Estimating The Number of COVID-19 Deaths

Introduction

The goal in this project was to create a multiple linear regression model to estimate the number of COVID-19 deaths in the summer by state. To generate this model I gathered the Data.csv dataset which contains a row for each of the 46 states I could find data on, and columns for statistics on each state such as the number of hospital beds, the median salary, total deaths in the spring, etc. Links to the resources I used to compile this dataset can be found in the Resources section at the bottom of this report. The goal ofthe model was to predict the summer_tot_deaths column which was the total number of deaths in the summer due to COVID 19. Below is a list of the potential covariates used.

Covariates Used

Variable	Description
num_poverty	The number of people below the federal poverty threshold in each state
num_hosp_beds	The number of hospital beds available in each state.
num_lungs_death	The number of lung related deaths in 2019
Pop_ abv_65	The number of people in each state older than 65
media_salary	The median_salary in each state
population	The population of each state
pop_density	The population density for each state
recovered	The total people that recovered by the beginning of summer
Spring_death	The total deaths during spring. This helps with adjusting the data since not all states got their infection at the same time.

Model Creation

In order to select the best possible model, I used the both the stepwise, and best subsets approaches.

Stepwise Regression

```
#model without variable selection
model1 <- lm(summer_tot_deaths~., data = specificregoin)

#stepwise variable selection
summary(step(model1, direction="both", scope=~.))
```

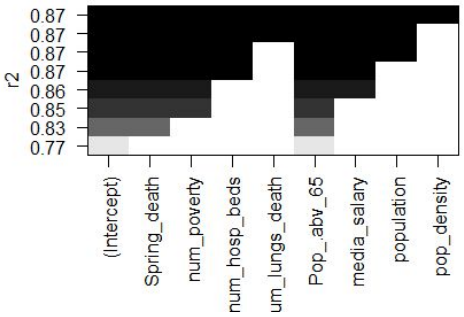
```
## Call:
## lm(formula = summer_tot_deaths ~ Spring_death + num_poverty +
##     num_hosp_beds + Pop_.abv_65 + media_salary, data = specificregoin)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1238.9  -406.1  -179.4   297.1  2249.3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.454e+03  8.140e+02  -1.786 0.082005 .
## Spring_death   1.627e-01  4.538e-02   3.586 0.000943 ***
## num_poverty   -3.209e-04  1.208e-04  -2.657 0.011478 *
## num_hosp_beds  1.277e-02  8.356e-03   1.528 0.134867
## Pop_.abv_65    1.401e+00  1.195e-01  11.730 3.39e-14 ***
## media_salary   1.650e-02  8.464e-03   1.950 0.058616 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 757.1 on 38 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8702, Adjusted R-squared:  0.8531
## F-statistic: 50.94 on 5 and 38 DF,  p-value: 8.029e-16
```

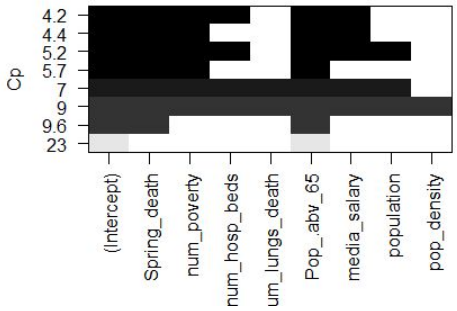
Based on this method the selected variables will be Spring_death, num_poverty, num_hosp_beds, pop_.abv_65, and media_salary. Next I will try the best subsets approach.

Best Subsets Regression

```
#best subsets variable selection
best_subsets <- regsubsets(summer_tot_deaths~., data = specificregoin)
plot(best_subsets, scale = "r2")

plot(best_subsets, scale = "Cp")
```





Judging by the plots shown above, the model with 6-variables and 5-variable has the highest R-squared and model with 5- variables as Ill has the highest Cp.

