The Relationship between COVID-19 Incident Rate and Weather in OECD Countries.

Eun Lim - 1005280839

April 26, 2021

Abstract

The COVID-19 pandemic has caused numbers of death and incurred huge social costs. The COVID-19 belongs to the coronavirus subfamily and therefore shares similar symptoms to the common cold and flu which are also in the coronavirus subfamily. Since the common cold and flu are much more infectious in the winter, we suspect that the relationship between the infectivity of the COVID-19 and the weather. The infectivity is measured by the incident rate and the temperature is considered as a main factor as the weather. The weather effect on the incident rate is analyzed by fitting a linear model and it is found that the incident rate decreases as the temperature increases.

Introduction

The world has been suffered from the COVID-19 pandemic since 2020. The World Health Organization (WHO) on March 11, 2020, declared the novel coronavirus (COVID-19) outbreak a global pandemic. By the end of March, most of the world became seriously affected by the disease. The virus that causes the COVID-19 is SARS-CoV-2 and it is one of the viruses in the coronavirus subfamily (Corman et al., 2018, p. 175). Such a group of viruses causes respiratory tract infections that can range from mild to lethal, and two of the most popular diseases that are caused by coronavirus are the common cold and flu (Similarities and Differences between Flu and COVID-19., 2021). It is well-known that the common cold infections and flu infections occur more frequently in the winter because the virus tends to get stronger and the human immune system tends to get weaker due to the low temperature.

Accordingly, in this paper, we examine the weather effect on the incident rate of COVID-19 over 37 OECD membership countries under the hypothesis that the incident rate of COVID-19 is affected by the temperature, and so there is a relationship between the incident rate and the weather. Consequently, the weather effect, in this paper, is defined by the effect of temperature. Especially, we define the temperature as the air temperature recorded by weather stations on land and sea as well as some satellite. Furthermore, we define the incident rate of COVID-19 as a ratio of the number of confirmed cases to the population that is converted into the percentage scale.

The weather effect on the incident rate of COVID-19 over 37 OECD membership countries is important and meaningful to analyze for three major reasons. Firstly, among various factors that affect the incident rate, the weather is almost the only natural factor that one has no control over. However, at the same time, the weather, namely temperature, exhibits substantial spatial and interannual variability. Thus, when all the other factors are excluded from analysis or assumed to be held constant, the weather effect determines the infectivity of the COVID-19 and behaves as a baseline indicator that does not depend on social, economic, and political factors. The second reason highlights the context of why the incident rate of COVID-19 is used as some type of indicator in this paper. The COVID-10 pandemic is risky and considered the worst among all other pandemics in history, such as H1N1 Swine Flu and Zika Virus, because of the infectivity and the fatality. The fatality, however, is somewhat under control if a government has a certain level of the public health care system and the ability to cope with risks. For example, as long as a country's government can

give appropriate treatments to a patient, it is not as lethal unless one has had an underlying disease or one is old. Nevertheless, the infection itself is harder to control even with some preventive measures. Consequently, to examine the characteristics of the virus, analyzing the incident rates is more appropriate than analyzing any other indicator. Lastly, although the COVID-19 is a global shock, as previously explained, governments may have different level of the public health care system and perhaps have implemented different preventive measures. The membership countries in OECD meet a certain social, economic and political standards, i.e., not too poor, underdeveloped or in a war, but still captures some variability. What this means is that the gap between the countries in OECD is not like the gap between the wealthiest country and the poorest country in the world, but the gap is like the gap between a country in G7 and a country in G20. Therefore, the weather effect can be more accurately analyzed and remove some unobserved bias and confounding variables.

Data

To conduct the analysis, we gathered 3 different datasets, COVID-19 cases per country, average monthly temperatures per country, and demographics per country. The COVID data is from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (COVID-19 Portal, 2021). Also, the average monthly temperature data is from the World Bank Climate Change Knowledge Portal (World Bank Climate Change Knowledge Portal, n.d.), and the global demographic data is from the OECD Statistics (Demography - Population - OECD Data, n.d.).

The original COVID-19 data contains information including but not limited to the number of cumulative confirmed cases and deaths on a daily basis, and geographical information from the day that the virus is first identified up to April 25th, 2021.

The demographics data contains the population of each of the 37 OECD countries and used to compute the incident rate of the COVID-19 per country. This data is based on the report in 2018.

The weather data contains the average monthly temperatures of each country from 1991 to 2016. It would be much preferred if there were data of average monthly temperatures from April 2020 to March 2021 for each of the countries, but it was impossible to find that from a credible source, and so our weather data is chosen instead.

Each dataset is cleaned, following the two main criteria; dropping unnecessary variables and converting the unit and time.

