TS1

Dylan Hayashi

MSDS 413, Fall 2022, 55

Northwestern University, Time-Series Analysis & Forecasting

September 25, 2022

# 1. Gauss-Markov Assumptions Review

## 1.1 Research question

Can a stores' expenditures, number of accounts, number of competitors, and district potential be used to explain and predict their sales?

## 1.2 - MLR Model

The following is a Multiple Linear Regression equation with sales as the dependent variable and expenditures, numbers of accounts, number of competitors, and district potential as the independent variables.

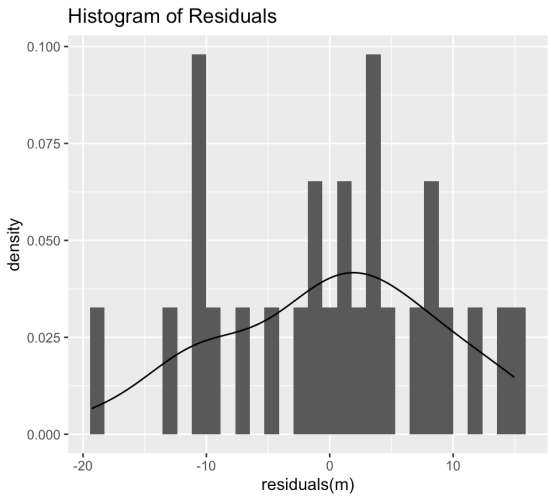
$$sales = \beta_0 + \beta_1 expenditures + \beta_2 account + \beta_3 competitors + \beta_4 potential + \varepsilon$$

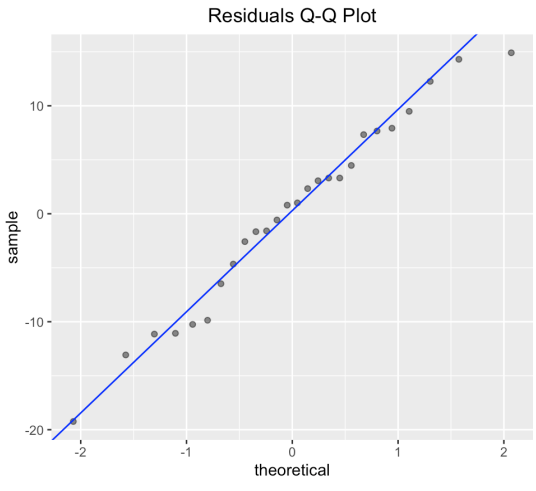
The assumptions for this model are as follows:

1. Linearity: There is a linear relationship between the dependent variable sales and the independent variables expenditures, number of accounts, number of competitors, and district potential.
2. Multicollinearity: The independent variables are not correlated with each other.
3. Independence: The observations are independent of one another.
4. Normality: The error terms are normally distributed.
5. Homoscedasticity: The error terms have a constant variance.

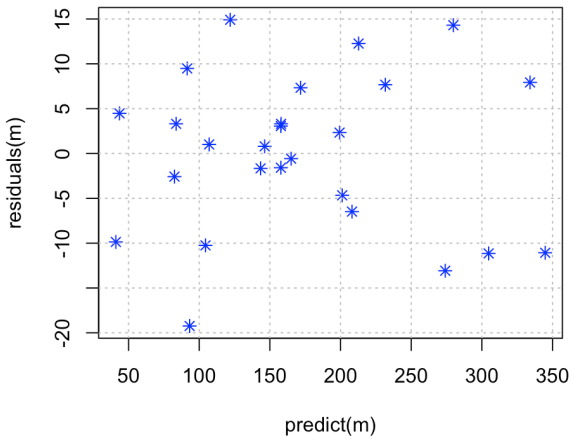
## 1.3 - Testing Assumptions of Normality and Constant Variance

The residuals of this regression are not normally distributed. This is most obviously demonstrated by looking at their histogram, shown below on the left. Other methods of assessing normality make the residuals appear approximately normal. The residuals in the QQ Plot on the right do fall close to the line. Similarly, the 95% confidence intervals for both skew (-0.805, 0.4334) and excess kurtosis (-1.4, 0.6) include 0.



The variance of the residuals appear approximately constant. As the scatter plot on the right shows, the minimums and maximums of residuals appear evenly distributed across the x-axis, aside from the global minimum in the bottom left corner.