Best Subsets Options

Type	R-Square	Mallows's Cp
6-variable	0.87	5.7
5-variable	0.87	4.2
4-variable	0.86	4.4

The 6-variable model has a high R-squared and high Cp. The 4-variable model has a loIr Cp than 6-variable but also has loIr R-squared. The 5-variable has the highest R-squared and the loIst Mallows's Cp and is therefore the best model betIen the options

Final Variable Selection

Regression subset and stepwise regression both favored the 5-variable model.the final model includes the variable Spring_death, num_poverty, num_hosp_beds, Pop_.avb_65, and media_salary.

Final Model Generation and Summary

```
#Selected model
model1 <- lm(formula = summer_tot_deaths ~ Spring_death + num_poverty +
               num_hosp_beds + Pop_.abv_65 + media_salary, data = specificregoin)

#summary
summary(model1)

##
## Call:
## lm(formula = summer_tot_deaths ~ Spring_death + num_poverty +
##     num_hosp_beds + Pop_.abv_65 + media_salary, data = specificregoin)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1238.9  -406.1  -179.4   297.1  2249.3
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.454e+03  8.140e+02  -1.786 0.082005 .
## Spring_death   1.627e-01  4.538e-02   3.586 0.000943 ***
## num_poverty   -3.209e-04  1.208e-04  -2.657 0.011478 *
## num_hosp_beds  1.277e-02  8.356e-03   1.528 0.134867
## Pop_.abv_65    1.401e+00  1.195e-01  11.730 3.39e-14 ***
## media_salary   1.650e-02  8.464e-03   1.950 0.058616 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 757.1 on 38 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8702, Adjusted R-squared:  0.8531
## F-statistic: 50.94 on 5 and 38 DF,  p-value: 8.029e-16
```

Regression significance

According to the F-test, the overall regression model is significant with a p-value of 8.029e-16. This is less than alpha = 0.05 and leads us to reject the null hypothesis that the fit of the intercept only model and predictor model are equal. Spring deaths, the population in poverty, and the population above 65 Ire found significant at an alpha = 0.5 using t-tests. The low p-values means I reject the null hypothesis that the coefficients of these predictor variables are equal to zero, and therefore changes in the predictor variables are related to the changes in the response variable (COVID-19 deaths in summer). The other variables had p-values over 0.05. A 95% confidence interval on the coefficients can be found in the table below.

95% Confidence Interval on Coefficients

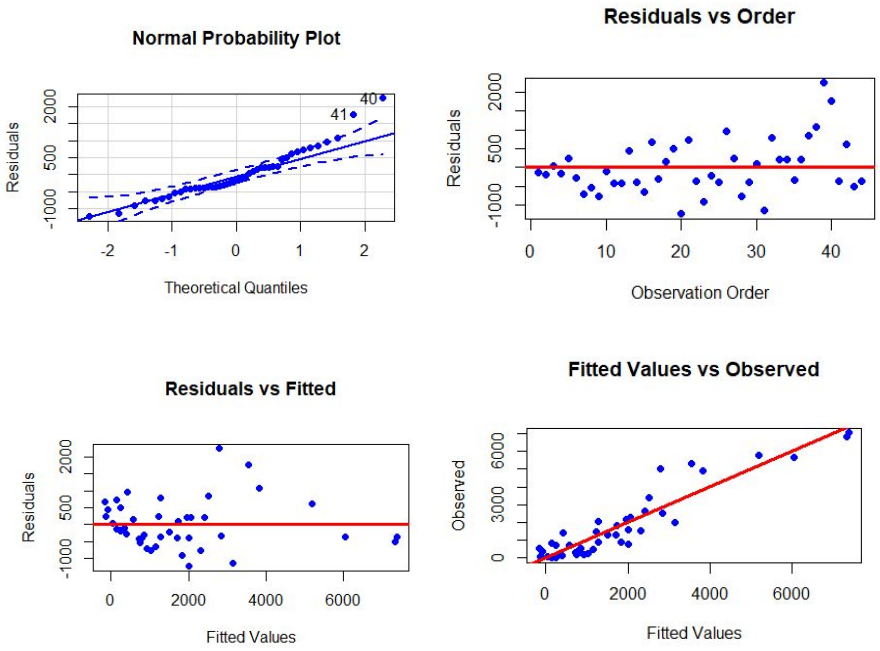
Coefficient	LoIr Bound	Upper Bound
(Intercept)	-3.10E+03	1.94E+02
Spring_death	7.09E-02	2.55E-01

