Covid Deaths Analysis and Prediction

Peter Chu

6/6/2022

Abstract

This project was done on COVID-19 was done as my mother is a registered nurse. She has had this job for the past 32 years and had to work 60 hours a week during the height of the pandemic. I did this in honor of her risking her life to help others as she has done her whole life. The observed data composes of 2 categorical and 2 numeric variables which were all randomly sampled from the California Government database on COVID-19 statistics. The data was assessed using ANOVA tests, linear regression, and several t-tests to see if there was any correlation between Covid-19 deaths and the year, the number of cases in a county, and the county area. It was found that none of these variables of interest are strongly correlated enough thus there may be other good variables such as a person's age and health but this was not analyzed in this project.

Introduction

The dataset was obtained from a Government Operations Agency sponsored project which collects open data from multiple agencies and presents it in a readable format with no personal details included. It only includes data for California. This dataset focuses on Statewide COVID-19 cases, deaths, and tests. For this project the dataset was limited to the large cities of Los Angeles, San Diego, and San Francisco, which respectively have the largest, 2nd largest, and 4th largest city populations.

The main goal for this analysis is to find out whether the deaths from COVID-19 is uniform in certain counties and year, and if we can predict the number of deaths based on some variables. However, it should be noted that some changes in deaths may not be directly related to the COVID-19 pandemic, but rather due to interventions (CDC1 2022). For example social distancing has been shown to reduce the spread of COVID-19, but was not a direct result that came from the virus itself. The CDC and government had intervened and placed many restrictions related to social distancing and masking that could affect the number of deaths. There are many other variables that are not covered in this dataset that could also help predict the number of deaths. Clinical conditions, such as heart disease, cancer, diabetes, have also been shown to have a positive correlation with COVID-19 deaths, but again these variables are not in the dataset (Esai 2022). Some of these variables are clear indicators such as those with and without cancer. According to a research paper, there is a significant increase in death risk with COVID patients who have cancer (Lunski 2020). Overall there are many factors that play a role into predicting a COVID-19 death that there starts to become trade-offs with how many variables we use and goodness of model fit. Thus we will stick to our 4 variables of interest: Cases and Deaths in a single day, the Year the data point was recorded in, and the County where the data point was recorded in.

Exploratory Analysis

The dataset is very large with nearly 52,100 observations for 18 variables. Thus in order to achieve the tidied dataset we first created a new variable to show which year the data was recorded in. Then we subset the variables and areas of interest before removing all missing values. From our histogram of Deaths it does not appear normal, but due having 2,559 observation Central Limit Theorem applies so it is normal. The same applies to the variable Cases. Our scatterplot of Deaths vs Cases shows that there is some correlation between the two and appears to be going in two directions as well. Our boxplots show that there are a lot of outliers and from our boxplots without outliers there is indication that the mean deaths for each year is possibly the same, but not for each county. (Appendix 1 & 2). Our correlation matrix shows that there is some correlation among the variables and linear relationship between them as well.

Figure 1

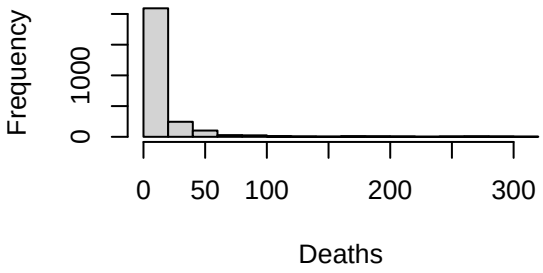


Figure 2

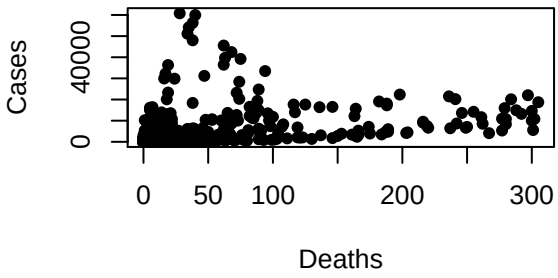


Figure 3

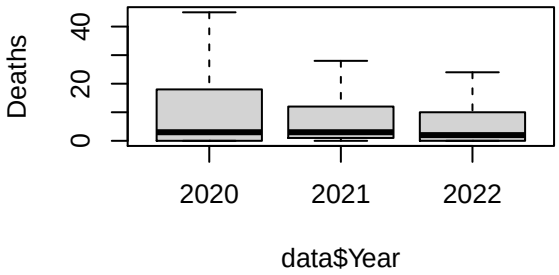


Figure 4

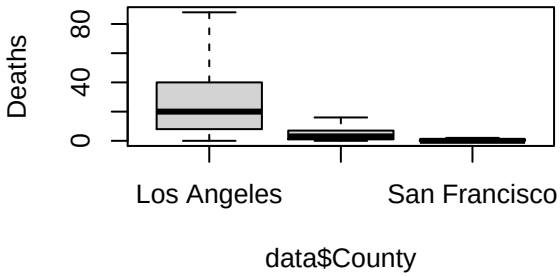
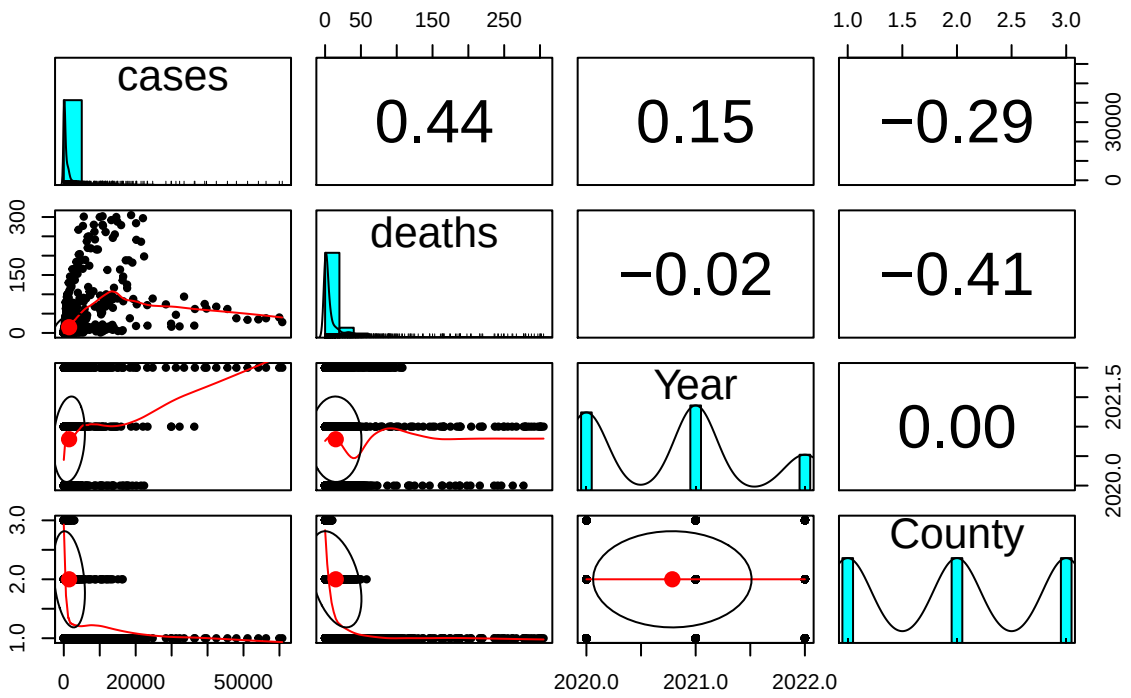