As a result, from the COVID-19 dataset, we dropped all other variables except the variables denoting the number of confirmed cases of the 37 OECD countries and date. Also, since the dataset is constructed on daily basis, we only took the data from the last day of each month from March 2020 to March 2021, then computed the difference between the number of confirmed cases of one month and the number of confirmed cases of the month right before. As a result, we obtained the number of confirmed cases reported every month from April 2020 to March 2021.

Likewise, the original temperature data contains the average monthly temperatures from 1991 to 2016, i.e., 12×24 monthly values. Therefore, we calculated the 24 years average of each month's temperature.

Additionally, the population from the demographics data is extracted and is used to divide the number of confirmed cases from the COVID-19 data to get the incident rate. For simplicity, the rates are scaled to percentage.

Then, we merged each dataset to create a big dataset that contains all the information needed for the paper. The merging process is done by merging the COVID-19 data with the weather datasets indexing by country names and months.

The merged dataset, therefore, contains four variables, country, date, incident rate, and temperature. Among the variables, the most important ones are incident rates and average monthly temperatures. At this point, the countries and dates are unnecessary.

The minimum, mean and maximum incident rates from April 2020 to March 2021 over 37 OECD countries

are 0.01%, 1.95%, and 14.42%. Likewise, the minimum, mean and maximum temperature from the same period is $-23.01°c$, $9.94°c$.

```
temp_plot <- ggplot(data = covid, aes(x=c(temperature), y=incident_rate))+
  geom_point(color = 'skyblue', size = 1) +
  xlab("temperature (degrees celsius)")+
  ylab("incident rate (%)")+
  ggtitle("Fig 1. Incident Rate vs. Temperature")

temp_his <- ggplot(data=covid, aes(x=temperature)) +
  geom_histogram(fill="orange",colour="white") +
  xlab("temperature (degrees celsius)") +
  ggtitle("Fig 3. Temperature")

rate_his <- ggplot(data=covid,aes(x=incident_rate)) +
  geom_histogram(fill="light green", colour = "white") +
  xlab("incident rate (%)") +
  ggtitle("Fig 2. Incident Rate")

temp_plot|rate_his|temp_his
```

