Statistical Analysis of Covid-19 Deaths in the US

In this project, I created a code in order to compare the trends in mortality rate due to Covid-19 and six different predictors as variables. These included: income, average age, unemployment rate, ratio of people to hospitals, population density, and percentage of uninsured people in each state. The purpose of this project is to find out whether or not the Covid mortality rate was correlated to any of the variables stated above. In order to do so, I created a model for this data and graph it in order to see whether the data would fit the model or not. Based on the results, I would check whether there was or there wasn´t a correlation between the deaths and the variables.

1. Significance of the Regression

First, I will use code in order to create a model for the data, in which I will see if the error is significant, as well as which variables are significant and which aren´t.

Code:

```
library(AER)
data <- read.csv("C:\\Users\\galon\\OneDrive\\Escritorio\\College Stuff\\Year2\\IE330\\Project2\\project2.csv")
state <- data[,1]
mortalityRate <- data[,2]
income <- data[,3]
averageAge <- data[,4]
unemployment <- data[,5]
ratioPH <- as.numeric(data[,6])
population <- as.numeric(data[,7])
uninsured <- data[,8]

model <- lm(mortalityRate~income + averageAge + unemployment + ratioPH +population + uninsured)
summary(model)
```

Output:

```
> model <- lm(mortalityRate~income + averageAge + unemployment + ratioPH +population + uninsured)
> summary(model)

Call:
lm(formula = mortalityRate ~ income + averageAge + unemployment +
    ratioPH + population + uninsured)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041609 -0.013293 -0.001651  0.009608  0.066197

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.145e-01  9.514e-02   3.306 0.002076 **
income       -2.125e-06  5.289e-07  -4.018 0.000268 ***
averageAge   -4.627e-03  1.899e-03  -2.437 0.019626 *
unemployment  1.081e-03  1.829e-03   0.591 0.557817
ratioPH       7.216e-07  4.862e-07   1.484 0.146007
population    1.657e-04  2.390e-05   6.932 3.04e-08 ***
uninsured    -1.085e-03  1.525e-03  -0.711 0.481159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02385 on 38 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6075,    Adjusted R-squared:  0.5455
F-statistic: 9.803 on 6 and 38 DF,  p-value: 1.585e-06
```

From this output of code, we will be focusing more specifically on the following output line:

```
F-statistic: 9.803 on 6 and 38 DF,  p-value: 1.585e-06
```

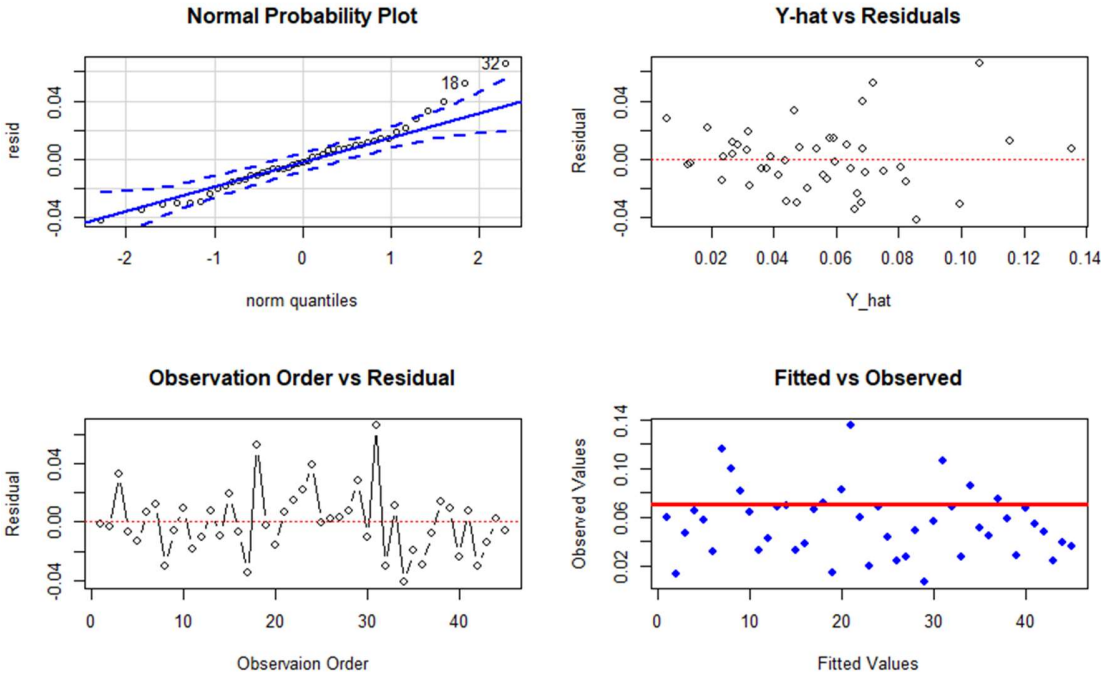
The model is significant if the p-value of the F-Statistic is less than a certain value which is deemed to be an accurate measurement of the significance of a model. Generally 0.05 is used as a p-value that determines if a model is significant. Because the p-value for F-statistic is 1.585e-6 < 0.05. Therefore, we reject the null hypothesis at $\alpha = 0.05$ and conclude that the regression is significant.

