Sales Analysis

Maeva_Braeckevelt

30/12/2020

Executive summary

This analysis aimed at analyzing the pattern of association between registered COVID-19 cases and registered death due to COVID-19 until the 26 October 2020 among countries. The data used was gathered both by the Center for Systems Science and Engineering and the World Bank'sites. The main variables that I used were: total of registered number (per thousand people), total of death (per thousand people) and the population (per millions). The regression model chosen was the Log-Log Weighted Linear regression. It showed a linear pattern between death and confirmed case, such as, for people among all countries, for +10% change in registered COVID-19 case, there is an association of +9,5% change in registered Covid-19 deaths. However, this analysis is subject to the the politics testing of every country (mass testing, few tests, etc.)

Introduction

What factors could impact the sales of a restaurant during the Covid-19 pandemic? That's the question I want to answer. Restaurants are living a rough period : in Belgium, from the 20 october 2020, new sanitarian measures, among others,appeared : the restaurant can not welcome sitting clients anymore (only take-away meals) and there is a curfew at 10pm. The restaurants of my company are struggling to be profitable and the managers need as much information as possible to be able to take the right decision in a unstable environment. Maybe knowing how the productive hours or the day of the week (and other variables) are correlated to the sales, in this particular period, could be insightful. I have taken the data of one of our restaurants and I will use it for my analysis. There is not such analysis done under "normal" cirucumstances yet but it would be interesting to compare those in the future.Due to recentness of the situation I only have 42 observations, from the 20 october to the 30 november 2020. Thus, the unlikely circumstances and the lack of observations may be problematic during the analysis and for external validity. Unfortunately, the actual circumstance will remain so the analysis is worth being done even with few observations. Moreover, once the model is set, adding observations and see the evolution will be interesting as well.

