.000       8.442      10.107
IncomeGroup[T.Upper middle income]     -0.2348      0.277     -0.848      0.397      -0.779       0.309
Temperature                            -0.0003      0.010     -0.032      0.974      -0.020       0.020
Population                           2.528e-08   2.66e-09      9.511      0.000    2.01e-08    3.05e-08
Daylength                              -0.3047      0.031     -9.781      0.000      -0.366      -0.244
Tests                                1.124e-06   2.29e-07      4.905      0.000    6.74e-07    1.57e-06
==============================================================================
Omnibus:                        6.380   Durbin-Watson:                   0.326
Prob(Omnibus):                  0.041   Jarque-Bera (JB):                6.280
Skew:                          -0.272   Prob(JB):                       0.0433
Kurtosis:                       3.078   Cond. No.                     3.35e+08
==============================================================================
```

Ordinary least squares (OLS) regression analysis was performed using these parameters and revealed significant associations of daylength with cases/week but not temperature (Table 4). However, this model showed some evidence of non-constant variance (Figure 7), indicating there could be structures to the data that are not accommodated by the parameters of this linear regression model. The residuals did appear to be approximately normally distributed as indicated by comparison of their quantiles against those of the normal distribution (Figure 6)

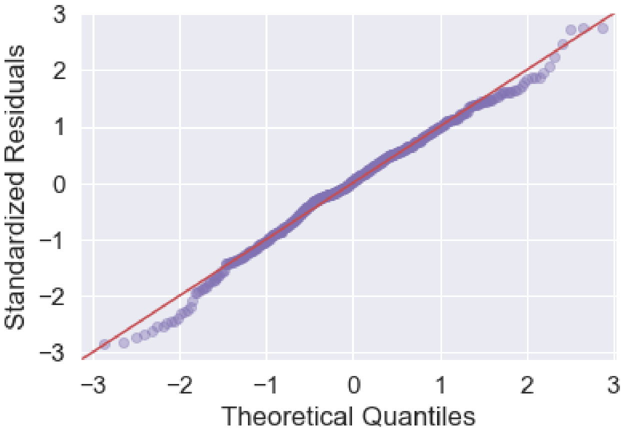


Figure 6 QQ-plot comparing distribution of residuals against normal distribution

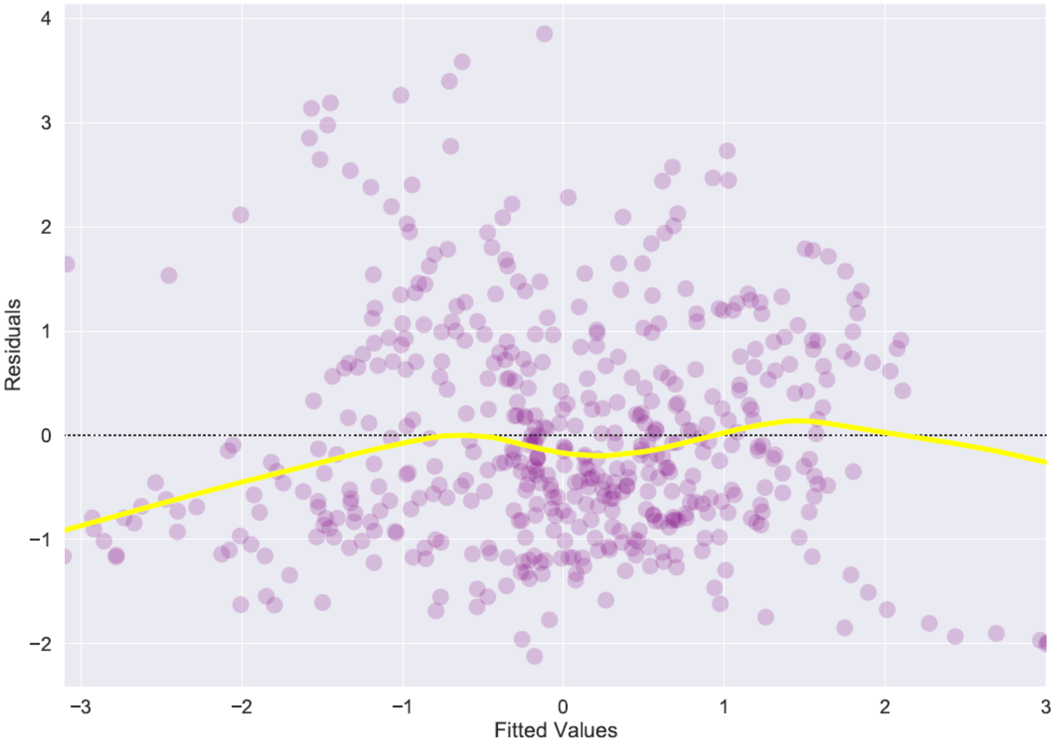


Figure 7 Residuals plotted against fitted values for Model 1, that did not account for time or country effects. There appears to be some structure to the distribution of the residuals around zero, although this is not severe, still indicates the model could be improved

The time and country factors in these data are likely to account for the missing variance, although themselves of little relevance to the study hypothesis. Data from each week of the study, and from each country are not likely to be independent of each other, and non-constant variance of the residuals reflects the fact is not considered in the linear model presented here.