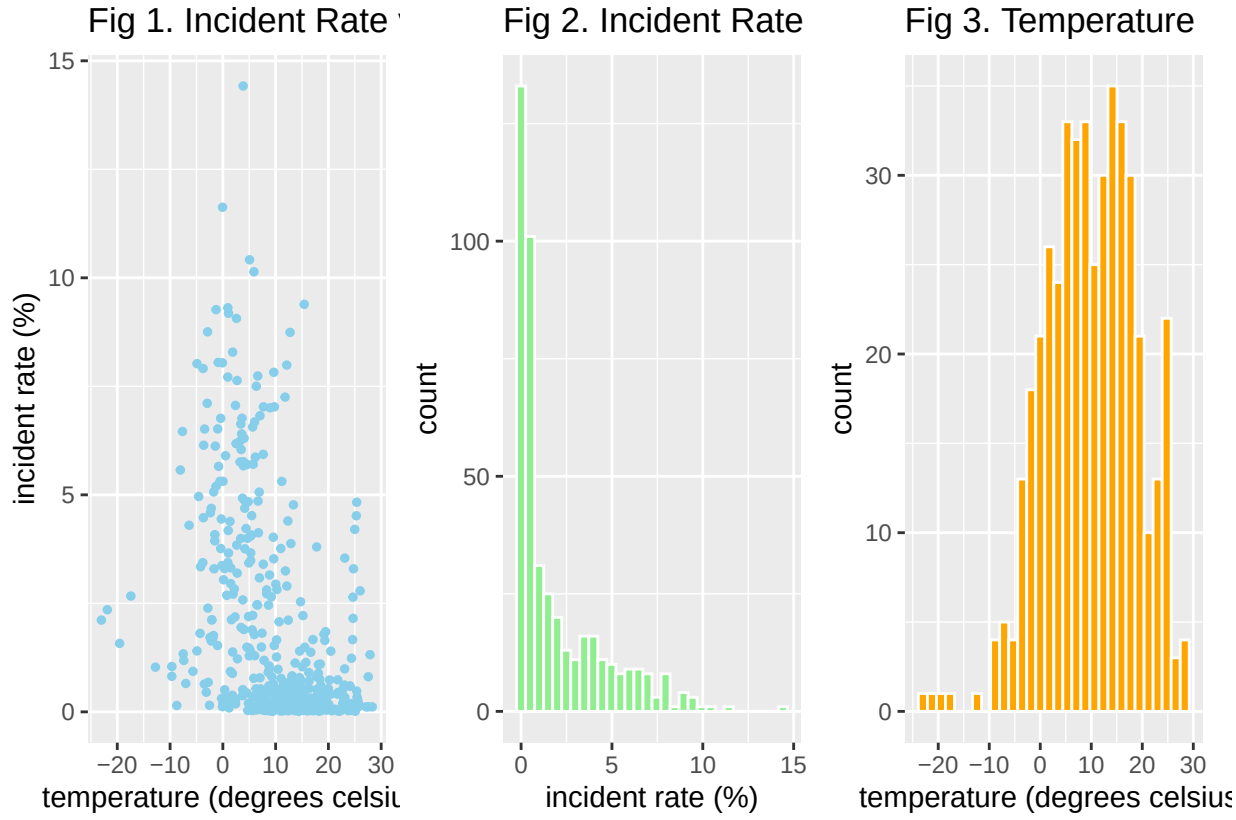


Figure 1. shows the relationship between the temperature and the incident rate. The scatter plot slightly shows some negative and linear relationships. But, at a relatively lower temperature, the points are more sparsely scattered while at a relatively higher temperature, the points are more concentrated. It can later be a problem if we were to fit a linear model, because the residuals can show some patterns, instead of randomly scattered.

Figure 2. illustrates the distribution of the monthly incident rate over 37 OECD countries. It is obvious that the data is not normally distributed with a right-skewed distribution. It almost looks like a typical log-normal distribution which is also a problem if we were to fit a linear model.

The last one, figure 3., shows the distribution of the average monthly temperature of the 37 OECD countries. It does not look like a typical bell shape of a normal distribution, because it is slightly left-skewed. The skewness is not as extreme as the skewness found in figure 2., but it is still hard to conclude that it is perfected normally distributed.

All analysis for this report was programmed using `R version 4.0.4`.

Methods

We will find the weather effect on the incident rate by regression analysis method. The linear model has a slope coefficient which describes the linear relationship. The slop coefficient is estimated by the OLS and the MLE approach. Also, the significance of the slope coefficient is tested by the hypothesis test. To determine that the model does not violate the linear model assumptions, we will conduct the goodness of fit test.

Linear Regression

Let Y denote the monthly incident rate and let X denote the average monthly temperature. Equivalently, we define the average monthly temperature as the independent variable and the incident rate as the dependent variable.

Suppose that $Y_1, Y_2, \ldots, Y_{444}$ are independent realizations of the random variable Y that are observed at the values $x_1, x_2, \ldots, x_444$ of a random variable X. Assume that the regression of Y on X is linear, so for $i = 1, 2, \ldots, 444$ $Y_i = E(Y|X = x) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i is the random error in Y_i and $E(e|X) = 0$.

The equation of the least-squares line of best fit is $y = \hat{\beta}_0 + \hat{\beta}_1 x$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the slope and intercept estimator of the regression line. The slope coefficient in the model is significantly important and contains a piece of meaningful information. The slope coefficient defines the linear relationship between the dependent and independent variables. If the slope coefficient is negative, then it implies that there exists a negative relationship. If the slope coefficient is positive, it implies that there exists a positive relationship. If the slop coefficient is zero, then it implies there is no relationship. The intercept coefficient is rather unimportant, especially in this paper, because it denotes the incident rate when the temperature is at 0 degrees celsius. The temperature can be both negative and positive values, so the value of 0 does not imply a significant meaning.

It the following section of Goodness of Fit Test, we will first check if the regression assumptions are satisfied. If the assumptions are not satisfied, then the model can be redefined with some transformed data.

Goodness of Fit Test

We will conduct two goodness of fit tests. Firstly, we will conduct a Kolmogorov-Smirnov Test to check if the normality of random errors assumption is satisfied. After the KS test, we will conduct a Breusch-Pagan Test to check the constant variance assumption of random errors.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test compares the empirical cumulative distribution function of residuals from our model with the cumulative distribution function of normal residuals. The test statistic takes the largest absolute difference between the two distribution functions across all x values. In this paper, we will conclude the test result using the `ks.test` command from the `stats` library. The Kolmogorov-Smirnov test with the significance level of $\alpha = 0.05$ will be conducted under the hypothesis,

$$H_0 : \text{Errors follow a normal distribution.} \tag{1}$$
$$H_1 : \text{Errors do not follow a normal distribution.} \tag{2}$$

If the p-value of the test result is less than 0.05, then we reject the null hypothesis and conclude that the residuals are not normally distributed and so the model assumption is violated.

Breusch-Pagan Test

The Breusch-Pagan test examines the non-constant variance of residuals by regressing the squared residual on the independent variable, x, which should not make sense if the assumption of the constant variance of residuals is satisfied. The Breusch-Pagan test with the significance level of $\alpha = 0.05$ will be conducted under the hypothesis,

$$H_0 : \text{Variances are constant.} \tag{3}$$
$$H_1 : \text{Variances are not constant.} \tag{4}$$

by the `bptest` function from the `lmtest` package. If the p-value is computed to be less than 0.05, then we reject the null hypothesis and conclude that the residual variances are not constant, and so it violates the model assumption.

Based on the results of the goodness of fit tests, we may transform the data and fit a new model.