Figure 5



Statistical Methods

Notes: For our regression analysis we will have our predictors be Year, County, and Cases and test if they predict Deaths. $\alpha = 0.05$ *for all tests*

For this project we will use 3 statistical methods, t-test, one-way ANOVA and multiple linear regression.

t-test

a t-test compares the means of two groups. Since our two samples come from the same population and time we will use a paired t-test. The null hypothesis for a paired t-test is that $H_0 : \mu_d = 0$ and the alternative hypothesis is $\mu_d \neq 0$ where mu_d is the average difference between the populations. The assumptions of the paired t-test are that

1: The samples are the same size 2: The samples are dependent 3: There is a one-to-one correspondence between the two samples 4: The sample is random 5: the data is normal 6: The variance is homogeneous

For our analysis we know the sample is random as all observations were equally as likely to be collected, the residuals are normal by the Central Limit Theorem, and from our "residuals vs fitted plot", shown later in the analysis, since the data points seem to follow no pattern and look similar in terms of width we can conclude that our variances are homogeneous.

One-way ANOVA

ANOVA (Analysis of Variance) is the statistical test that tests whether at least one mean from a set of 3 more samples is different from the others. It also tests whether the within group or among group variance is larger. The null hypothesis for a one-way ANOVA is $H_0 : \mu_i = \mu_j$ and the alternative hypothesis is $H_A : \mu_i \neq \mu_j$ where $i \neq j$. There are three assumptions in order to perform the one-way ANOVA.

1: The sample must be a random sample. 2: The residuals must be normally distributed. 3: Equality of variance

The assumptions for our one-way ANOVA are already satisfied and explained the the **t-test** section.

Linear Regression

Linear regression has two models, the simple and multiple regression. For this analysis we will use multiple regression with the model equation $Y = \beta_0 + \beta_i X_i$ for $i = 1, 2, 3, ...n$. Linear regression is used to show which variables of interest can predict another one or in a sense find a "relationship" between these variables. The null hypothesis is $H_0 : \beta_i = 0$ for $i = 1, 2, 3, ...n$ and alternative hypothesis $H_A : \beta_i \neq 0$ for $i = 1, 2, 3, ...n$.

There are two things to consider when comparing multiple models: the goodness of fit, which describes how well the model fits the data, and complexity, which describes how many parameters are in the model. In general the more parameters (variables of interest) in the model, the more complex it is, but usually the less goodness of fit. There is an inverse relationship between these two, but we want to find the model with a high goodness of fit and low complexity. Thus we shall only use a handful of predictors. For this analysis we will compare the models with AIC (Akike's Information Criterion), BIC (Bayes' Information Criterion), and the adjusted R^2 value rather than the R^2 value. A lower AIC and BIC with a high adjusted R^2 value is what we are looking for. The four assumptions of multiple linear regression are:

1: The samples are random 2: There is no collinearity between predictors 3: There is some linear relationship between predictors 4: The residuals are normal 5: The residuals have equal variance

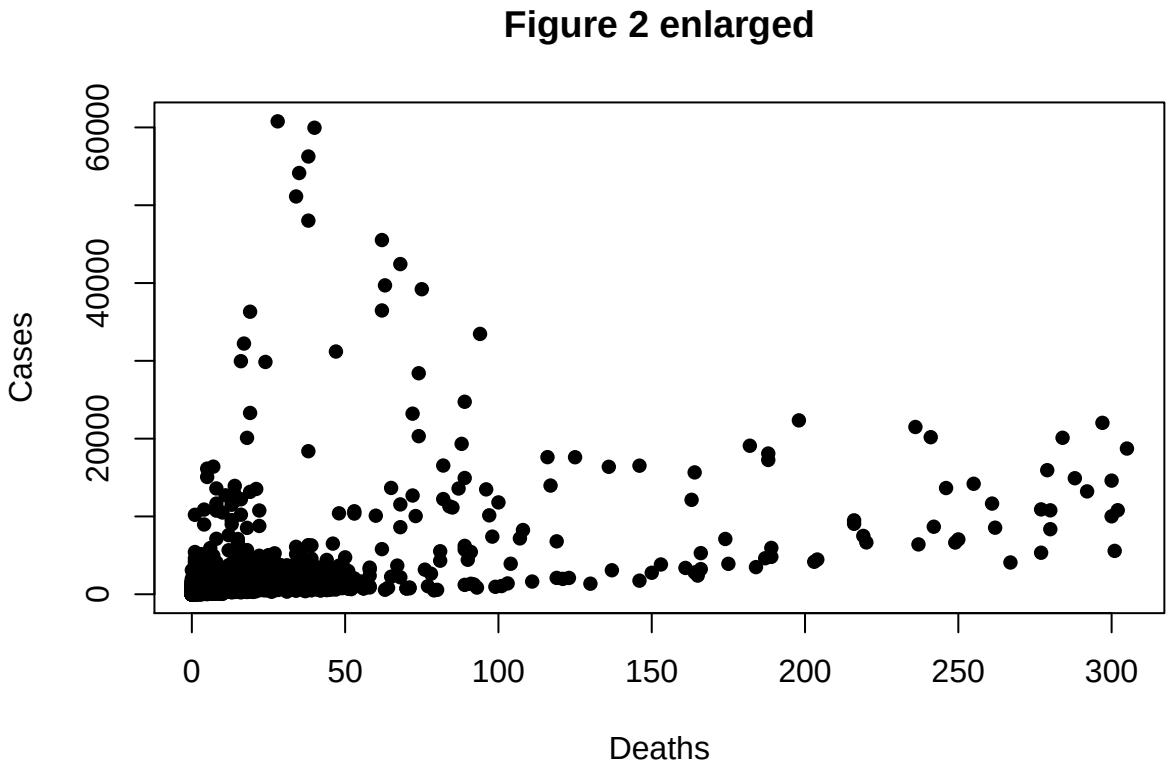
Our assumptions for 1, and 5 are already satisfied and explained in the **t-test** section.

Looking at our correlation matrix (Figure 5) we can see that there is no collinearity among our predictors Cases, Year, and County. In addition there is some linear relationship among predictors, so our assumptions for multiple linear regression are satisfied. From our shapiro-tests on the resdiuals of each model we get a w-value > 0.05, so the residuals are normal and the assumption is met (Appendix 6).