### 1.4 - Parameter Estimation and F-Test

Executing the regression produced the summary table shown below. The resulting equation is:

$$sales = 179.69 + 1.821Expenditures + 3.3434Accounts - 21.348Competitors + 0.327potential$$

The results of a Global F-Test are also shown in the summary table. A Global F-Test assesses the overall significance of a model by comparing it to an intercept-only model. The F-Statistic for this model was 479.1, which indicates that this MLR model is considerably more significant than the intercept-only model, and therefore has explanatory power.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 179.6928    13.0601  13.759 5.62e-12 ***
expend        1.8210     1.0894   1.672    0.109
accounts      3.3434     0.1641  20.368 2.60e-15 ***
comp        -21.3480     0.7940 -26.887  < 2e-16 ***
potential     0.3270     0.4714   0.694    0.495
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.678 on 21 degrees of freedom
Multiple R-squared:  0.9892,    Adjusted R-squared:  0.9871
F-statistic: 479.1 on 4 and 21 DF,  p-value: < 2.2e-16
```

### 1.5 - Interpretation of the Parameter for Number of Competitors

The parameter estimate for the variable ‘number of competitors’ produced by the MLR was -21.348. This means that, on average, an increase of one competitor was associated with a decrease in store sales of $21,348 with all other things equal.

### 1.6 - Performance and Interpretation of a t-test for Number of Competitors

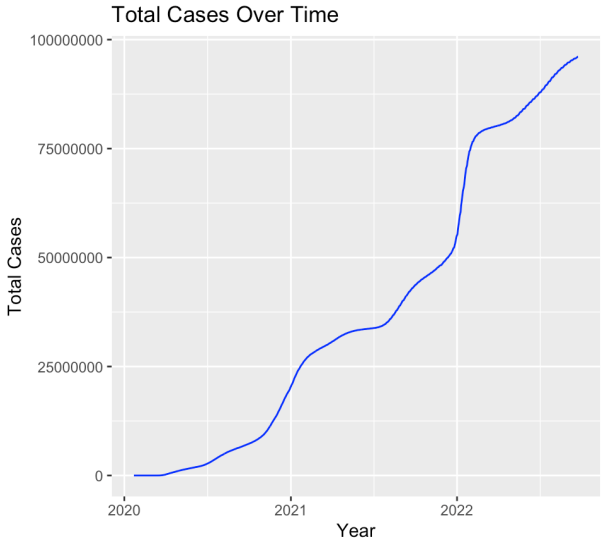
By executing the regression analysis, a t-test for the variable ‘number of competitors’ was performed, and the test value was $2.6e^{-15}$. The null hypothesis of this test is that the parameter value for ‘number of competitors’ is equal to 0. The resulting test value indicates that we can

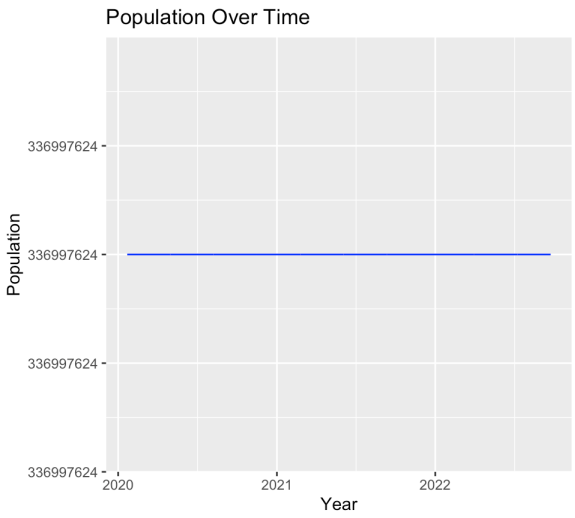
reject the null hypothesis, even at the 0.001 level of significance. This means that the variable, number of competitors, has significance.

## 2. Identifying a Time Series

### 2.1 - Time Plots

The time plots below show time on the x-axis and the variables cases and population on the y-axis.




### 2.2 - Time-Ordered Sequence

In order to be classified as time-series data, the Covid dataset must contain two variables (total cases and population), indexed by date, and there must be a value of total cases and poulation for each date. Thus, our null hypothesis is H10 : xit, i ∈ {1, 2}t ∈ {1, 2, .., n}, and our alternative hypothesis is H1a: not H10. To test this, we can check several features:

1. Each date in the date variable is unique. The code to the right shows that there are 987 observations in the date index, and 987 unique observations of dates, indicating that each date is unique.
2. There is an observation of total cases and population for each date. As the code also shows, there are 978 unique observations in the total cases and population variables.