Maximum Likelihood Estimator

Consider a simple linear regression model, as defined in the Linear Regression section, in which the regression function is linear, i.e., $f(x) = \beta_0 + \beta_1 x$, and the response variable Y is defined as, $Y = f(X) + \epsilon = \beta_0 + \beta_1 X + \epsilon$, where the random error ϵ is normally distributed with a zero mean and a variance, σ^2. Suppose that we are given a sequence $(X_1, Y_1), ..., (X_n, Y_n)$ that is described by the above model, $Y_i = f(X_i) + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We, thus, have three unknown parameters, $\beta_0, \beta_1, \sigma^2$ and we want to estimate them using the given sample. Also assume that $X_1, X_2, ..., X_n$ are fixed values, therefore the only randomness of the function is from the random errors, ϵ_i. Then, we can say that $Y_i \sim N(f(X_i), \sigma^2)$ with a probability density function,

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(\frac{-(y-f(X_i))^2}{2\sigma^2})}$$

, and so we can state the likelihood function of the sequence $Y_1, ..., Y_n$ as,

$$f(Y_1, ..., Y_n) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}$$

The maximum likelihood estimator (MLE) approach is used to estimate the slope of the regression line, β_1. The MLE of β_1 is $\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, which equals the OLS estimator of β_1, according to the book, A Modern Approach to Regression with R by Simon J. Sheather (Sheather, 2009).

All derivations regarding the MLE can be found in Section 1 of the Appendix.

Confidence Interval

We will compute a 95% confidence interval for β_1, the slope of the regression line. The 95% confidence interval means that if thousands of samples of n items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population slope is 0.95. In general, the range of the confidence interval is defined by the sample statistic $\pm$ margin of error.

By the definition of the OLS estimator of slope coefficient of a simple linear regression, $\hat{\beta}_1$, $E[\hat{\beta}_1|X] = \beta_1$ and $Var[\hat{\beta}_1|X] = \frac{\sigma^2}{SXX}$, where $SXX = \sum_{i=1}^{n}(x_i - \bar{x})^2$. So, by the definition. $\hat{\beta}_1|X \sim N(\beta_1, \frac{\sigma^2}{SXX})$. If we standardize the conditional distribution of $\hat{\beta}_1$, we get $Z = \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{SXX}}} \sim N(0,1)$. However, by the assumption, σ^2 is the unknown but constant variance of random errors, so, we cannot use it to test hypothesis and find confidence intervals for β_1. Since σ is unknown, replacing σ by S, the standard deviation of the residuals results in,

$$T = \frac{\hat{\beta}_1 - \beta_1}{\frac{s}{\sqrt{SXX}}} \tag{5}$$

$$= \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2} \tag{6}$$

Note that the $n-2$ degree of freedom follows the formula sample size $\breve{}$ number of mean parameters estimated, and since we estimate two parameters, β_0, β_1, we have $n-2$ degree of freedom. So the sample statistic for the 95% confidence interval for β_1 is $\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$. Also, since the level of confidence is 95%, the alpha is 0.05. Thus, the margin of error is $\pm t_{n-2,0.025}$.

Putting all together, the 95% confidence interval for β_1, the slope of the regression line is $\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \pm t_{n-2,0.025}$. Equivalently $\beta_1 \in (\hat{\beta}_1 - t_{n-2,0.025} \times se(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,0.025} \times se(\hat{\beta}_1))$.

All derivations regarding the distribution of $\hat{\beta}_1$ can be found in Section 2 of the Appendix.

Hypothesis Test

We will conduct a two-tailed hypothesis test about the mean of $\hat{\beta}_1$ under the hypothesis,

$$H_0 : E[\hat{\beta}_1] = 0 \tag{7}$$

$$H_1 : E[\hat{\beta}_1] \neq 0 \tag{8}$$

By the definition and by the assumption of the linear model from the above, $E[\hat{\beta}_1|X] = \beta_1$. By the law of total expectation, $E[E[\hat{\beta}_1|X]] = E[\hat{\beta}_1] = \beta_1$, which is the slope of the regression line. So, we can redefine the hypothesis as,

$$H_0 : \beta_1 = 0 \tag{9}$$

$$H_1 : \beta_1 \neq 0 \tag{10}$$

What this test implies is the existence of the linear relationship between the temperature and the incident rate. Also, we will conduct the hypothesis test with the 95% confidence level.

Since the variance of random errors are unknown as in the section for confidence interval, the test statistics if defined as,

$$T = \frac{\hat{\beta}_1 - \beta_1^0}{se(\hat{\beta}_1)} \tag{11}$$

$$= \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-2} \tag{12}$$

Then, we can also define the rejection region as $RR : \{T \leq -t_{n-2,0.025}\}$ or $\{T \geq t_{n-2,0.025}\}$

Bayesian Credible Interval

Assume we are interested in finding a 95% credible interval of the parameter β_1. Suppose our data is a random sample of Normal random variables with mean, $\beta_0 + \beta_1 x_i$ and fixed variance, σ^2, i.e., $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$. Assume that the prior distribution of Y_i is proportional to $1/\sigma^2$ and the prior distribution of β_1 is also proportional to $1/\sigma^2$. The posterior distribution of β_1 is T distribution with $n-2$ degree of freedom centered at $\hat{\beta}_1$ and has the scale parameter, $se(\hat{\beta}_1)^2$, where the $\hat{\beta}_1$ is the OLS estimator for β_1.