Results

Paired t-test: Deaths and Cases

For our paired t-test we will analyze whether the mean difference between these two variables is equal to 0. If it is then we can conclude that the average number of cases is similar to the average number of deaths. Thus we will run a paired t-test with H_0: the mean difference between deaths in a single day and cases in a single day is 0 and H_A : the mean difference between deaths in a single day and cases in a single day is not 0. Equivalently we can say $H_0 : \mu_d = 0$ and $H_A : \mu_d \neq 0$

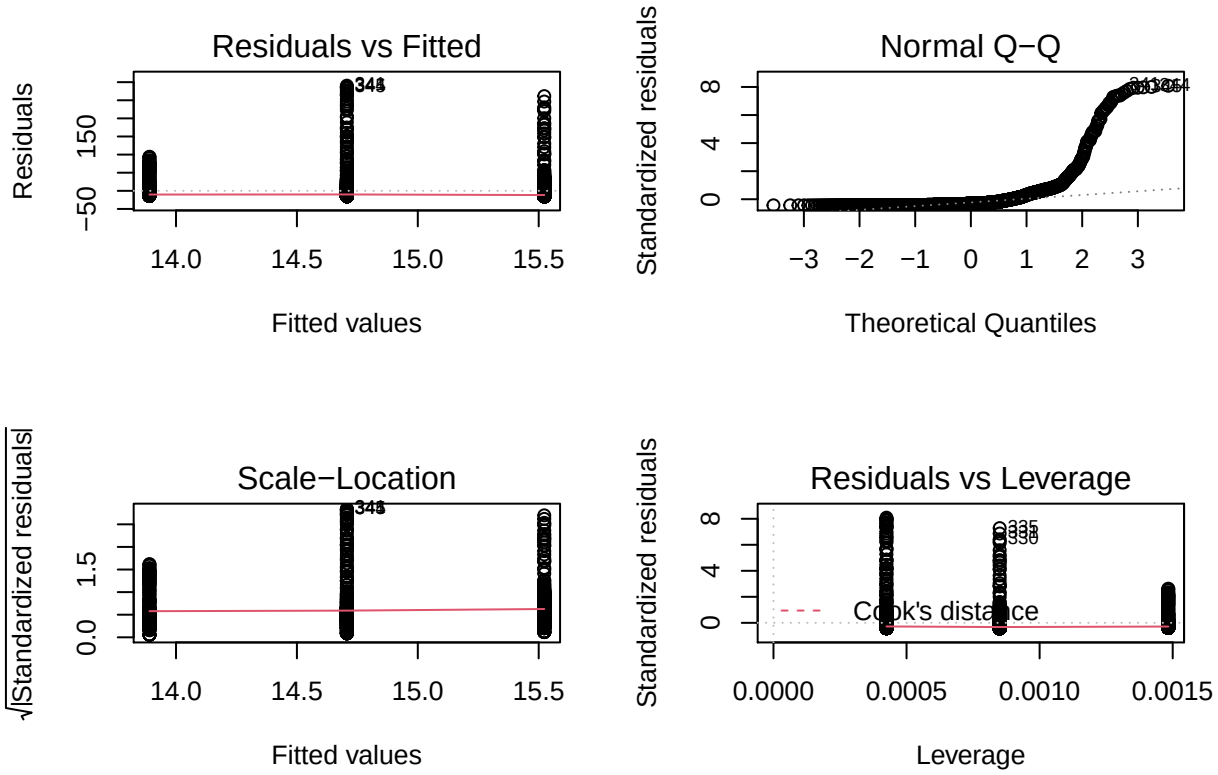


From our paired t-test we get a p-value $< 2.2e\text{-}16$ (Appendix 9). Thus we reject our null hypothesis and can conclude that there is a difference between the mean number of deaths and mean number of cases. Looking at our scatterplot we can also observe that there is a difference between cases and deaths is not equal to 0.

One-way ANOVA: Deaths versus Years and County

Death vs Years

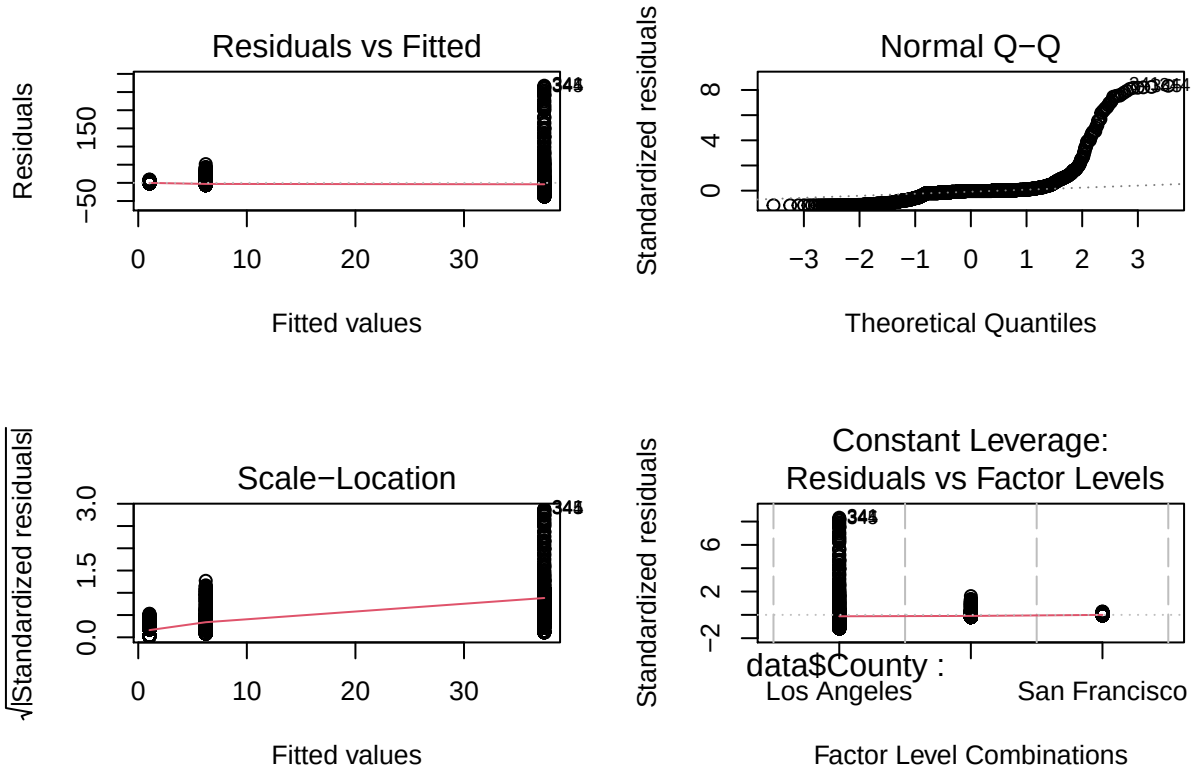
For our one-way ANOVA on Deaths and County, we will analyze whether the mean number of deaths in the same across all years. We are interested in these variables as differing number of populations and density of population could affect how fast COVID-19 spreads which leads to deaths. The Year is important as over time more and more of population became vaccinated. We will run a one way anova H_0: The mean deaths in a single day is the same for each year and the H_A: The mean deaths in a single day is not the same for each year. Equivalently we can say the $H_0 : \mu_{2020} = \mu_{2021} = \mu_{2022}$ and the $H_A : \mu_{2020} \neq \mu_{2021} \neq \mu_{2022}$ where μ is the average deaths on a single day.



Upon our initial observation, the means for deaths on a single day appears to be the same across all 3 years. However, running a one-way ANOVA on the variables Death and Year, we get a p-value of 0.404 (Appendix 3). Thus we can conclude that while the means may appear similar, they are not equal to each other and thus we reject the null hypothesis. Thus the average number of deaths in a single day is different for all three counties.

Deaths vs County

For our one-way ANOVA on Deaths and County, we will analyze whether the mean deaths on a single day the same across all counties. The counties in this dataset are Los Angeles, San Francisco, and San Diego. While all the cities have some of the highest populations in California, but are not equally dense. Thus we will run a one way ANOVA with the $H_0 : \mu_{LA} = \mu_{SD} = \mu_{SF}$ and $H_A : \mu_{LA} \neq \mu_{SD} \neq \mu_{SF}$.



Running our one-way ANOVA we can confirm our initial observation as we get an extremely small p-value less than 0.05 (Appendix 4). Thus we reject the null hypothesis and can confirm our initial observation that our average death on a single day is not the same for the counties.