```
> length(unique(X$date))
[1] 978
> length(X$date)
[1] 978
> length(X$total_cases)
[1] 978
> length(X$population)
[1] 978
```

Thus, we accept our null hypothesis.

### 2.3 - Constant Interval

In order to be classified as time-series data, the Covid dataset must also have identical intervals across its time axis. Thus our null hypothesis is that H20 : $(t + 1) - t = c, t \in 1,2,..,n$, and our alternative hypothesis is H2a: not H20. To test this, we can sum the difference between each observation and its successive observation. If this value is equal to the number of observations minus one, then we can accept our null hypothesis that the dates are evenly spaced. This is because for each observation except the last, the difference between itself and the next date should be 1. The code shown on the right run executes this test and returns a value of 977, one less than the number of observations, 978. Thus, we can accept our null hypothesis.

```
> dif <- diff(as.Date(X$date,"%Y-%m-%d"))
> nrow(X)
[1] 978
> table(dif)   # nrow - 1?
dif
  1
977
```

### 2.4 - Time Series Classification and Justification

Based on the results of the last two sections, the Covid dataset meets the requirements to be classified as time-series data. It contains two variables, total case and time. It is indexed by time with even intervals, and there is an observation of total cases for each date.

## 3. Time Series as a Stochastic Process

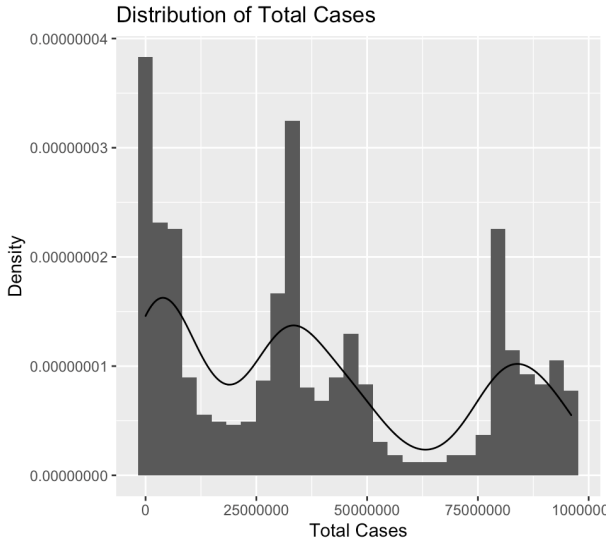
### 3.1 - EDA

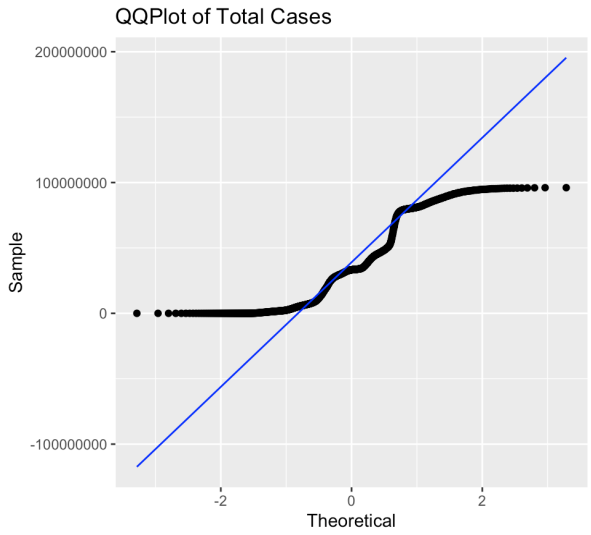
The two tables below contain summary statistics for the total cases and population variables.

| | X..X.total_cases |
|---|---|
| nobs | 977.000000 |
| NAs | 0.000000 |
| Minimum | 1.000000 |
| Maximum | 96065161.000000 |
| 1. Quartile | 6914121.000000 |
| 3. Quartile | 71152549.000000 |
| Mean | 38106313.611055 |
| Median | 33266251.000000 |
| Sum | 37229868398.000000 |
| SE Mean | 1011801.612860 |