All derivations regarding the posterior distribution can be found in Section 3 of the Appendix.

Results

Overall, through the analysis, we have captured a linear relationship between the incident rate of the COVID-19 and the temperature. The linear relationship is statistically significant. According to the linear model used for the entire paper, we see that the incident rate drops by 9.085% for one degree celsius increase in temperature.

Goodness of Fit Test

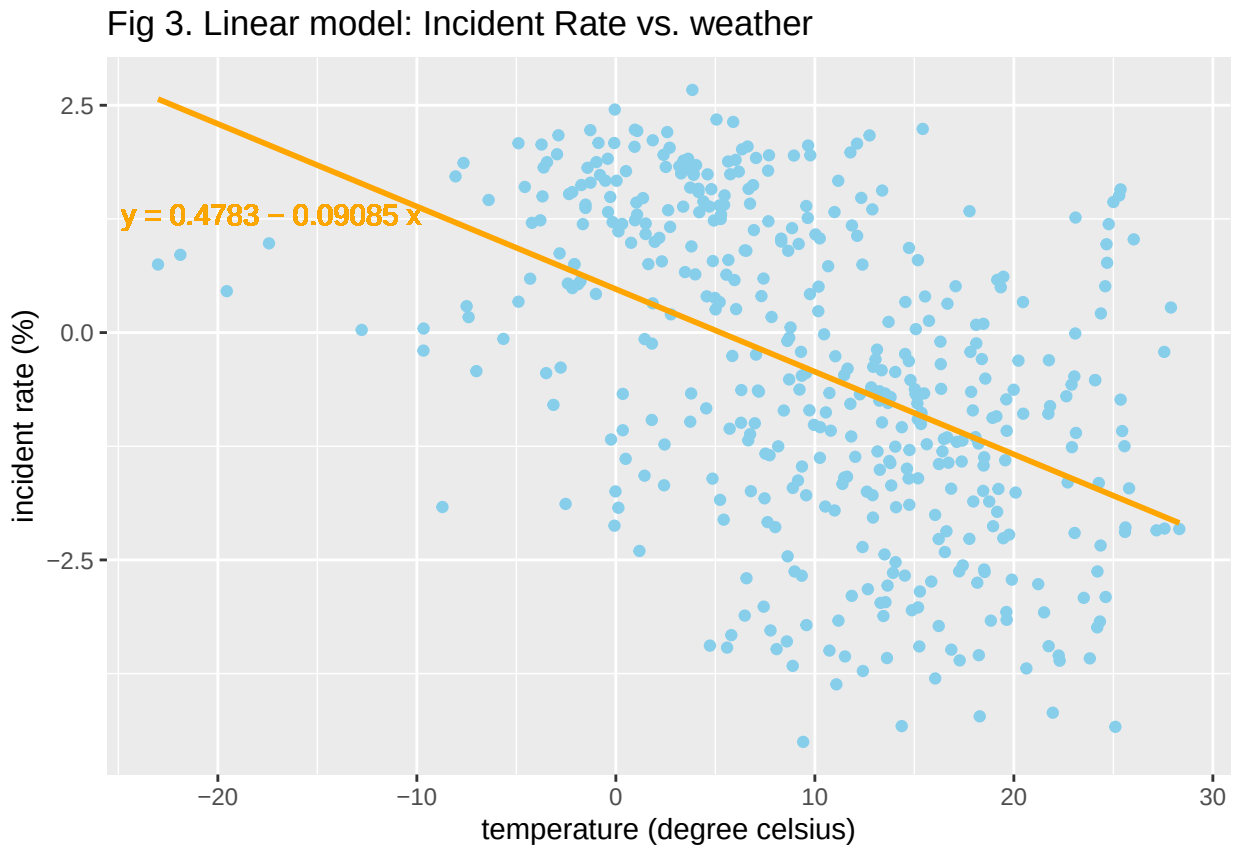
We fit a tentative simple linear model (level-level model). The equation of the line of best fit of the model is $y = 3.144 - 0.1203x$. Goodness of fit tests, the Kolmogorov-Smirnov test and the Breusch-Pagan test, are each conducted to check if the linearity assumption and the constant variance assumption of residuals is satisfied.

