An Phríomh-Oifig Staidrimh Central Statistics Office

John Hooper
Medal for Statistics

Demographic Predictors of the Incidence of COVID-19

Introduction

During the COVID-19 pandemic, claims about certain areas or parts of society being particularly likely to contract the novel coronavirus have abounded in the media. From claims that the virus is particularly prevalent in border counties to questions about whether young people or tourists are to blame for its spread, there has been no shortage of speculation about what might drive the differential rates of COVID-19 around the country.

We wanted to examine whether any of these popular claims were true. To make the task manageable, we took the total number of cases per county from March 2020 to January 2021 as the dependent variable. We then used linear regression and data visualisation to examine the following questions:

1. Do the border counties have more cases?
2. Do counties with younger populations have more cases?
3. Do counties with higher population density have more cases?
4. Do counties with more domestic tourists have more cases?

Statistical Procedure

Data were compiled from official sources and used to analyse the chosen questions. Rates of incidence of COVID-19, the dependent variable, were obtained from Ireland's COVID-19 data hub. Demographic information, used as the independent variable, was obtained from the online database of the Central Statistics Office (CSO).

The following data were used to analyse each of the claims:

1. Donegal, Monaghan, Cavan, Louth and Leitrim were classified as border counties. Counties were then encoded with a dummy variable which was '1' for a border county and '0' otherwise.
2. The average age for each county in Ireland was obtained from the 2016 census. Despite the five-year gap between then and now, these were accepted as a reasonable approximation.
3. The population for each county was obtained from the 2016 census. The area of each county was obtained from the GeoHive system. This was used to calculate the population density. The logarithm of density was used in analysis because of the wide variation in the data (in which Dublin was an outlier).
4. The number of trips made to each county in Ireland per year was downloaded from the CSO database. The average between 2010 and 2019 was calculated.

Once the data were obtained, ordinary least-squares linear regression was used to regress cases of COVID-19 per 100,00 onto the variable. The adjusted R^2 was used to measure the variance explained by each model.

It should be noted that linear regression was used, even though the data were found to be heteroscedastic. This was justified as an approximation, but means that all results should be treated cautiously.

Regression Analysis

All the regression analysis was performed in R. The method used was ordinary least-squares regression. Significance was assessed using the Student's t-statistic of the coefficient of regression, in a two-tailed test. The critical value chosen was 0.05. The significance codes used below are: <0.05*; < 0.01**; < <0.001***.

Border Counties

	Estimate	Standard Error	t-value	p-value
Intercept	3182.8	188.8	16.86	8.28e-15***
Border County	1416.3	430.4	-3.29	0.00308**

As can be seen in the table above, border counties had a significantly higher relative incidence of COVID-19. As a result, this variable was included as a covariate in the subsequent analysis of age, tourism and density. The adjusted R^2 value was 0.2822.

Average Age

	Estimate	Standard Error	t-value	p-value
Intercept	11304.2	5315.0	2.127	0.0439*
Average Age	-208.1	140.8	-1.478	0.1525

When border counties were included as a covariate, the following results were obtained. The adjusted R^2 value was 0.3596.

	Estimate	Standard Error	t-value	p-value
Intercept	11778.3	4354.9	2.705	0.0126*
Average Age	-228.1	115.4	-1.975	0.0603
Border County	1455.2	407.0	3.575	0.0016**

Level of Tourism

	Estimate	Standard Error	t-value	p-value
Intercept	3009.3	275.6	10.920	1.42e-10***
Annual Trips	0.4361	0.5025	0.868	0.39443
Border County	1526.3	450.8	3.385	0.00255**

Population Density

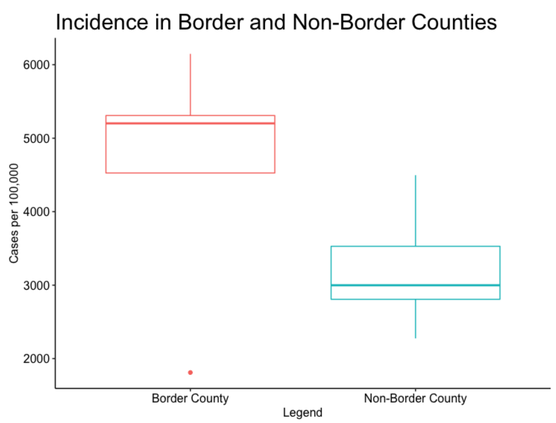
	Estimate	Standard Error	t-value	p-value
Intercept	1586.3	955.0	1.661	0.1097
Logarithm of Density	466.9	233.9	1.997	0.0573

When border counties were included as a covariate, the following results were obtained. The adjusted R^2 value was 0.4652.