| LCL Mean | 36120756.598744 |
| UCL Mean | 40091870.623365 |
| Variance | 1000196426199163.000000 |
| Stdev | 31625882.220093 |
| Skewness | 0.443616 |
| Kurtosis | -1.152647 |

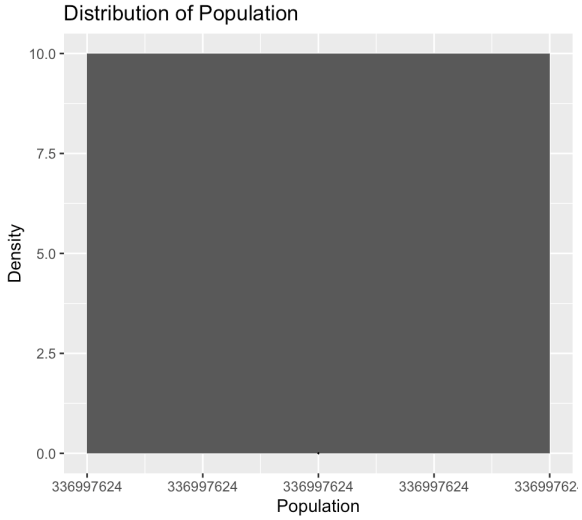
| | X..X.population |
|---|---|
| nobs | 977 |
| NAs | 0 |
| Minimum | 336997624 |
| Maximum | 336997624 |
| 1. Quartile | 336997624 |
| 3. Quartile | 336997624 |
| Mean | 336997624 |
| Median | 336997624 |
| Sum | 329246678648 |
| SE Mean | 0 |
| LCL Mean | 336997624 |
| UCL Mean | 336997624 |
| Variance | 0 |
| Stdev | 0 |
| Skewness | NaN |
| Kurtosis | NaN |

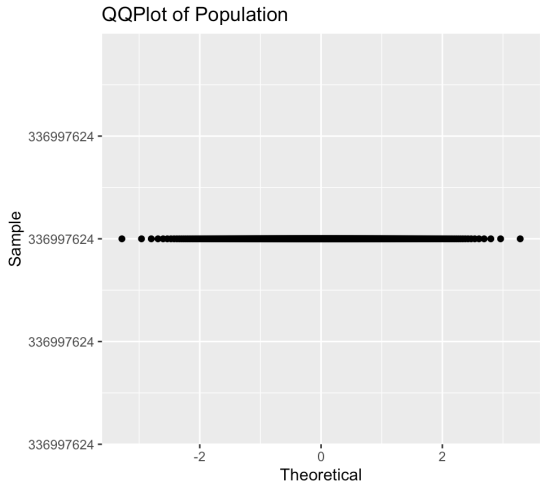
It can be clearly demonstrated by looking at the histograms and QQ plots of total cases and population that neither variable is normally distributed. As one can see by its histogram, the distribution of total cases is tri-modal, rather than unimodal. As the QQ plot shows, these observations do not fall close to the straight line that represents normality.




The case for lack of normality in population is similarly obvious. Population is constant, hence why its histogram appears like a rectangle with an x-axis dimension of 1. This is far from the bell-shaped normal curve. Similarly, the observations in the QQ plot of population form almost a horizontal line, which would intersect the 45 degree line of normality, rather than fall on it.



### 3.2 - Estimation of Mean and Confidence Intervals

A t-test for the population mean of total cases being equal to 0 produced a p-value of $2.2e^{-15}$ and a 95% confidence interval of (36178652, 40152512). This p-value, far smaller than 0.01, allows us to reject the null hypothesis that the population mean was 0.

A t-test for the population mean of population was not possible to execute because the variable is constant. Given the sample mean and sample standard deviation of 0, a 95T confidence interval would be (336997624, 336997624). This interval does not contain 0.