The KS test has the p-value $< 2.2e^{-16}$, thus we reject the null hypothesis and conclude that the residuals of the model are not normally distributed. As well, the result of the BP test shows the p-value $< 9.291e^{-8}$. Thus, we reject the null hypothesis, again, and conclude that the residuals have non-constant variance. These goodness of fit test results show that the model assumptions are not satisfied. Therefore, we will now fit another model by transforming the data. We will try to log scale the incident rate and fit the model.

The new model has a log-scaled dependent variable, $log(y)$, i.e., log scaled incident rates. The model is a log-level model and the equation of the line of best fit of the model is $y = 0.4783 - 0.09085x$. We, again, conduct the KS test and the BP test.

With the log scaled data, the KS test reports a p-value $< 3.169e^{-8}$ and the BP test reports a p-value $= 0.2035$. As a result, we still reject the null hypothesis of the KS test and conclude that the residuals are not normally distributed. On the other hand, we fail to reject the null hypothesis for the BP test since the p-value is greater than 0.05, and so we can conclude that the residuals have constant variance. According to the book, A Modern Approach to Regression with R (Sheather, 2009), the normality assumption is rather a weak assumption and for large sample sizes, the assumption is less important due to the central limit theorem. Therefore, we will continue with the new model.

Linear regression.



The simple linear model is illustrated by the scatter plot and the regression line above. The orange line is the line of best fit computed by the OLS method, using the `lm()` function.

According to the model summary, the equation of the line of best fit is,

$$log(y) = 0.4783 - 0.09085x$$

. The intercept coefficient estimate is 0.4783 and the slope coefficient estimate is -0.09085. These numbers, respectively, imply that when the monthly average temperature is at 0 degrees celsius, the log of the monthly incident rate is 3.144% and for one degree celsius increase in the monthly average temperature, the monthly incident rate decreases by 9.085%.

Maximum Likelihood Estimator

We computed the maximum likelihood estimator of the slope of the regression line, $\hat{\beta}_1$. The maximum likelihood estimate of the slope coefficient is -0.09085 and it is identical to the OLS estimate of the slope coefficient.

Confidence Interval

The upper bound and the lower bound of the 95% confidence interval for β_1, the slope of the regression line are -0.107 and -0.075 respectively. The 95% confidence interval implies that if thousands of samples of 444 items are drawn from a population using simple random sampling and a confidence interval is calculated for each sample, the proportion of those intervals that will include the true population slope is 0.95.

Hypothesis Test

The test statistic is computed to be -11.24381 and the reject region is $RR: \{T \leq -1.96535\}$ or $\{T \geq 1.96535\}$. Since our test statistic is smaller than -1.96535, we reject the null hypothesis and conclude that the slope coefficient is significantly different from 0, i.e., there exists a clear statistical evidence of linear relationship between the incident rate and the temperature.

Bayesian Credible Interval

These intervals are actually identical to the OLS confidence intervals. The key difference, however, is the interpretation. For example, for the credible interval, we can say that there is a 95% chance that the incident rate will decrease by 10.67% up to 7.5% for every additional one degree celsius increase in the temperature.

Conclusions

In this paper, we examined the weather effect on the incident rate of the COVID-19 over the 37 OECD countries. We adopted the method of simple linear regression analysis to capture and analyze the relationship between the incident rate and the temperature.

Before formally analyzing the data with statistical tools, such as fitting a linear model, conducting hypothesis tests and goodness of fit tests, we looks at the data by checking the numerical summary and visualizing the data. This step alarmed some problems of the data and so we conducted goodness of fit test to verify assumptions of a linear model before selecting a model. As a result, the tests proved that the base model (level-level model) is not appropriate, Thus, we log transformed the dependent variable, the incident rates, and then fit the new, log-level model. The model satisfied the assumptions, so we continues the analysis with the model.

The most important information we obtained from the linear model is the slope estimator. The slope estimator estimates the slope of the population regression line. The estimate is computed by the OLS method and the MLE approach. The results were both -0.09085, as expected. The hypothesis test of the mean of the slope estimator was conducted because the mean of the slope estimator is the slope coefficient by the definition. The hypothesis test, therefore, was used to test the existence of the linear relationship between the incident rate and the temperature and by the test results, we concluded that there is a statistically significant linear relationship. The value of the slope estimator shows that the incident rate decreases by 9.085% if the temperature increases by 1 degree celsius.

The 95% confidence interval that is computed by the OLS method and by the t-value and the 95% credible interval computed by the Bayesian approach resulted in the same value, (-0.10673,-0.07497).

Weaknesses

The most obvious weakness is that the model, in fact, do not meet the assumption that the random errors are normally distributed. The violation of this assumption is considered insignificant and we kept the model because the random errors have a constant variance. Although the normality assumption is not a very powerful assumption and often overlooked because of the CLT. It is better if the assumption was satisfied.

Another weakness is that we only consider the temperature as a main factor of weather. Actually, there are other components of weather, such as wind, humidity, temperature, etc. Simply treating temperature as weather may incur bias and oversimplify the model.