Multiple linear regression: Predicitng number of deaths

Note: For our multiple linear regression there are 3 models.

For all of our models out p-values are less than 0.05. Thus we reject the null hypothesis and can conclude that for each of our models $\beta_i \neq 0$ for $i = 1, 2, 3, ...n$ (Appendix 5).

Table 1: Table of models' formulas

Model Name	Model Structure (y x)
Model 1	Deaths Year
Model 2	Deaths Year + County
Model 3	Deaths Year + County + Cases

Table 2: Table of models' degrees of freedom, AIC, BIC, R squared, and adjusted R squared values

	df	AIC	BIC	R Squared	Adjusted R squared
Model 1	3	25593.88	25611.42	0.00	0.3
Model 2	5	25026.46	25055.70	0.20	0.3
Model 3	6	24668.62	24703.71	0.31	0.3

As we can see from the table our third model is the best model as it has the smallest AIC and BIC values and the highest R^2 value. Thus we will use this model to fit our multivariate linear model (Appendix 7).

Best Multivariate Linear Model: Model 3

Table 3: Model 3 parameter values

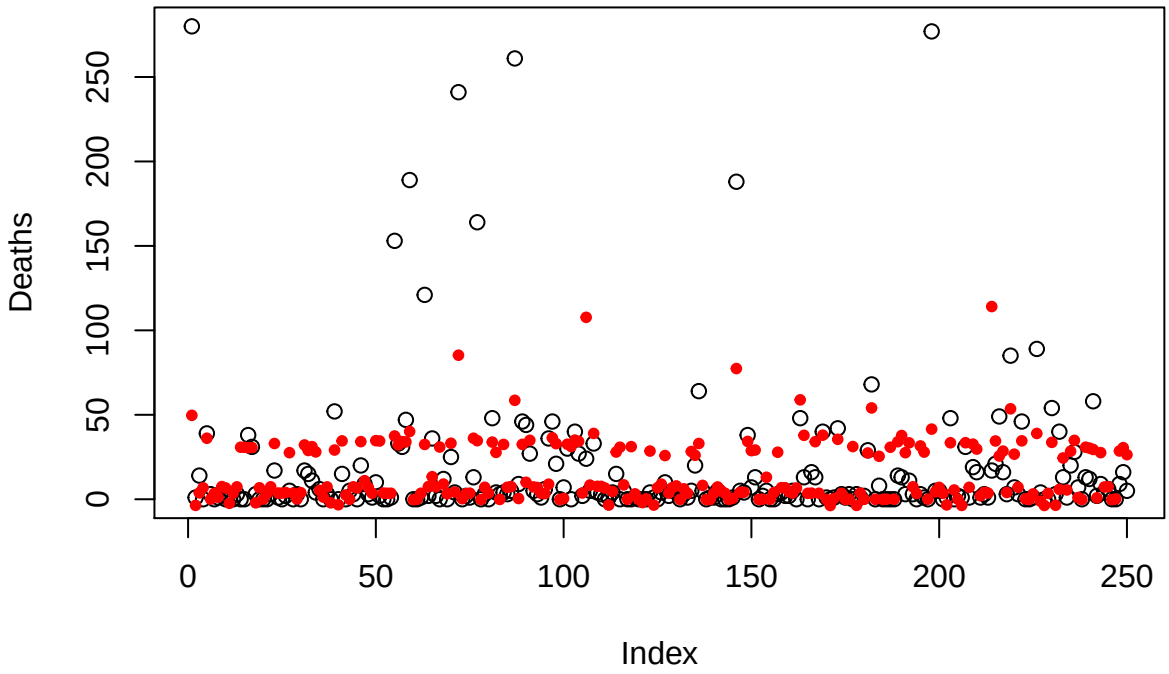
Component	Estimate	Standard Error	t-value	p-value
Intercept	7.021e+03	1.671e+03	4.201	2.75e-05
Year	-3.460e+00	8.270e-01	-4.184	2.96e-05
County (San Diego)	2.441e+01	1.491e+00	-16.377	<2e-16
County (San Francisco)	-2.746e+01	1.520e+00	-18.071	<2e-16
Cases	2.811e-03	1.431e-04	19.637	<2e-16

From the table we can now fit our model $Y = \beta_0 + \beta_i X_i + \epsilon_i$ with these values and thus our model becomes $Death = 7021 + -3.4600 Year + 24.41 SanDiego + -27.46 SanFrancisco + 0.002811 Cases$ where Deaths is the number of deaths in a single day, cases are the number of cases in a single day, Year is 2020 to 2022, and San Diego and San Francisco are binary variables. For example, if the county is Los Angeles then San Diego and San Francisco = 0. Thus we then have the model $Death = 7021 + -3.4600 Year + 24.41 * 0 + -27.46 * 0 + 0.002811 Cases = 7021 + -3.4600 Year + 0.002811 Cases$

Prediction of best multivarite linear model: Model 3

We then take the 2559 observations and then 250 observations are used to test the fitted model.The red dots are predicted values and the unfilled circles are the actual values.

Figure 7



As we can see most of the red dots either fill or are close to the unfilled circles with the exception of a handful of outliers. The plot shows that the predicted values are close to the actual values, but can not closely predict any outliers. Thus the model is still decent at predicing the number of deaths in a single day. In general, since the Year and County are a given, the number of cases is the only variable with a lot of variability. However, it is useful as this is the only variable that can cause large changes in the predicted value

Discussion

Real World Application

From our analysis we concluded that the average number of deaths was not the same for each year across the counties, but was very close. Thus researchers may be interested in discovering underlying reasons for why this may be despite the introduction of vaccines and widespread inoculation. Researchers could also explore if these underlying reasons affect the predictions of deaths.

Limitation

According to the CDC (CDC1 2022), more than 81% of COVID-19 deaths occurred to individuals who were over the age of 65. This dataset did not include the age or health of an individual which could be tested as potential predictors. In addition, there was simply too much data to go through each county in California. In the raw dataset there were 52094 observations and 18 variables. Thus out analysis is limited to only the three counties of Los Angeles, San Diego, and San Francisco. Another limitation is that we know how many people were vaccinated, but not which vaccination they got. There are differences in effectiveness of each

vaccine and this could have been analyzed if the data was included. Overall, with the data given, we were able to do a very general analysis on the three counties, but could have gone much more in depth given more data.

Final Take-away

From the analysis, I can confidently say that we have only scratched the surface on predicting deaths due to COVID-19. There is a lot of factors, such as which vaccine people got, that could be used to create a better model. There is missing important information such as the quality of hospitals treatments to the health of individuals that could all be analyzed. However, with our limited data I can conclude that the year and county affects the number of deaths. A small town of 1000 people can have drastically different results than a large town with 100,000 people. There is not set equation that can be transferred from one county to another to predict these deaths.

References

Hao Zhu (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra

José Pinheiro, Douglas Bates and R Core Team (2022). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-157. https://CRAN.R-project.org/package=nlme

Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.37.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Revelle, W. (2022) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.2.5.

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Kirill Müller (2020). here: A Simpler Way to Find Your Files. R package version 1.0.1. https://CRAN.R-project.org/package=here

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Hadley Wickham, Jim Hester and Jennifer Bryan (2022). readr: Read Rectangular Text Data. R package version 2.1.2. https://CRAN.R-project.org/package=readr

CDC1 "People with Certain Medical Conditions." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 2 May 2022, https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html.