### 3.3 - Hypothesis Testing

The sample skew for total cases was 0.5536, and its excess kurtosis was -1.1527. The population variable did not have skew or kurtosis. To test the population values, we can make a null hypothesis that the population skew and kurtosis are equal to zero. The alternative hypothesis is that they are not.

For total cases, the resulting 95% confidence interval for skew is (0.438, 0.445), and the interval for kurtosis is (-1.157, -1.148). Neither of these intervals include zero, so we can reject the null hypotheses that the population skew and kurtosis of total cases are equal to 0.

R was not able to make the same calculations to test the skew and kurtosis of the population, nor was it able to make the 95% confidence interval. This was because the population variable had no variance, and thus skew and kurtosis could not be calculated.

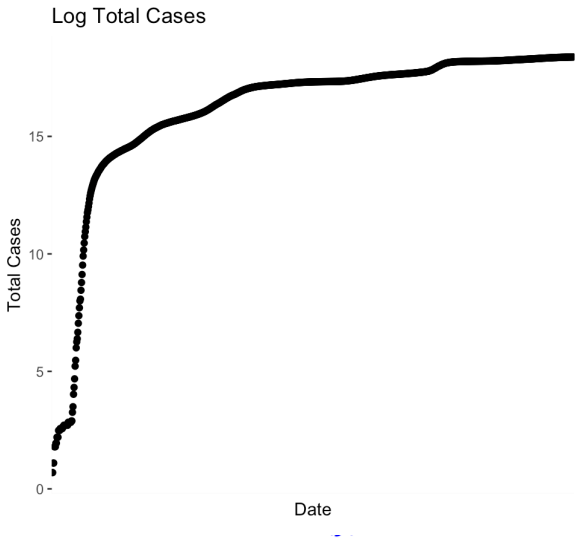
### 3.4 - Stochasticity and Gaussian Classification and Justification

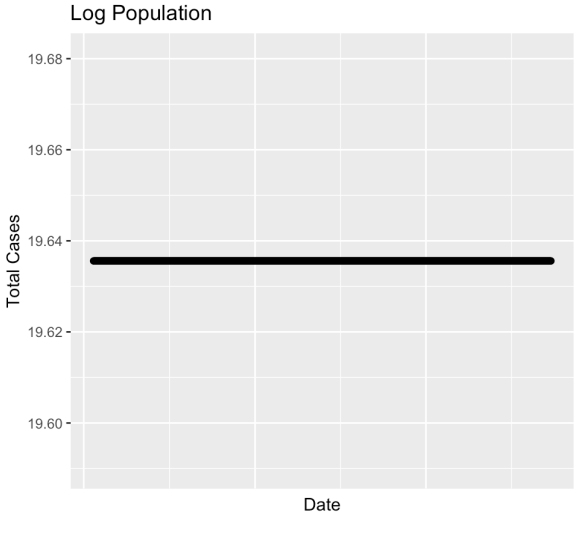
As shown in the summary statistics tables, the variable Total Cases does have variance (1002608166874907) and the variable population does not (0). Thus, we can conclude that the total cases variable does have stochasticity and the population variable does not. This means that total cases dataset all three criteria and is a valid time-series, whereas population fails to satisfy the third criterion and is not a valid time-series.

## 4. Obtaining a Linear Additive Model

### 4.1 - EDA

The following two line plots are of log total cases and log population over time. The regular and log total cases plots behave similarly in the sense that the overall value of the variable increases, but the slopes are different. Unsurprisingly, the line in the plot looks close to a logarithmic function. The regular and log population variables also behave identically, they are both constant.




| | y |
|---|---|
| nobs | 977.000000 |
| NAs | 0.000000 |