Next Steps

As mentioned previously, since the variable used to denote weather is confined. It is recommended to include more variables indicating weather condition, such as humidity, wind, precipitation, etc. Adding more variables does not always results in a better model. However, we can find a better model by comparing how well candidate models with different weather variables explain the data.

Discussion

The weather effect on the incident rate of the COVID-19 is analyzed by a simple linear model. Firstly, the level-level model is used to test the goodness of fit of the residual normality and the constant variance. The level-level model is rejected. Accordingly, the log-level model is accepted by the test and used throughout the paper. The linear model reports that for one degree celsius increase in the temperature, the incident rate of the COVID-19 decreases by 9.085%, which is computed by the OLS method. The slope is statistically proven to be significant. The MLE approach is also used to estimate the slope and it is the same as the OLS estimator. The confidence interval and the credible intervals are estimated as (-0.10673,-0.07497). The negative relationship shows that the COVID-19 is less infectious when the temperature is low and therefore shows a similar infection pattern as the common cold and flu. The infection pattern is important because the pattern can be taken into account when governments implement preventive measures. Until now, the COVID-19 preventive measures are implemented as a response to spikes and drops in the number of confirmed cases, rather to prevent in advance. It is not because the governments are not smart but because the COVID-19 a new virus, so they lack the information. Therefore, this result can be used to implement more stringent preventive measures in the so called "flu season".

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)
4. *Similarities and Differences between Flu and COVID-19.* (2021a, January 27). Centers for Disease Control and Prevention. https://www.cdc.gov/flu/symptoms/flu-vs-covid19.html
5. Corman, V. M., Muth, D., Niemeyer, D., & Drosten, C. (2018). *Hosts and Sources of Endemic Human Coronaviruses. Advances in Virus Research*, 163–188. https://doi.org/10.1016/bs.aivir.2018.01.001
6. Sheather, S. (2009). *A Modern Approach to Regression with R (Springer Texts in Statistics)* (2009th ed.). Springer.
7. *COVID-19 Portal*. (2021, April 25). Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/
8. *World Bank Climate Change Knowledge Portal*. (n.d.). For Global Climate Data and Information! https://climateknowledgeportal.worldbank.org/watershed/154/climate-data-historical
9. *Demography - Population - OECD Data.* (n.d.). The OECD. https://data.oecd.org/pop/population.htm

Appendix

Section 1

Given any data set $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, probability density of the simple linear model is,

$$\prod_{i=1}^{n} p(y_i|x_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{(y-(\beta_0+\beta_1 x_i))^2}{2\sigma^2})}$$

Since the parameters, $\beta_0, \beta_1, \sigma^2$ are unknown, we replace the unknown parameters by (b_0, b_1, s^2). Thus,

$$\prod_{i=1}^{n} p(y_i|x_i; b_0, b_1, s^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} e^{-(\frac{(y-(b_0+b_1 x_i))^2}{2s^2})} \tag{13}$$

$$= (\frac{1}{\sqrt{2\pi s^2}})^n e^{-\frac{1}{2s^2}\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2} \tag{14}$$

$$\tag{15}$$

This is the likelihood, a function of the parameter values. It's just as informative, and much more convenient, to work with the log-likelihood,

$$L(b_0, b_1, s^2) = \log \prod_{i=1}^{n} p(y_i|x_i; b_0, b_1, s^2) \tag{16}$$

$$= \log(\frac{1}{\sqrt{2\pi s^2}})^n e^{-\frac{1}{2s^2}\sum_{i=1}^{n}(y_i-b_0-b_1x_i)^2} \tag{17}$$

$$= -\frac{n}{2}\log 2\pi - n\log s - \frac{1}{2s^2}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 \tag{18}$$

$$\tag{19}$$

Given the log-likelihood function, the value of b_1 that maximizes the function is computed by taking a partial derivative of the function w.r.t. b_1 and setting it equal to 0. However, the b_1 is a function of b_0, so we will find the value of b_0 that maximizes the log-likelihood function first by the same partial derivative method.

$$\frac{d}{db_0} L(b_0, b_1, s^2) = -\frac{1}{2s^2}\sum_{i=1}^{n} -2(y_i - (b_0 + b_1 x_i)) \tag{20}$$

$$= \frac{1}{s^2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) \tag{21}$$

$$= 0 \tag{22}$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \tag{23}$$

$$\Rightarrow \sum_{i=1}^{n} y_i - nb_0 - b_1\sum_{i=1}^{n} x_i = 0 \tag{24}$$

$$\Rightarrow b_0 = \bar{y} - b_1\bar{x} \tag{25}$$

$$\frac{d}{db_1}L(b_0, b_1, s^2) = -\frac{1}{2s^2}\sum_{i=1}^{n} -2x_i(y_i - (b_0 + b_1 x_i)) \tag{26}$$

$$= \frac{1}{s^2}\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) \tag{27}$$

$$= 0 \tag{28}$$

$$\Rightarrow \sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0 \tag{29}$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{30}$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - (\bar{y} - b_1\bar{x}) \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{31}$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i + b_1\bar{x} \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{32}$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - b_1(\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i) = 0 \tag{33}$$

$$\Rightarrow b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} \tag{34}$$

$$\Rightarrow b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{35}$$

The maximum likelihood estimator of β_1 is $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$.