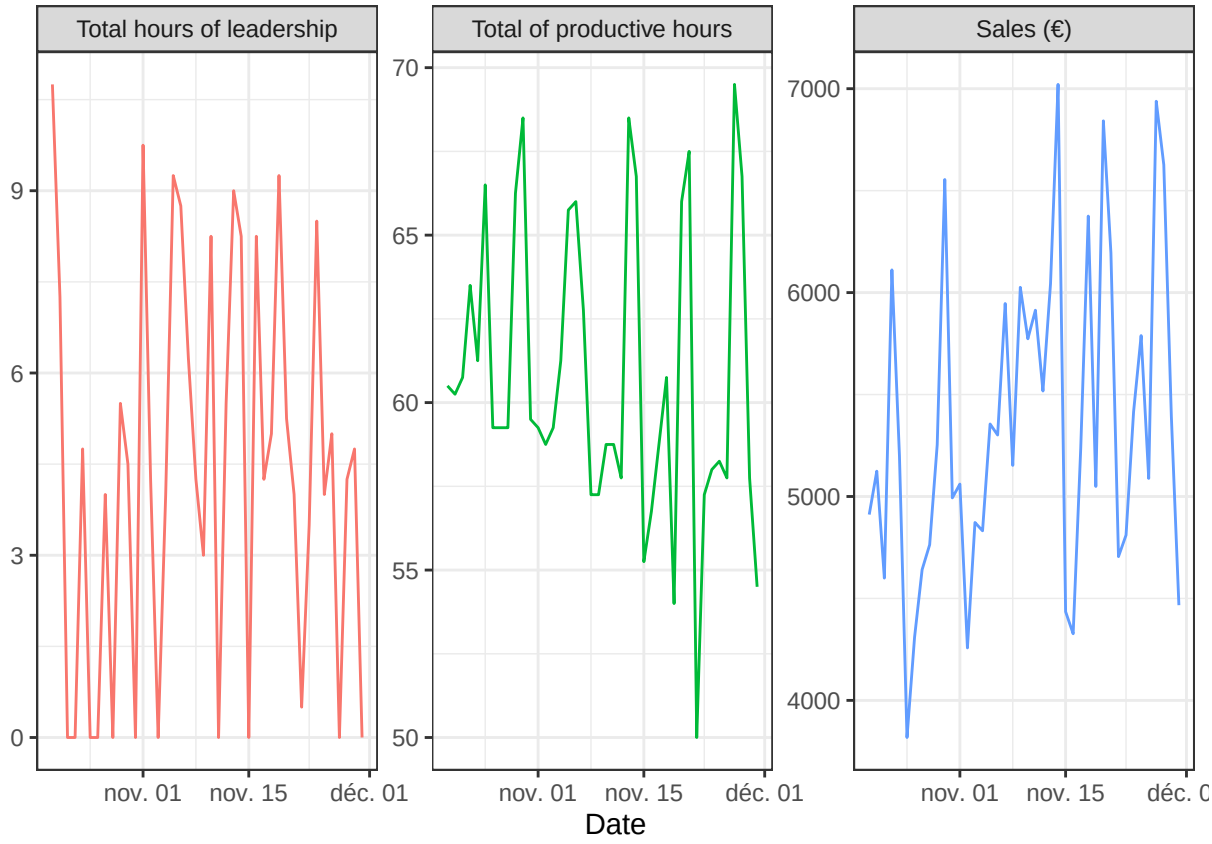
Data

Working hours and Sales

Staffing a restaurant according to the demand is a difficult decision. Providing statistical directives could be an advantage to the manager. My first analysis aim to analyze the correlation between working hours and sales. My goal is to observe how increasing or decreasing the productive hours could impact the sales and secondly if the presence of the manager has and how an impact on the sales. I used the Sales in EURO (HT) per day (y), the productive hours (hours of kitchen's employee and waiters) as a first variable and the leadership hours (hours of the managers working either in the kitchen or as a waiter) as the second variable.

Table 1: Summary statistics

variable	n	mean	median	min	max	sd	skew
Sales	42	5358.266	5217.607	3817.94	7021.001	789.843909	0.4048451
Productive_hours	42	60.750	59.250	50.00	69.500	4.479969	0.2248577
Leadership_hours	42	4.375	4.250	0.00	10.750	3.366464	0.0809803



the day of the week the hours of downtime of a delivery platform : When employees feel swamped by on site client, they can pause the delivery platform The sales are in EURO (HT) per day. The variable that I will use are the productive hours : hours of kitchen's employee and waiters. the leadership hours : hours of the managers working either in the kitchen or as a waiter the day of the week the hours of downtime of a delivery platform : When employees feel swamped by on site client, they can pause the delivery platform

Histogram and summary statitics

The summary statistic does not take to account the size of the population of the countries, so the interpretation is less meaningful. There is very small and very big population. But I observed some similarities between the two variables : they both are skewed with a right tail and some extreme values. The median of both are significantly smaller than the mean. The majority of the countries has low value. Although those variable share the same tendency, the scale is very different between them. I can already sense that I will have to find a way to weight the population to be able to have a relevant analysis.

Transformation of the variables

To uncover the trend of the pattern association, I investigated four non-parametric regressions : level-level (figure A1 in the appendix), log-level (A2), level-log (A3), log-log(A4). I chose to use the fourth model (A4) : use Log transformation for both of my variables.

Substantive reasons : the COVID-19 is a very contagious disease, one person can infect multiple people, so the variable are affected in multiplicative ways. The graph itself fits better for the interpretation. Also, we are looking for percentage association.

Statistical reasons: The distributions of the variables are skewed with a long right tail. So taking the log is good solution to make the distribution of my transformed variable more symmetric. In addition,my variables have the same metric so for the purpose of comparison, taking log for both is easier. The graph (A4) can be interpret in a meaningful way and give a good prediction.

Model choice and interpretation

The regression model that I have chosen is the Log-Log Weighted Linear regression : weight by coutries' population. It sounded pertinent that bigger population has a bigger impact on the slope. Please find in the appendix the estimation of the different models and the argumentation.

Formula : ln_death = -3,39 + 0,95 * ln_confirmed, weights: countries'population

Alpha : -3,39 is the average of ln_death when the confirmed case is one (ln(1)=0). **Beta** : the deaths is 9,5 percent higher on average for observation having 10 percent higher case confirmed In log-log transformation, alpha is usually not meaningfull. The deaths are not increasing as fast as the confirmed cases. The graph (A8) shows that bigger the population is bigger are the cases of deaths. It will be very interesting to transform the variable per capita. This way, I could take out the bias having more deaths by higher population.