The Durbin Watson statistic is a measure of autocorrelation in residuals from a regression, with values less than 2 indicative of positive autocorrelation, and in this case the Durbin Watson statistic was 0.326 (Table 4). This suggests serial correlation between data points, such as might be caused by similarity of successive data points across the 42 week time series in this study.

3.4 Mixed Model Regression Analysis

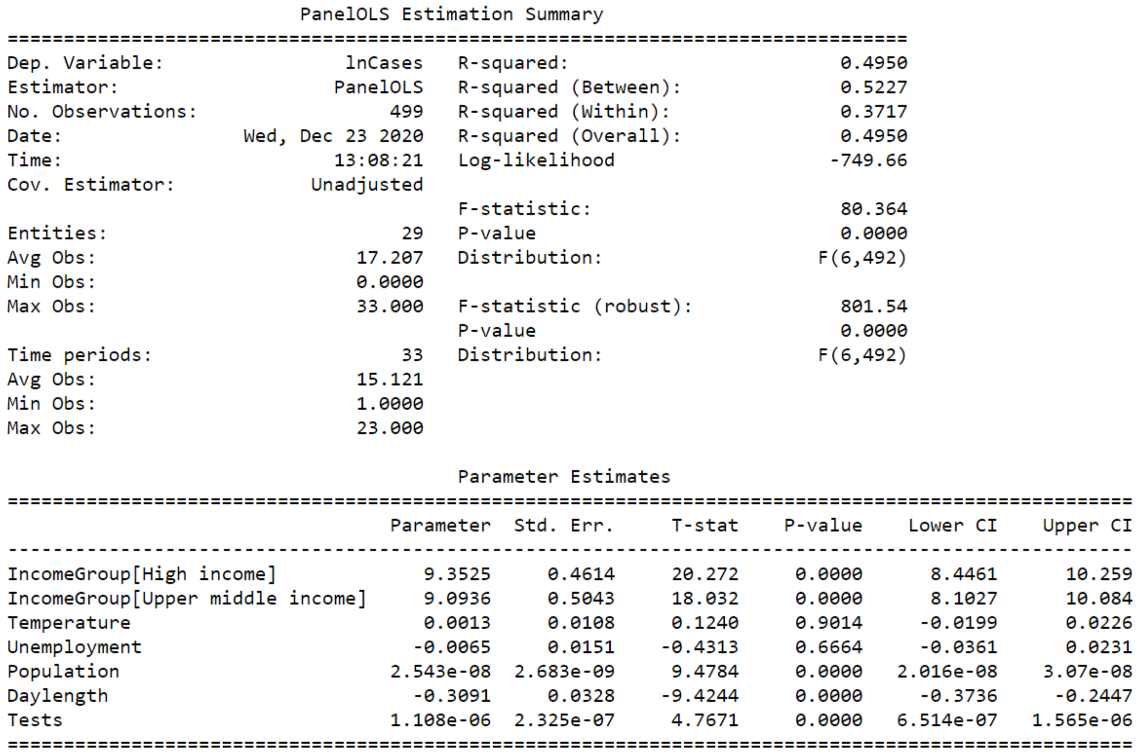
Models that include time and country as random effects might accommodate the hierarchical structure that these parameters bring to the model.