num_poverty	-5.65E-04	-7.64E-05
num_hosp_beds	-4.15E-03	2.97E-02
Pop_.abv_65	1.16E+00	1.64E+00
media_salary	-6.32E-04	3.36E-02

Model validity?

To check the validity of the model, I ran a series of tests on the residuals of the model in R. The output is shown below.

Model Diagnostic Plots



Analysis of Diagnostic Plots

- **Normality:** The Normal Probability plot shows that most data points stay within the 95% CI. There are a few points outside of the CI interval, but not enough to invalidate the entire model. This upholds the normality assumption.
  Fix: Take out the outliers.
- **Independence of residuals:** There is a slight pattern across the x-axis of the Residuals vs Order plot, indicating the independence of residuals assumption is invalid.
  Fix: Reduce the collinearity by removing variables that are strongly related to one another thus creating more independent residuals.
- **Linearity and Homoscedasticity:** The Residuals vs. Fitted graph seems to have an N-shape pattern signalling a higher order relationship that has not been accounted for. This violates the linearity assumption. In addition, a funnel shape is present on the Residuals vs. Fitted plot which indicates a violation of homoscedasticity.

Fix: Transform the data using a log, exponential, or other function.

Model Goodness of Fit

The adjusted R-squared value of the model is 0.8531 which shows a good fit as it is above 0.85. In addition, the difference bettween the R-squared (0.8702), and the adjusted R-squared is small thus showing minimal to no overfitting in the data.

Take Away

Equation:
(0.1627*Spring_death) - (0.0003209*num_poverty) + (0.01277*num_hosp_beds) + (1.401*Pop_.abv_65) + (media_salary * 0.0165) -1454

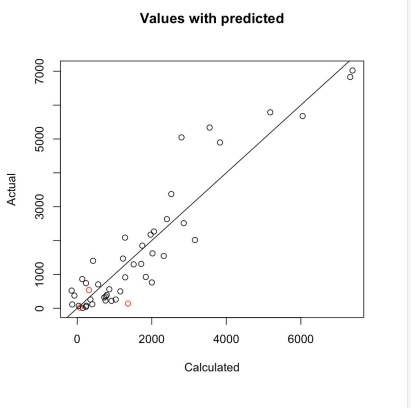
Multicollinearity and Impact on inference:

```
> vif(model1)
 Spring_death   num_poverty num_hosp_beds   Pop_.abv_65  media_salary
     1.239035      1.234183      1.051372      1.454807      1.068912
```

All multicollinearity variables are below 10. This points to low inference which means the predictors can be trusted.

Testing the model

I left out 3 of the states when I created the model. I can now testthe model on this set of data the model has not been trained on. The predicted (calculated) values are displayed as a line, and the actual values of the 3 new states are plotted in red.



I can see thatthe model's prediction is close to what the actual values are; this further validates the model.

I foundthe model was significant, and a good predictor of the COVID-19 related deaths in summer. This shows that the number of COVID-19 related deaths in a given state during the summer can be accurately predicted given the number of deaths in Spring in the state, the number of people in poverty in the state, the quantity of available hospital beds in the state, the population above 65 in the state, and the median salary within the state. Furthermore, due to the low VIF values, I can interpret the coefficients of the model to determine a negative or positive correlation with the COVID-19 related deaths in summer. HoIver, because many of the coefficients have both positive and negative values for their coefficients within the 95% CI, this does not give much insight. Overall,the model is a good predictor of the number of COVID-19 related deaths for a state in summer.