Esai Selvan, Myvizhi. "Risk Factors for Death from Covid-19." Nature News, Nature Publishing Group, 27 May 2020, https://www.nature.com/articles/s41577-020-0351-0.

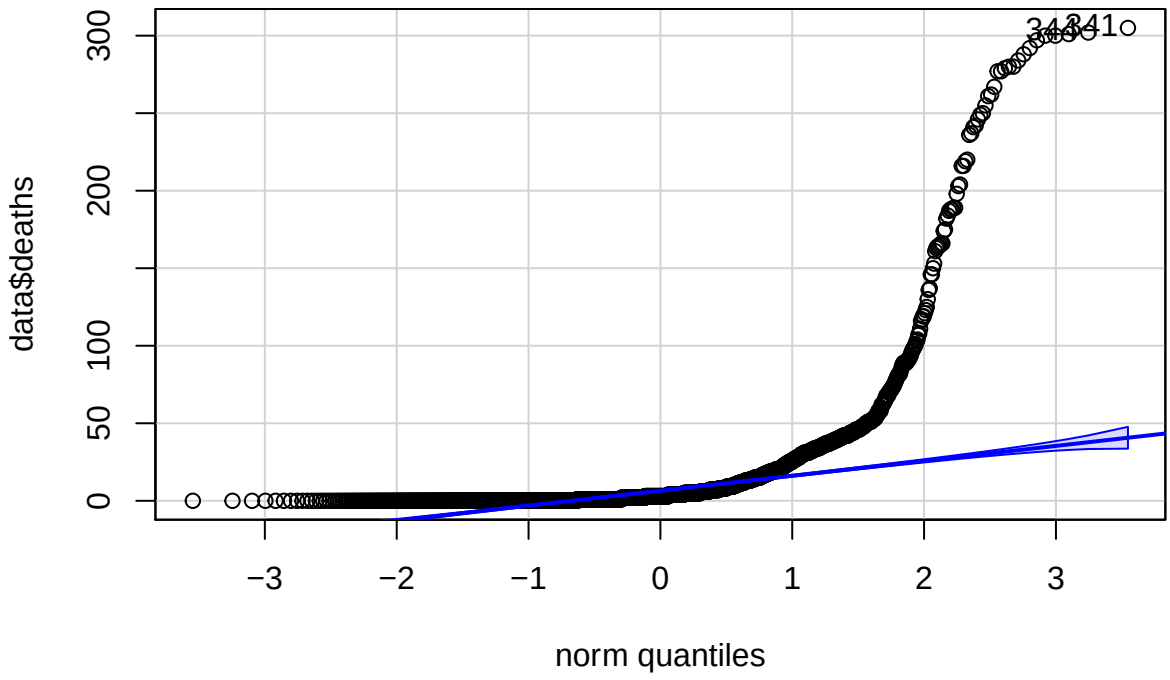
Lunski, Michael J., et al. "Multivariate Mortality Analyses in COVID-19: Comparing Patients with Cancer and Patients without Cancer in Louisiana." Cancer, vol. 127, no. 2, 2020, pp. 266–274., https://doi.org/10.1002/cncr.33243.

Appendix

```
#Shapiro test for Deaths
shapiro.test(data$deaths)
```

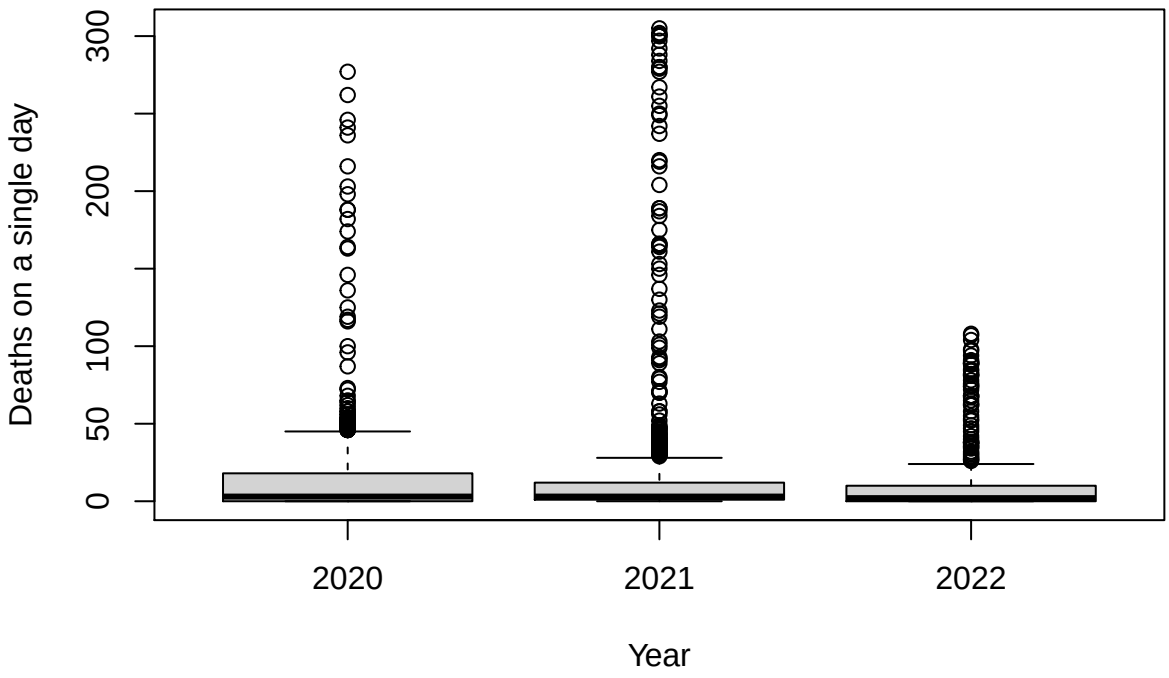
```
##
##  Shapiro-Wilk normality test
##
## data:  data$deaths
## W = 0.42299, p-value < 2.2e-16
```

```
# qq plot of Deaths
qqPlot(data$deaths)
```