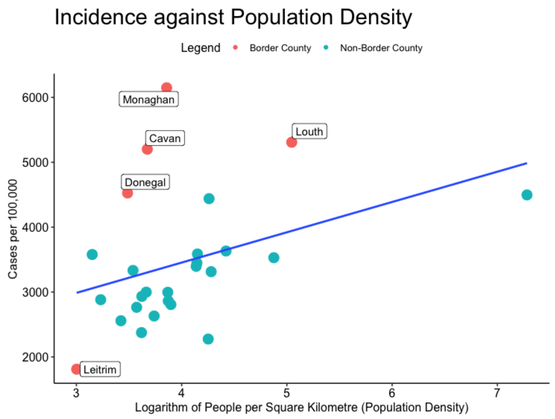
	Estimate	Standard Error	t-value	p-value
Intercept	944.7	755.0	1.251	0.223438
Logarithm of Density	552.9	182.1	3.036	0.005873**
Border County	1546.0	374.0	4.134	0.0004***

Graphs

Border Counties



Population Density



When all the independent variables with some promise of an effect were combined into one linear model, the following results were obtained. The adjusted R^2 value was 0.4491.

	Estimate	Standard Error	t-value	p-value
Intercept	3993.49	5395.43	0.740	0.467029
Logarithm of Density	482.88	221.89	2.176	0.040553*
Border County	1542.10	379.62	4.062	0.000518***
Average Age	-73.37	128.52	-0.571	0.573880

Discussion and Evaluation

These results show that the border counties experienced excess COVID-19 cases inexplicable by mere chance. This is especially true of the Ulster counties, whereas Leitrim and Louth are not so high. In addition, a county's having a higher population density appears to increase its incidence of COVID-19. The relationship is highly significant once counties' being on the border is controlled for as a covariate; in that model, an increase by a factor of 10 in population density increases the incidence per 100,00 by about 553 cases.

There seems to be little evidence to support the claim that those areas which receive more tourists have a higher incidence. However, the analysis employed here may miss out on important effects: it only takes an aggregate of all COVID-19, rather than looking at how they vary over time; and only average data from 2010 to 2019 was used to estimate an area's popularity as a tourist destination, which does not take into account the differences due to the pandemic itself in 2020.

Equally, there may be some effect due to age since, though it is not significant, the sample size is small enough that an important effect could easily be missed. However, the fact that the average age is very far from significance in the linear model which includes population density implies that age is probably not an important influence; this is further indicated by the decrease in the adjusted R^2.

All results should be treated with caution, as the assumptions of linear regression were not upheld. If this project were to be done again, more time would be put into finding a better regression model. However, since the most natural alternatives, Poisson and quasi-Poisson regression, were too over-dispersed to be meaningful, linear regression was the best that could be obtained within the temporal constraints.

Conclusions

To answer the four questions posed by the project:

1. The border counties clearly have a disproportionate number of cases; this is even more apparent if only the Ulster counties are included.
2. There is some evidence that counties with younger populations are more likely to spread COVID-19, but the data are not significant. Therefore, the idea that young people are primarily responsible for the spread of the virus may be overblown.
3. High population density seems predictive of high COVID-19 incidence, and is significantly so when being on the border is included as a covariate.
4. There is little evidence that counties which are tourist destinations have higher levels of COVID-19, though a time series analysis would be a more precise way to analyse this and could show a relationship.

Bibliography

1. Central Statistics Office (CSO) Database: datasets EP001, E8084, HTA11 (https://data.cso.ie/);
2. Ireland's COVID-19 Data Hub (https://covid19ireland-geohive.hub.arcgis.com/);
3. R packages: tidyverse, matrixStats, lubridate, stringr, ggthemes, ggpubr, haven, ggrepel;
4. 'Statistics: An Introductory Analysis' (Third Edition) by Taro Yamane.