Section 2

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{36}$$

$$= \frac{SXY}{SXX} \tag{37}$$

$$\text{Since } \sum_{i=1}^{n}(x_i - \bar{x}) = 0, \text{ we can find,} \tag{38}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) \tag{39}$$

$$= \sum_{i=1}^{n}(x_i - \bar{x})y_i \tag{40}$$

$$c_i = \frac{x_i - \bar{x}}{SXX} \Rightarrow \hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \tag{41}$$

$$\tag{42}$$

Using the alternative equation of $\hat{\beta}_1$ just defined, we can find the distribution of $\hat{\beta}_1$.

$$E(\hat{\beta}_1|X) = E[\sum_{i=1}^{n} c_i y_i | X = x_i] \tag{43}$$
$$= \sum_{i=1}^{n} c_i E[y_i|X = x_i] \tag{44}$$
$$= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) \tag{45}$$
$$= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \tag{46}$$
$$= \beta_0 \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{SXX}) + \beta_1 \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{SXX}) x_i \tag{47}$$
$$= \beta_1 \tag{48}$$

$$Var(\hat{\beta}_1|X) = Var[\sum_{i=1}^{n} c_i y_i | X = x_i] \tag{49}$$
$$= \sum_{i=1}^{n} c_i^2 Var[y_i|X = x_i] \tag{50}$$
$$= \sigma^2 \sum_{i=1}^{n} c_i^2 \tag{51}$$
$$= \sigma^2 \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{SXX})^2 \tag{52}$$
$$= \frac{\sigma^2}{SXX} \tag{53}$$

By the assumption, the random errors of the linear model are normally distributed. Also, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $Y|X$ is normally distributed. Since $\hat{\beta}_1|X$ is a linear combination of y_i's, $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{SXX})$

Section 3

By the assumption that the errors, $\epsilon_i \sim N(0, \sigma^2)$, we can state that $Y_i|x_i, \beta_0, \beta_1, \sigma^2 \sim N(\beta_0 + \beta_1 x_i, \sigma^2), \forall i = 1, 2, ..., n$. So, the likelihood of $Y_i|x_i, \beta_0, \beta_1, \sigma^2$ is,

$$p(y_i|x_i, \beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{(y-(\beta_0+\beta_1 x_i))^2}{2\sigma^2})}$$

and the likelihood of $Y_1, Y_2, ..., Y_n$ is the product of each likelihood $p(y_i|x_i, \beta_0, \beta_1, \sigma^2)$, which is,

$$\prod_{i=1}^{n} p(y_i|x_i; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{(y-(\beta_0+\beta_1 x_i))^2}{2\sigma^2})}$$

By the assumption, we defined the prior density $p(\beta_0, \beta_1, \sigma^2) \propto \frac{1}{\sigma^2}$. Then we apply the Bayes' rule to derive the joint posterior distribution after observing data $y_1, y_2, ..., y_n$

$$p(\beta_0, \beta_1, \sigma^2|y_1, y_2, ..., y_n) \propto (\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\frac{(y-(\beta_0+\beta_1 x_i))^2}{2\sigma^2})}) \times \frac{1}{\sigma^2} \tag{54}$$

$$\propto (\frac{1}{\sqrt{2\pi\sigma^2}})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2} \times \frac{1}{\sigma^2} \tag{55}$$

$$\propto \frac{1}{\sigma^{(n+2)}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2} \tag{56}$$

Integrating out β_0, σ^2 from the joint posterior distribution, we get the marginal posterior distribution of β_1,

$$p(\beta_1|y_1, y_2, ..., y_n) = \int_0^\infty (\int_{-\infty}^{\infty} p(\beta_0, \beta_1, \sigma^2|y_1, y_2, ..., y_n) d\beta_0) d\sigma^2 \tag{57}$$

$$\beta_1|y_1, y_2, ..., y_n \sim t(n-2, \hat{\beta}_1, se(\hat{\beta}_1)^2) \tag{58}$$

The second line means that marginal posterior distribution of β_1 follows a T-distribution with $(n-2)$ degree of freedom, centred at the OLS estimator of β_1 and scaled by the standard error squared.

Therefore, the credible interval of β_1 is $\beta_1 \in (\hat{\beta}_1 - t_{n-2,0.025} \times se(\hat{\beta}_1), \hat{\beta}_1 + t_{n-2,0.025} \times se(\hat{\beta}_1))$.