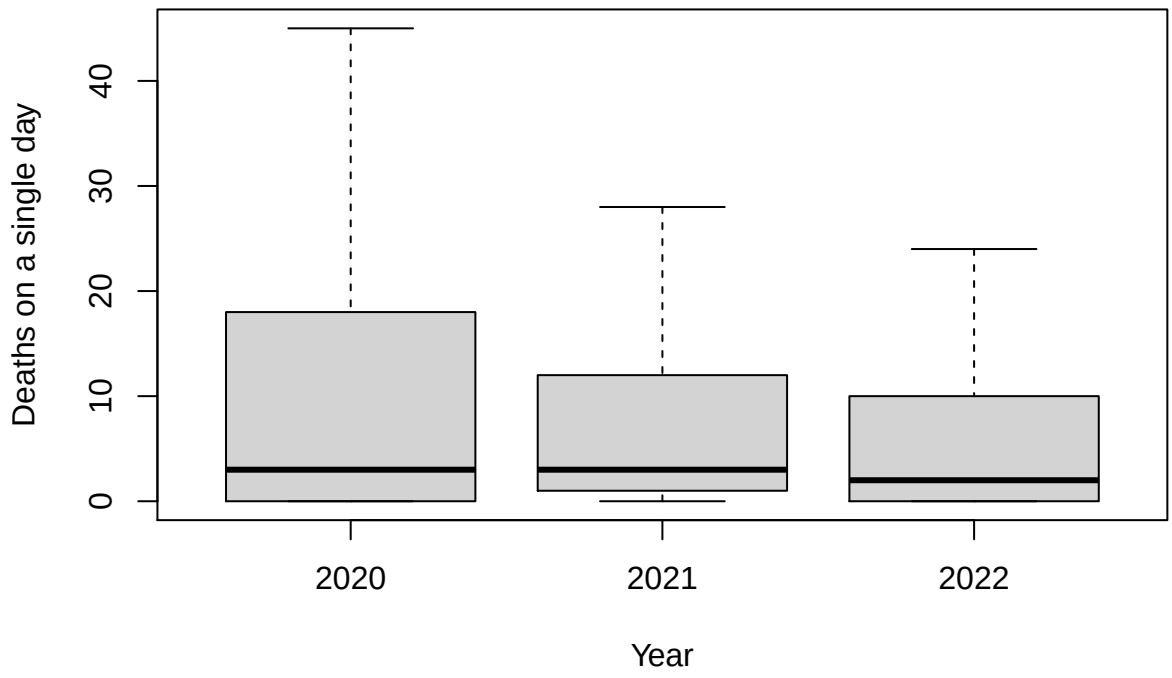



```
## [1] 341 344
```

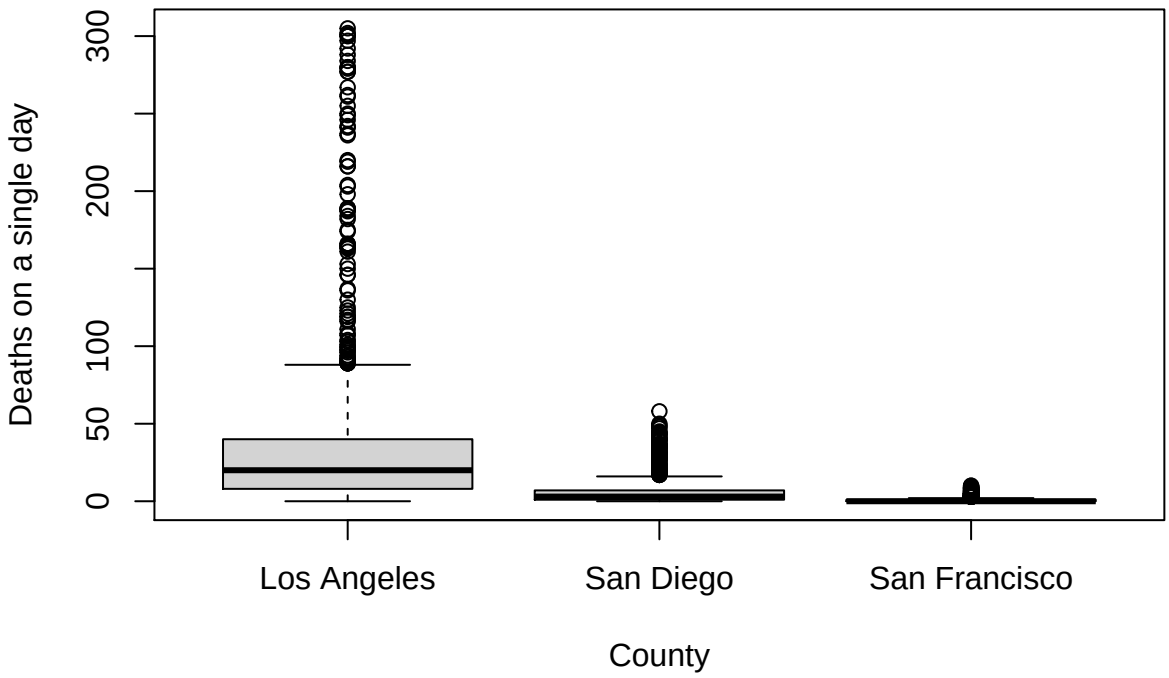
```
#Boxplots of Deaths and Years. (appendix 1)
boxplot(data$deaths~data$Year, xlab = 'Year', ylab = 'Deaths on a single day')
```



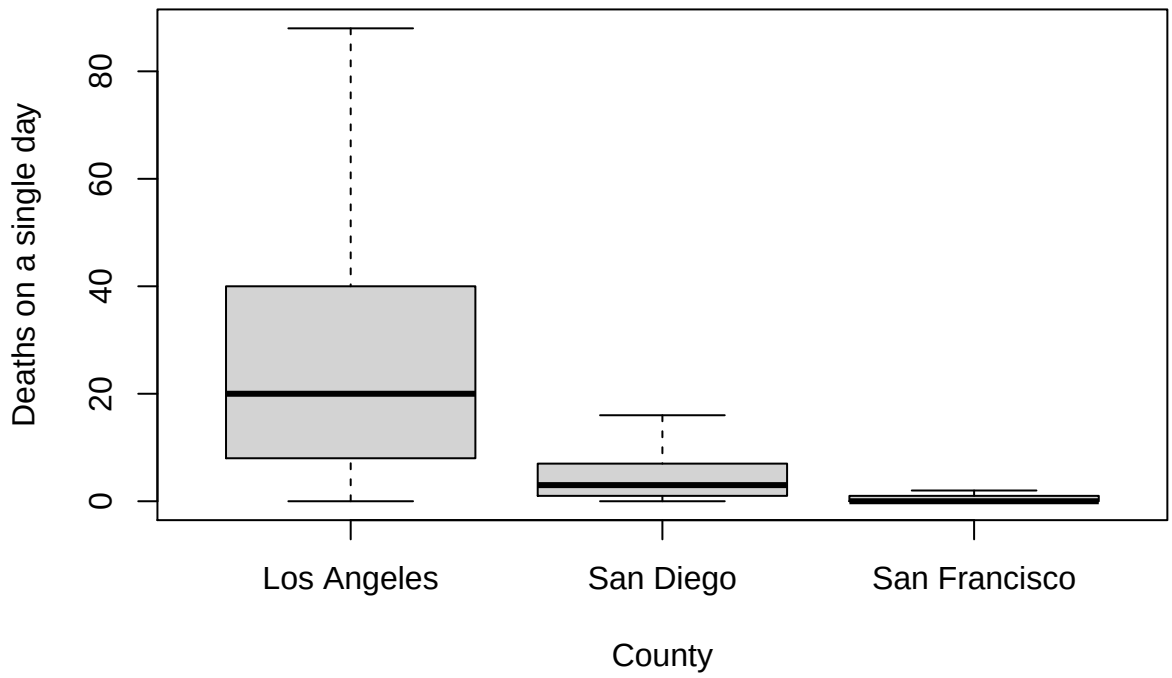

```
boxplot(data$deaths~data$Year, xlab = 'Year', ylab = 'Deaths on a single day', outline = FALSE)
```