Hypothesis testing on Beta

H0 : there is no pattern association between deaths and confirmed cases.

HA : there is a pattern of association between deaths and confirmed cases.

I chose 95% confidence interval for the hypothesis testing. My confidence interval is not crossing 0, so it means that it is significant. My p value is $< 0{,}05$ so I can reject my HO and therefore there is a pattern of association between deaths and confirmed cases on the 26/10/2020.

$$H_A : \beta \neq 0$$

Analysis of the residuals

For this 5 countries, the model overestimated the deaths. I can see that the predicted value is smaller than the real value. For example, for Liechtenstein , the ln_deaths is equal -6,91, but the predicted value for is -4.35 so the predicted value is off by -2,56. So those countries have less death than the average. There could be a lot of explanation for it : better healthcare, less sensitif to covid19 (young people), isolation, etc.

For this 5 countries, the model underestimated the deaths. I can see that the predicted value is bigger than the real value. For Italy, the ln_deaths is equal to 3,62, the predicted value is 2,55 so the predicted value is off by +1,06. So, Italy has more deaths than the average. It could be due to the way they handle the pandemic, or the bad healthcare, or old population, etc

Appendix

This appendix contains the documentation of the analysis annoted.

Investigation of the transformation of the variable

Estimating different models

I estimated fourth model : Simple linear regression (A5), quadric linear regression (A6), Piecewise linear spline regression (A7) and Weighted linear regression : weight with population (A8).

Simple Linear regression (A5)

This graph represents the simple linear regression between the confirmed cases and the death due to COVID-19 until the 26/10/2020.

The formula is ln_death = -4,12 + 1,03 * ln_confirmed

Alpha : -4,01 is the average of ln_death when the confirmed case is one (ln(1)=0).

Beta : the deaths is 10,3 percent higher on average for observation having 10 percent higher case confirmed

In log-log transformation, alpha is usually not meaningful. The deaths are increasing faster than the confirmed cases.

The adjusted R squared is 0,89. That's a trustful model.

Quadric Linear regression (A6)

This graph represents the quadric linear regression between the confirmed cases and the death due to COVID-19 until the 26/10/2020.

The formula is ln_death = -4,01 + 0,88 * ln_confirmed + 0,02 * ln_confirmed^2

Quadric Linear Regression is very hard to interpret The deaths are not increasing as fast as the confirmed cases.

The adjusted R squared is 0,89. That's a trustful model

Piecewise linear spline regression (A7)

This graph represents Piecewise linear spline regression between the confirmed cases and the death due to COVID-19 until the 26/10/2020. I chose two cut-off, the first at one, the second at 200.

The formula is ln_deaths = -4,06 + 0,91 * ln_confirmed * 1(confirmed < 1) + 0,99* l n_confirmed * 1(1 <= confirmed < 200) + 1,28 * ln_confirmed* 1(confirmed >= 200)

On the graph A7, thanks to the cut off, I can observed that the line is becoming steeper. The deaths are not increasing as fast as the confirmed cases. The adjusted R squared is 0,89. That's a trustful model

Weighted Linear regression : weight with population

This graph represents the Weighted Linear regression : weight with population between the confirmed cases and the death due to COVID-19 until the 26/10/2020.

The formula is ln_death = -3,39 + 0,95 * ln_confirmed

Alpha : -3,39 is the average of ln_death when the confirmed case is one (ln(1)=0).

Beta : the deaths is 9,5 percent higher on average for observation having 10 percent higher case confirmed

In log-log transformation, alpha is usually not meaningful. The deaths are not increasing as fast as the confirmed cases. I can observe than the countries with the more deaths and confirmed cases are the country with a large population. The adjusted R squared is 0,93. That's a trustful model.

Model Chosen Weighted Linear regression : weight with population

I chose the Weighted Linear regression : weight with population.

Statistical reasons: The adjusted R square of the model is the highest, so it's the model that fits the best. I excluded the PLS and the quadric function, for model complexity reasons.

Sustantive reasons : I can see a 0.8 difference between the slope of the linear regression and the slope of the weighted regression, it means that not taking into account the different size of population increase the slope and the predicted deaths. So, I think the pertinent choice is the weight regression.

Graphs and Model summary statistics