| Minimum | 0.693147 |
| Maximum | 18.380537 |
| 1. Quartile | 15.749077 |
| 3. Quartile | 18.080337 |
| Mean | 16.129814 |
| Median | 17.320054 |
| Sum | 15758.828255 |
| SE Mean | 0.110881 |
| LCL Mean | 15.912221 |
| UCL Mean | 16.347407 |
| Variance | 12.011835 |
| Stdev | 3.465809 |
| Skewness | -2.930575 |
| Kurtosis | 8.450659 |

| | y |
|---|---|
| nobs | 978.00000 |
| NAs | 0.00000 |
| Minimum | 19.63559 |
| Maximum | 19.63559 |
| 1. Quartile | 19.63559 |
| 3. Quartile | 19.63559 |
| Mean | 19.63559 |
| Median | 19.63559 |
| Sum | 19203.60354 |
| SE Mean | 0.00000 |
| LCL Mean | 19.63559 |
| UCL Mean | 19.63559 |
| Variance | 0.00000 |
| Stdev | 0.00000 |
| Skewness | NaN |
| Kurtosis | NaN |

### 4.2 - Estimation of Mean and Confidence Intervals

The sample mean for log total cases was 16.13, and the sample mean for log population was 19.636. Applying t-tests to test the null hypotheses that the population means of these variables are zero, and the alternative hypotheses that they are not, produced the following results:
Log total cases had a p-value of $2.2e^{-15}$ and a 95% confidence interval of (15.915, 16.35). This p-value allows us to reject the null hypothesis that the mean population is 0. As shown above, the interval does not contain 0.

Log population, because it is constant, was not able to be subjected to a t-test. Thus, there is no p-value or confidence interval.

### 4.3 - Hypothesis Testing

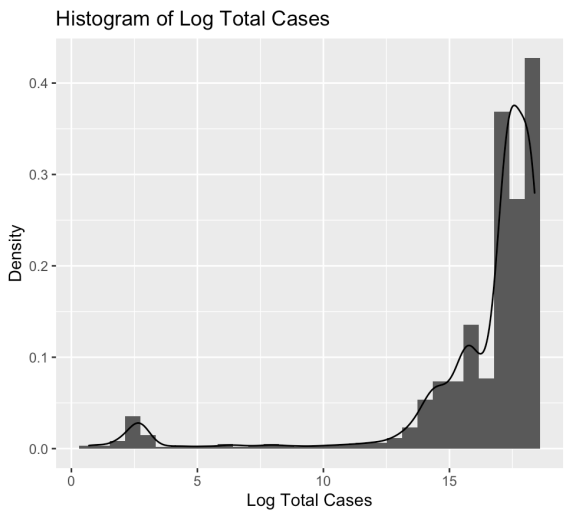
The sample skew for log total cases was -2.931 and its excess kurtosis was 8.46. The log population variable did not have skew or kurtosis. To test the population values, we can make a null hypothesis that the population skew and kurtosis are equal to zero. The alternative hypothesis is that they are not.

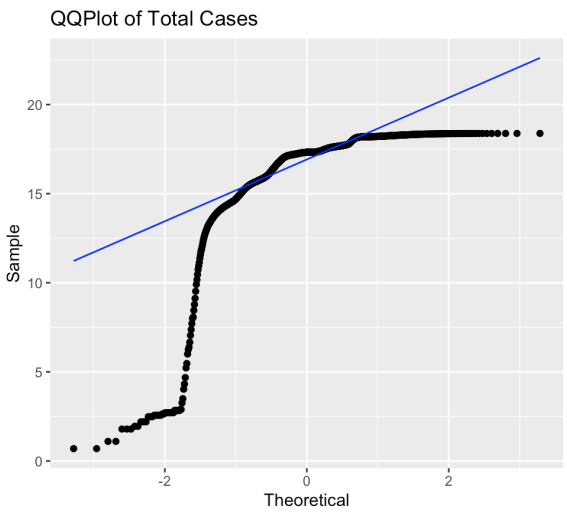