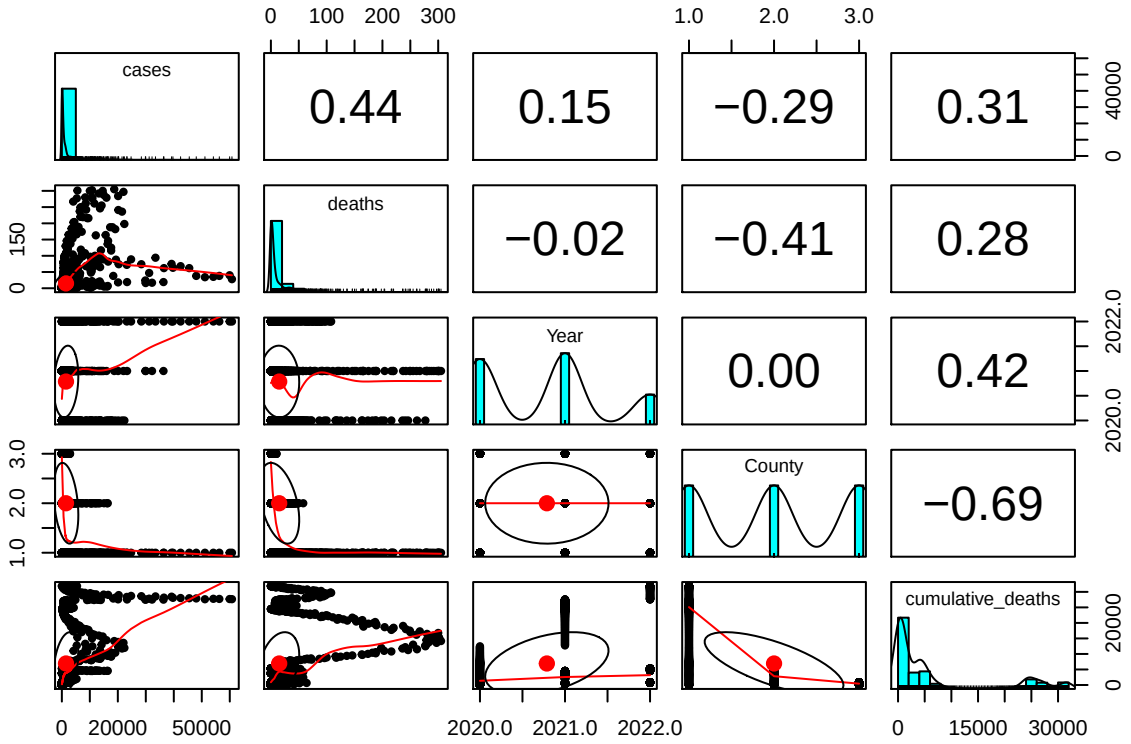

```
#Boxplots of Deaths and Years. (appendix 2)
boxplot(data$deaths~data$County, xlab = 'County', ylab = 'Deaths on a single day')
```




```
boxplot(data$deaths~data$County, xlab = 'County', ylab = 'Deaths on a single day', outline = FALSE)
```




```
#Correlation matrix
pairs.panels(data)
```




```
#one way ANOVA on Deaths and Years (appendix 3)
deathsYearAov <- aov(data$deaths~data$Year)

summary(deathsYearAov)
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## data$Year      1     898   898.1   0.696  0.404
## Residuals   2557 3297469  1289.6
```

```
#one way ANOVA on Deaths and County (appendix 4)
deathsCountyAov <- aov(data$deaths~data$County)
summary(deathsCountyAov)
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## data$County    2  659895  329948   319.6 <2e-16 ***
## Residuals   2556 2638471    1032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Models for multiple regression and their p-values (appendix 5)
model1 <- lm(data$deaths~data$Year)
model2 <- lm(data$deaths~data$Year+data$County)
model3 <- lm(data$deaths~data$Year+data$County+data$cases)

summary(model1)
```

```
##
## Call:
## lm(formula = data$deaths ~ data$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.523 -14.523 -11.707  -1.707 290.293
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1664.6132  1976.8813   0.842    0.400
## data$Year     -0.8164     0.9783  -0.835    0.404
##
## Residual standard error: 35.91 on 2557 degrees of freedom
## Multiple R-squared:  0.0002723,  Adjusted R-squared:  -0.0001187
## F-statistic: 0.6964 on 1 and 2557 DF,  p-value: 0.4041
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = data$deaths ~ data$Year + data$County)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.035  -6.043  -1.675   0.957 267.781
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1687.1252  1768.7339   0.954    0.340
## data$Year                  -0.8164     0.8753  -0.933    0.351
## data$CountySan Diego      -31.1758     1.5558 -20.039   <2e-16 ***
## data$CountySan Francisco  -36.3599     1.5558 -23.371   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.13 on 2555 degrees of freedom
## Multiple R-squared:  0.2003, Adjusted R-squared:  0.1994
## F-statistic: 213.4 on 3 and 2555 DF,  p-value: < 2.2e-16
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = data$deaths ~ data$Year + data$County + data$cases)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -166.659   -5.760   -2.073    1.769  258.067
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               7.021e+03  1.671e+03   4.201 2.75e-05 ***
```

```
## data$Year                -3.460e+00  8.270e-01  -4.184 2.96e-05 ***
## data$CountySan Diego     -2.441e+01  1.491e+00 -16.377  < 2e-16 ***
## data$CountySan Francisco -2.746e+01  1.520e+00 -18.071  < 2e-16 ***
## data$cases                2.811e-03  1.431e-04  19.637  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.95 on 2554 degrees of freedom
## Multiple R-squared:  0.3052, Adjusted R-squared:  0.3041
## F-statistic: 280.5 on 4 and 2554 DF,  p-value: < 2.2e-16
```

```
# checking residuals are norma. p-value < 0.05, but w > 0.05, since large data use w (appendix 6)
shapiro.test(model1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model1$residuals
## W = 0.42584, p-value < 2.2e-16
```

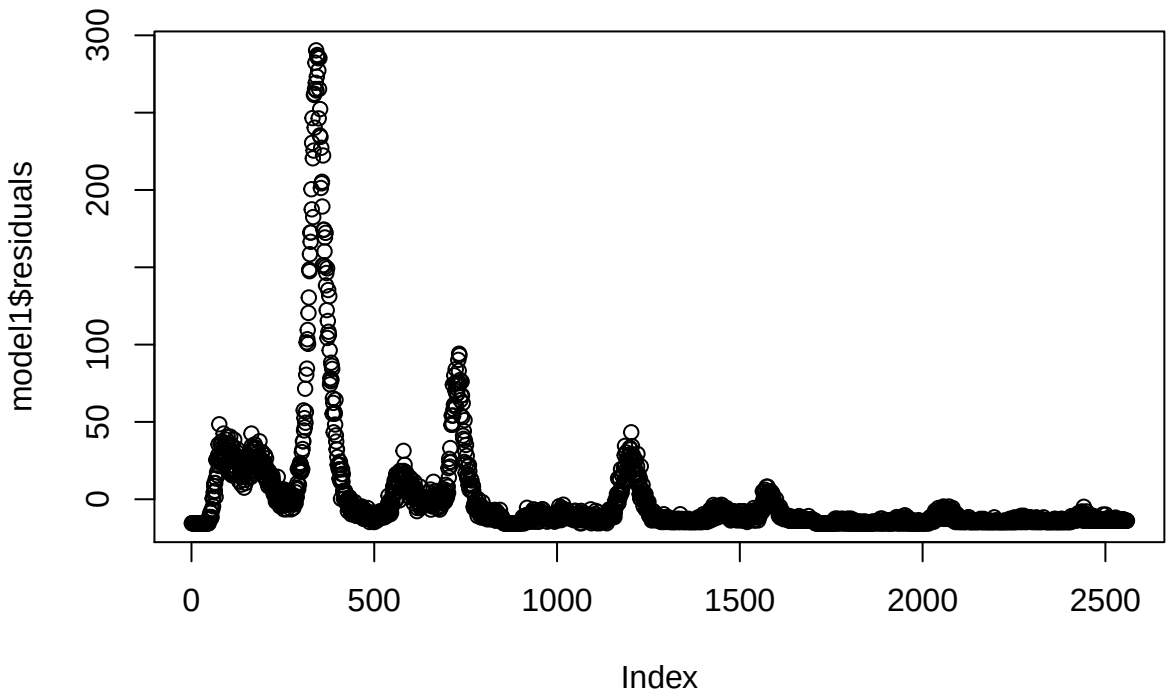
```
shapiro.test(model2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model2$residuals
## W = 0.50361, p-value < 2.2e-16
```