Model 2 was defined as:

$$y_{it}= \alpha_i + \beta x_{ijt} + \gamma_t + \epsilon_{it}$$

where *j* indexes Country and *t* indexes time, *β* is the slope, α_i is an entity (Country) -specific error and ϵ_{it} is the overall error or residual, + γ_t is the random characteristic of Country

Table 4 Regression analysis of the relationship between cases/week and the predictor variables in this study



```
                          PanelOLS Estimation Summary
================================================================================
Dep. Variable:                 lnCases   R-squared:                        0.4950
Estimator:                    PanelOLS   R-squared (Between):              0.5227
No. Observations:                  499   R-squared (Within):               0.3717
Date:                 Wed, Dec 23 2020   R-squared (Overall):              0.4950
Time:                         13:08:21   Log-likelihood                   -749.66
Cov. Estimator:             Unadjusted
                                         F-statistic:                      80.364
Entities:                           29   P-value                           0.0000
Avg Obs:                        17.207   Distribution:                   F(6,492)
Min Obs:                        0.0000
Max Obs:                        33.000   F-statistic (robust):             801.54
                                         P-value                           0.0000
Time periods:                       33   Distribution:                   F(6,492)
Avg Obs:                        15.121
Min Obs:                        1.0000
Max Obs:                        23.000

                                        Parameter Estimates
====================================================================================================
                                  Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
----------------------------------------------------------------------------------------------------
IncomeGroup[High income]             9.3525     0.4614     20.272     0.0000      8.4461      10.259
IncomeGroup[Upper middle income]     9.0936     0.5043     18.032     0.0000      8.1027      10.084
Temperature                          0.0013     0.0108     0.1240     0.9014     -0.0199      0.0226
Unemployment                        -0.0065     0.0151    -0.4313     0.6664     -0.0361      0.0231
Population                        2.543e-08  2.683e-09     9.4784     0.0000   2.016e-08    3.07e-08
Daylength                           -0.3091     0.0328    -9.4244     0.0000     -0.3736     -0.2447
Tests                             1.108e-06  2.325e-07     4.7671     0.0000   6.514e-07   1.565e-06
====================================================================================================
```

Accounting for the effects of country and time as random factors, did not change the interpretation of the relationship between the predictors and cases/week. Daylength remained negatively associated with cases/week, while temperature was not significant. Income positively predicted cases, with higher income associated with more cases/week. Similarly, higher values for population were associated with more tests per week.

4. Discussion

This study has provided some preliminary evidence that cases of COVID19 per week might be associated with daylength, independent of some other factors that might confound this relationship including testing rates, economic status and population numbers. This is preliminary evidence of seasonality of COVID19 infection, which has important implications for preventing and controlling future outbreaks of this disease.

Although daylength was associated with COVID19 cases, it was also strongly correlated with outdoor temperature, which precluded interpretation of the relative importance of these factors. Collinearity between daylength and temperature is one of the greatest obstacles to understanding the environmental factors that drive seasonality of infectious disease in humans. In animals, it is the number of hours of light per day (daylength) and not temperature, that drives seasonality of immune function through established neurobiological mechanisms (Stevenson, 2005). Collinearity between temperature and daylength make it difficult to study the importance of these mechanisms in humans at a population level. This problem has been addressed in humans using controlled-laboratory studies over several decades, all of which have failed to prove any effect of exposure to cold temperatures on vulnerability to viral infection (Eccles, 2015). Controlled studies of the biological effects of daylength on immunity would require months or years living in controlled conditions and are not feasible in humans. This biological question can only be addressed by analysis of real-world data collected over long time periods and at wide ranges of temperature and daylength that have the power to separate the effects of these two variables.

We elected to study the number of cases of COVID19 rather than positive tests or deaths, as this might be a better indicator of the spread of the disease rather than the demography of infection, which would be better represented by deaths. Nevertheless, the number of cases depends on both the number of tests, and the number of susceptible individuals that are exposed (population). Although we included these factors as covariables in the models, our findings might still be biased by differences in access to testing and healthcare.