For log total cases, the resulting 95% confidence interval for skew is (-2.943,-2.93), and the interval for kurtosis is (8.257, 8.409). Neither of these intervals include zero, so we can reject the null hypotheses that the population skew and kurtosis of total cases are equal to 0.

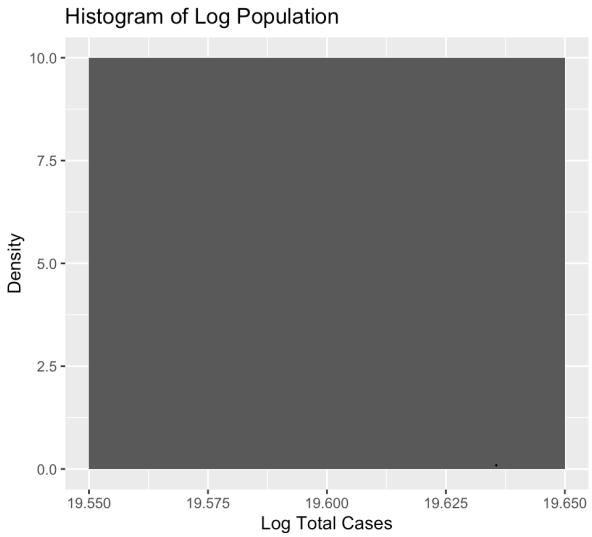
R was not able to make the same calculations to test the skew and kurtosis of the population, nor was it able to make the 95% confidence interval. This was because the log population variable had no variance, and thus skew and kurtosis could not be calculated.

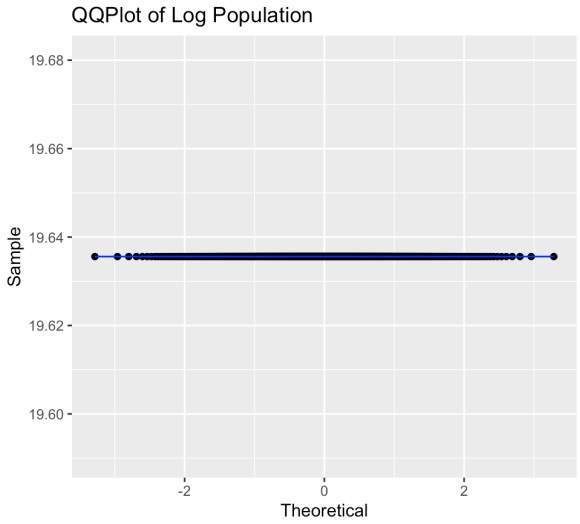
### 4.4 - Stochasticity, Gaussian, and Additive Classification and Justification

The following plots are histograms and QQ plots of log total cases and log population. Neither of the variables appear to be normally distributed. Log total cases is mostly unimodal, with a partial skew to the left. The observations do not lie on the line either. Similar to previous visualizations of population, the histogram and QQ plot show constancy.



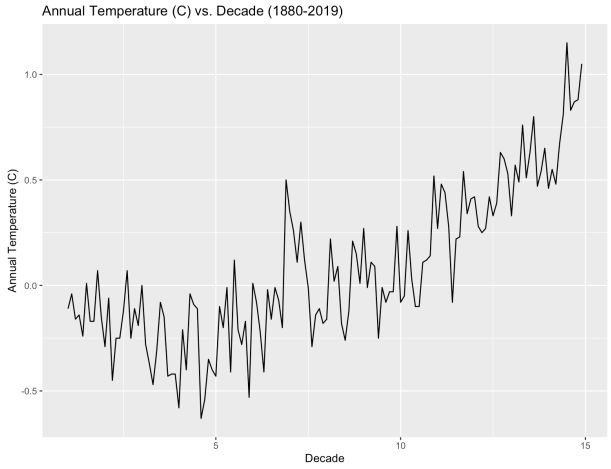



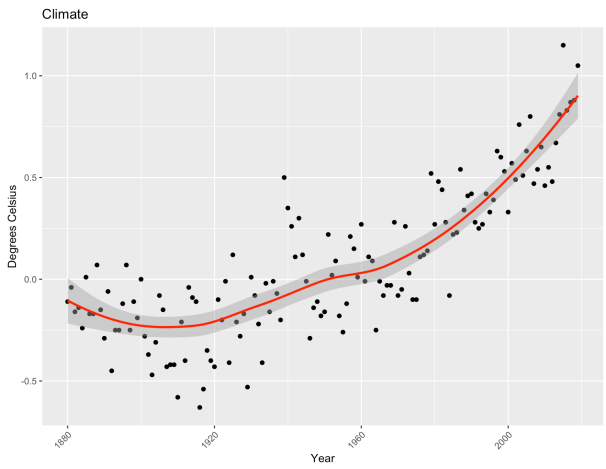

The data summaries in part 2 show that the log total cases variable does have variance (12), implying that it is stochastic, but log population does not have variance (0), implying it is not stochastic.

## 5. Decomposition of a Time Series

### 5.1 - Time Plot and Features

The two plots below are of climate over time. The left is a time plot in which each successive observation is joined by a line, and the right is a scatter plot with a line of best fit.

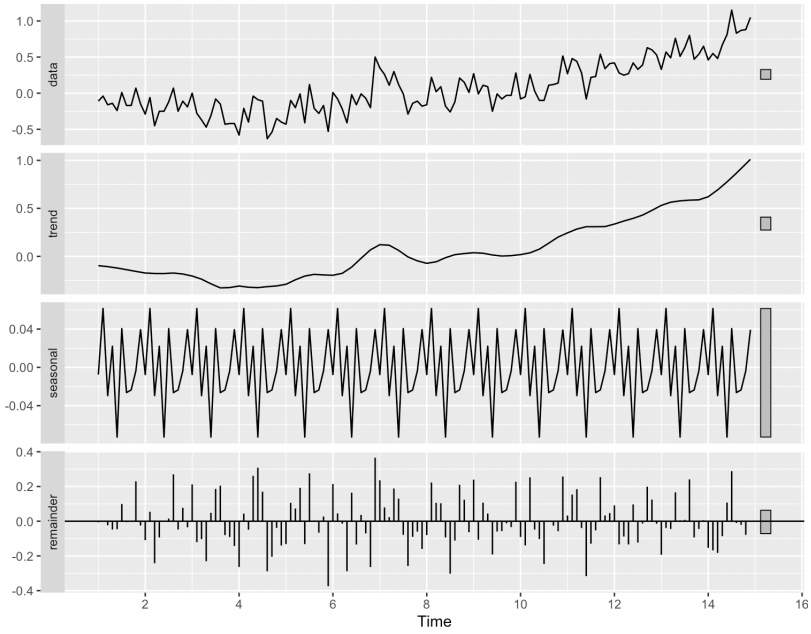



As demonstrated by the line of best fit, there appears to be an overall upward trend in the data over time. However, between decades 0 and 5, there is a downward trend, after which the slope of the line becomes positive. The line plot demonstrates that there is a large degree of variation, although it is not discernable what a seasonal component may look like. To me, the variance does not look constant. Instad, it appears greater in the first 7 decades than the latter 7.