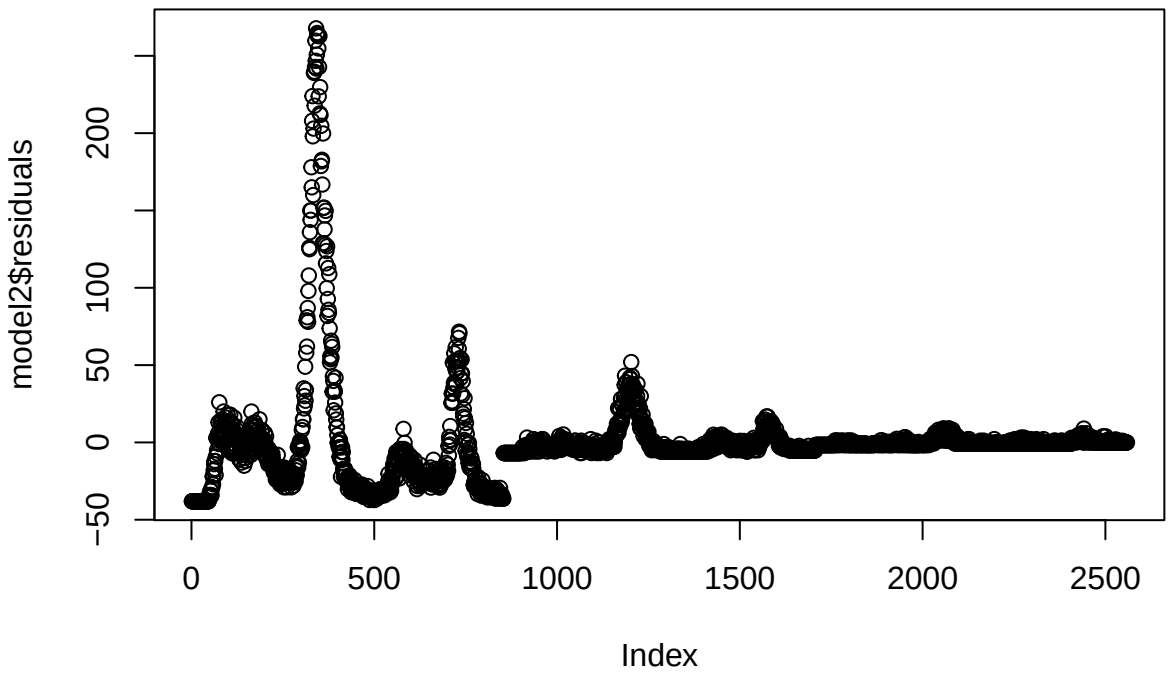
```
shapiro.test(model3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model3$residuals
## W = 0.52415, p-value < 2.2e-16
```

```
#plot of model residuals
plot(model1$residuals)
```




```
plot(model2$residuals)
plot(model2$residuals)
```




```
#AIC, BIC, R^2 and adjusted R^2 code (appendix 7)
results <- AIC(model1,model2,model3)

models <- list(model1,model2,model3)

results$BIC <- sapply(models,BIC)

models_summary <- lapply(models,summary)

for(i in 1:length(models)){
  results$R2[i] <- models_summary[[i]]$r.squared

  results$R2adj <- models_summary[[i]]$adj.r.squared

}

kable(results,digits = 2, align = 'c')
```

	df	AIC	BIC	R2	R2adj
model1	3	25593.88	25611.42	0.00	0.3
model2	5	25026.46	25055.70	0.20	0.3
model3	6	24668.62	24703.71	0.31	0.3

```
## model3 is the best
```

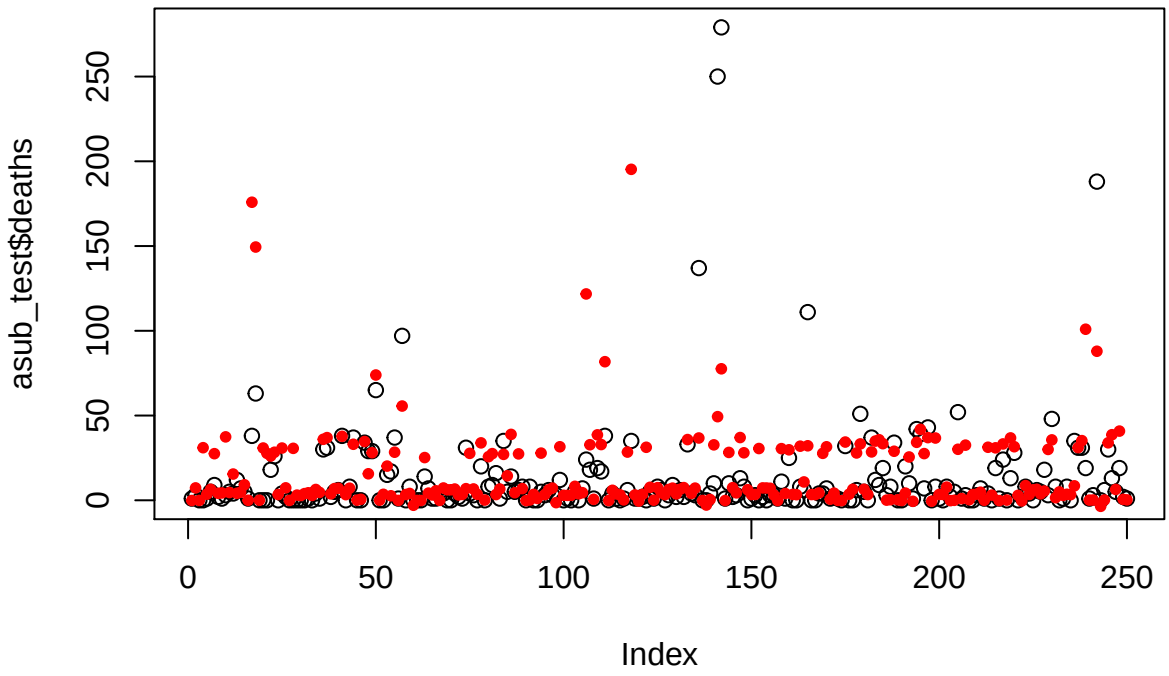
```r
#Plots for random sample testing fitted vs predicted values (appendix 8)
splitter <- sample(1:nrow(data),250, replace = F)

asub_train <- data[-splitter,]
asub_test <- data[splitter,]

model4 <- lm(deaths~Year+County+cases, data = asub_train)

prediction <- predict(model4, asub_test)

plot(asub_test$deaths, pch = 1)
points(prediction, pch = 20, col = 'red')
```




```r
#paired t-test for deaths and cases (appendix 9)
t.test(data$deaths,data$Year, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  data$deaths and data$Year
## t = -2824.3, df = 2558, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2007.299 -2004.513
## sample estimates:
```

```
## mean of the differences
##               -2005.906
```