2. Validity of the Regression

Code:

```
model <- lm(mortalityRate~income + averageAge + unemployment + ratioPH +population + uninsured)
summary(model)
resid <- model$residuals

par(mfrow=c(2,2))

qqPlot(resid, main='Normal Probability Plot')

plot(model$fitted.values, resid, main='Y-hat vs Residuals', xlab='Y_hat', ylab='Residual')
abline(h=0, lty=3, col='red')

plot(model$residuals, type='b', main='Observation Order vs Residual', xlab="Observaion Order", ylab="Residual")
abline(h=0, lty=3, col='red')

plot(predict(model),pch=19,col='blue',main='Fitted vs Observed',ylab = 'Observed Values',xlab = 'Fitted Values')
abline(0.07,0,lwd=3,col='red')
```

Output:



- From looking at the Normal Probability Plot, not all of the points fall within the area between the two dashed lines, the assumption is not met. The residual is not normally distributed.
- From the “Y-hat vs Residuals” graph, we can observe that there is kind of an open-funnel shape formed, which means the homoskedasticity is not satisfied.
- The data points gather around the dotted line, which indicates that the independence is satisfied.
- The linearity of the regression is not satisfied because the trend line between the observed and fitted values is horizontal, and the data scattered randomly around the line.

3. The model's ability to fit the data

Code:

```
library(AER)
data <- read.csv("C:\\Users\\galon\\OneDrive\\Escritorio\\College Stuff\\Year2\\IE330\\Project2\\project2.csv")
state <- data[,1]
mortalityRate <- data[,2]
income <- data[,3]
averageAge <- data[,4]
unemployment <- data[,5]
ratioPH <- as.numeric(data[,6])
population <- as.numeric(data[,7])
uninsured <- data[,8]

model <- lm(mortalityRate~income + averageAge + unemployment + ratioPH +population + uninsured)
summary(model)
```

Output:

```
> model <- lm(mortalityRate~income + averageAge + unemployment + ratioPH +population + uninsured)
> summary(model)

Call:
lm(formula = mortalityRate ~ income + averageAge + unemployment +
    ratioPH + population + uninsured)

Residuals:
      Min        1Q    Median        3Q       Max
-0.041609 -0.013293 -0.001651  0.009608  0.066197

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.145e-01  9.514e-02   3.306 0.002076 **
income       -2.125e-06  5.289e-07  -4.018 0.000268 ***
averageAge   -4.627e-03  1.899e-03  -2.437 0.019626 *
unemployment  1.081e-03  1.829e-03   0.591 0.557817
ratioPH       7.216e-07  4.862e-07   1.484 0.146007
population    1.657e-04  2.390e-05   6.932 3.04e-08 ***
uninsured    -1.085e-03  1.525e-03  -0.711 0.481159
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02385 on 38 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6075,    Adjusted R-squared:  0.5455
F-statistic: 9.803 on 6 and 38 DF,  p-value: 1.585e-06
```

From the data above, we will be focusing more specifically on this exact output line:

```
Multiple R-squared:  0.6075,    Adjusted R-squared:  0.5455
```

In order to check whether the model fits the data well, we should look at the R-squared value from the summary of the data. If the value is greater or equal to 0.85, then that means that the data fits the model well, on the other hand, if it's less than that value, it means the model doesn't fit data well. In our case, the model does not fit the data well because the adjusted R-squared is 0.5455 which is less than 0.85.

4. Takeaway message from our model

For this project, I used multiple predictors in different states of the USA, and their relation to the amount of Covid deaths, this were the overall results:

I found the model to be significant because the p-value associated with our F-Statistic is much less than 0.05. According to the graphs, the regression model is only valid for only one over four assumptions, which is independence. The normality, homoscedasticity, and the linearity assumptions of the model are violated. In terms of how well this regression fits the data it does not fit the model very well. This is shown by the very low adjusted R-squared value of 0.5455. The closer the R-squared value is to 1 the better an estimate of the data is created. Generally if a model has an R-squared value of 0.85 it is said to be an accurate estimate of the data. To sum up, the data do not have any precise relationship, which can be due to the fact that different predictors were collected from distinguished sources.

References

U.S. COVID-19 death rate by state | Statista. (2020). Retrieved 27 October 2020, from https://www.statista.com/statistics/1109011/coronavirus-covid19-death-rates-us-by-state/

Herman, Z. (n.d.). State Unemployment Rates: September 2020. Retrieved October 27, 2020, from https://www.ncsl.org/research/labor-and-employment/state-unemployment-update.aspx

List of States by Population Density. (n.d.). Retrieved October 31, 2020, from https://state.1keydata.com/state-population-density.php

Median Age by State 2020. (2020). Retrieved October 31, 2020, from https://worldpopulationreview.com/state-rankings/median-age-by-state

Median Household Income by State 2020. (2020). Retrieved October 31, 2020, from https://worldpopulationreview.com/state-rankings/median-household-income-by-state

Health Insurance Coverage of the Total Population. (2020, October 23). Retrieved October 31, 2020, from https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0

Number of hospitals and hospital employment in each state in 2019. (2020, April 06). Retrieved October 31, 2020, from https://www.bls.gov/opub/ted/2020/number-of-hospitals-and-hospital-employment-in-each-state-in-2019.htm