The results of this study have provided some evidence that COVID19 prevalence might be associated with daylength, but it was not possible to conclude that this was independent of temperature. Daylength was significantly associated with COVID19 cases, and temperature was not, but the reliability of the estimates derived from the model is compromised by the high collinearity between these two parameters. While it was not possible to demonstrate an independent effect of either temperature or daylength, the overall predictive power of the model is not affected by collinearity (O'Brien 2007), and it is likely that the combined effects of temperature and daylength are associated with COVID19 cases for the data on COVID19 prevalence in Europe in this study.

There are many limitations to this study that could implicate the validity and the general applicability of our conclusions. The multicollinearity of the parameters in the model compromises the reliability of the estimates of the effects of the predicator variables, although it does not affect the overall predictive power of the model, nor our overall conclusion of seasonality. The range of latitude, and consequently daylength was small, with no representation from tropical and equatorial countries. This is likely to decrease the chances of detecting a relationship between daylength and COVID19 cases, if one exists, by limiting the range of daylength in the sample. The numbers of COVID19 cases are modulated through a wide range of social, economic, health and demographic factors that were not considered in this study due to the lack of information. In particular, the procedures governing access to testing might drive the rates of COVID19 cases, but there was no way to incorporate this into the model, and it was assumed that the number of cases per week reflected the prevalence in each country. Data were available for less than one year, which does not represent the entire annual cycle of daylength. Temperature and daylength were taken as the mean values for each country which is a crude measure of the actual environmental conditions, particularly in countries that span a large latitudinal range.

4.1 Conclusion

There was some evidence that the number of COVID19 cases per week was higher in weeks with shorter days in Europe during the 2020 pandemic, but it was not possible to determine if this was independently associated with daylength or temperature. In conclusion, these data support increased prevalence of COVID19 in short daylengths and suggest that public health measures to prevent transmission of SARS-CoV-2 are particularly important in wintertime. Further research to identify the factors that mediate seasonality of susceptibility to viral infection will further understanding of immune function and guide public health measures to prevent transmission of infectious disease.

5. References

Bloom-Feshbach, K. et al. Latitudinal Variations in Seasonal Activity of Influenza and Respiratory Syncytial Virus (RSV): A Global Comparative Review. PLoS One 8, 3–4 (2013).

Dowell, S. F. Seasonal variation in host susceptibility and cycles of certain infectious diseases. Emerg. Infect. Dis. (2001) doi:10.3201/eid0703.017301.

Dowell, S. F. & Shang Ho, M. Seasonality of infectious diseases and severe acute respiratory syndrome - What we don’t know can hurt us. *Lancet Infect. Dis.* **4**, 704–708 (2004).

Eccles, J E Wilkinson Exposure to cold and acute upper respiratory tract infection. Rhinology 2015 Jun;53(2):99-106. doi: 10.4193/Rhin14.239.

Lee, W. M. et al. Human rhinovirus species and season of infection determine illness severity. Am. J. Respir. Crit. Care Med. 186, 886–891 (2012).

Li Y, Xin Wang, Harish Nair. Global Seasonality of Human Seasonal Coronaviruses: A Clue for Postpandemic Circulating Season of Severe Acute Respiratory Syndrome Coronavirus 2? J Infect Dis 2020 Sep 1;222(7):1090-1097. doi: 10.1093/infdis/jiaa436.

Lofgren, E., Fefferman, N. H., Naumov, Y. N., Gorski, J. & Naumova, E. N. Influenza Seasonality: Underlying Causes and Modeling Theories. J. Virol. 81, 5429–5436 (2007).

O'Brien R. A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity (2007) 41:673–690

Stevenson, T. J. & Prendergast, B. J. Photoperiodic time measurement and seasonal immunological plasticity. *Frontiers in Neuroendocrinology* (2015) doi:10.1016/j.yfrne.2014.10.002.

Tamerius, J. et al. Global influenza seasonality: Reconciling patterns across temperate and tropical regions. Environmental Health Perspectives (2011) doi:10.1289/ehp.1002383.