### 5.2 - STL Decomposition

The visualization on the right shows the results of STL decomposition. As I predicted, the trend line is decreasing between decades 0 and 5, after which point it inflects and develops a positive slope. This decomposition also makes apparent the structure of the seasonal component: it features four local maxima and three local minima, with the first maximum and second minimum being universal.

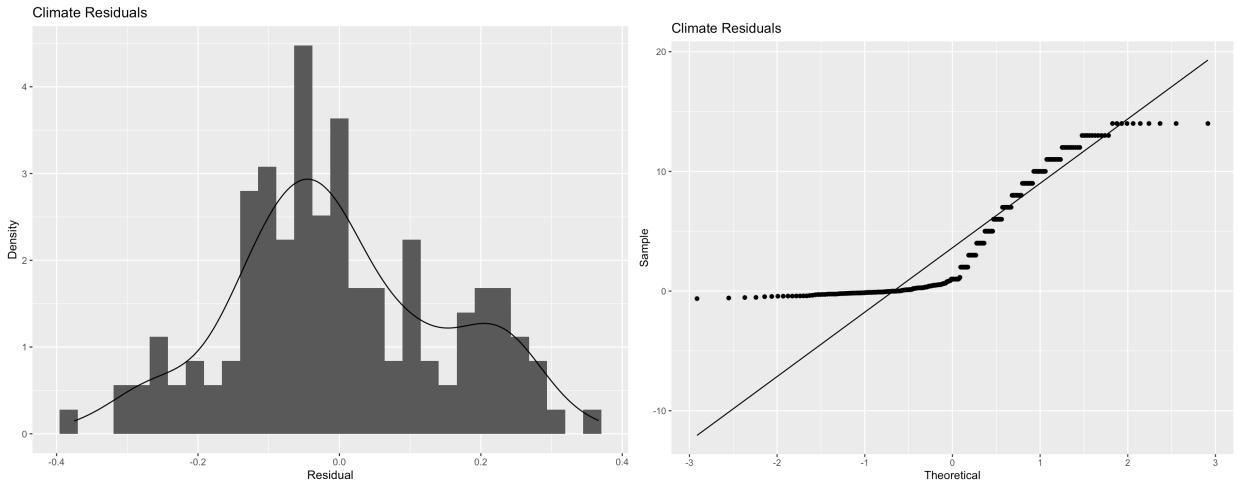



### 5.3 - Variation by Component

The seasonal component of the decomposition had a variance of 0.001512, which is 1.15% of the variance. Seasonality had much greater variariance, with a total of 0.106145, which is 80.75% of the variance. Lastly, the residuals had a variance of 0.022661, which is 17.24% of the variance.

### 5.4 - EDA of Residuals and Gaussian Classification and Justification

The residuals of this model are not normally distributed. While close to uni-modal, the histogram is almost bi-modal and does not have symmetry. The QQ plot shows that the observations deviate from the normal line.




The mean of the residuals is -0.0019. A t-test of the null hypothesis that the population residuals mean is equal to 0 returned a p-value of 0.8809 with a confidence interval of (-0.027,  0.0232). Given the size of the p-value, we are unable to reject the null hypothesis at the 95% confidence level.

The skew of the residuals is 0.149. A 95% confidence interval for the population residual skew is (0.137, 0.148), allowing us to reject the null hypothesis of a t-test that the population residual skew is equal to 0 at the 95% confidence level.

The excess kurtosis of the residuals is -0.429. A 95% confidence interval for the population residual skew is (-0.438, -0.409), allowing us to reject the null hypothesis of a t-test that the population residual kurtosis is equal to 0 at the 95% confidence level.