Regression

David Favela Corella

Linear Regression

Works by plotting X and Y and finding a relationship between them. This relationship is found by defining other parameters that are included in the linear regression formula, such as w and b. Some of the strengths include being a simple algorithm to implement and working well with data that tends towards a linear pattern. The weakness is that it has a high amount of bias because it assume the data will trend to be linear.

To begin, we get the data from the csv file. We are going to be working with covid-19 data.

```
df <- read.csv("covid_19.csv", header=TRUE)
str(df)
```

```
## 'data.frame':    18424 obs. of  9 variables:
##  $ Country.Region: chr  "Argentina" "Australia" "Australia" "Australia" ...
##  $ Lat           : num  -38.4 -35.5 -33.9 -12.5 -27.5 ...
##  $ Long          : num  -63.6 149 151.2 130.8 153 ...
##  $ Date          : chr  "1/22/2020" "1/22/2020" "1/22/2020" "1/22/2020" ...
##  $ Confirmed     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Recovered     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Active        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ WHO.Region    : chr  "Americas" "Western Pacific" "Western Pacific" "Western Pacific" ...
```

Then we clean and separate the data. We are going to use the Confirmed, Deaths, and Active columns for the data analysis. In this case, there's no cleaning as I made sure the data had no null values.

```
df <- df[,c(8,5,6)]
df$Confirmed <- as.integer(df$Confirmed)
df$Deaths <- as.integer(df$Deaths)
df$Active <- as.integer(df$Active)
dim(df)
```

```
## [1] 18424     3
```

```
head(df)
```

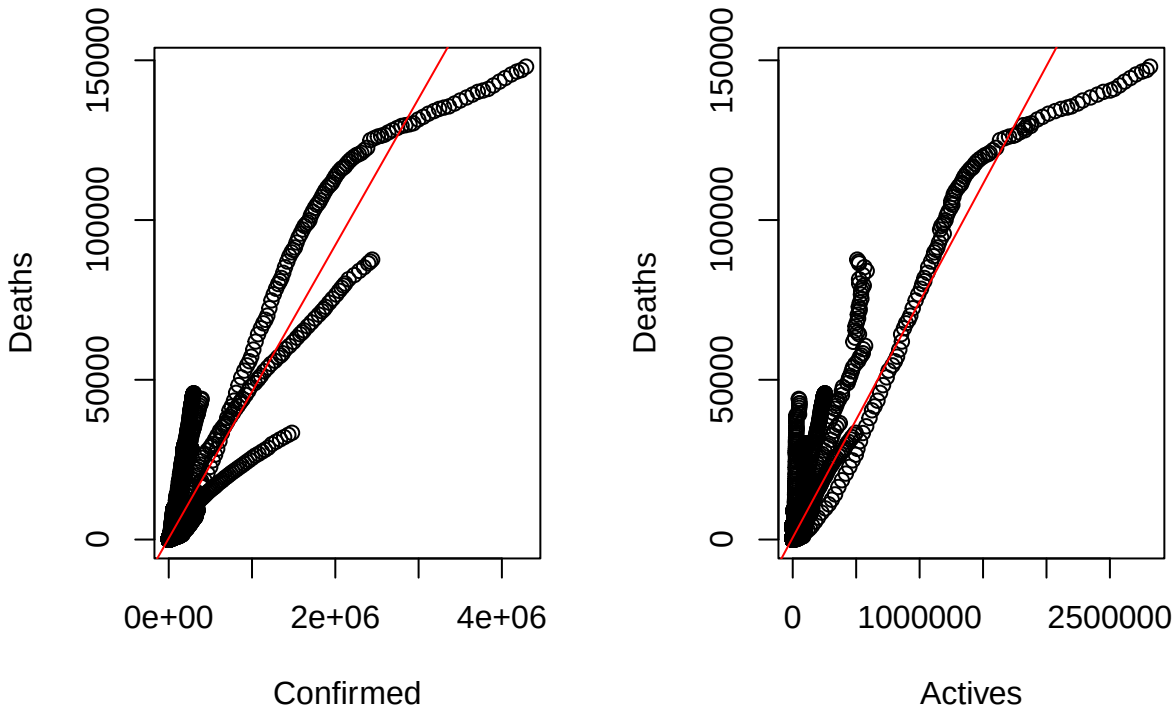
```
##   Active Confirmed Deaths
## 1      0         0      0
## 2      0         0      0
## 3      0         0      0
## 4      0         0      0
## 5      0         0      0
## 6      0         0      0
```

```
str(df)
```

```
## 'data.frame':    18424 obs. of  3 variables:
##  $ Active   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Confirmed: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Deaths   : int  0 0 0 0 0 0 0 0 0 0 ...
```

Data is then plotted. We plot confirmed cases against deaths and active cases against deaths.

```
par(mfrow=c(1,2))
plot(df$Deaths~df$Confirmed, xlab="Confirmed", ylab="Deaths")
abline(lm(df$Deaths~df$Confirmed), col="red")
plot(df$Deaths~df$Active, xlab="Actives", ylab="Deaths")
abline(lm(df$Deaths~df$Active), col="red")
```



We set the seed fo the model. The data is split in 80/20 for train and test.

```
set.seed(1234)
i <- sample(1:nrow(df), nrow(df)*0.80, replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

##Linear regression model algorithmn The model is created from the deaths and confirmed

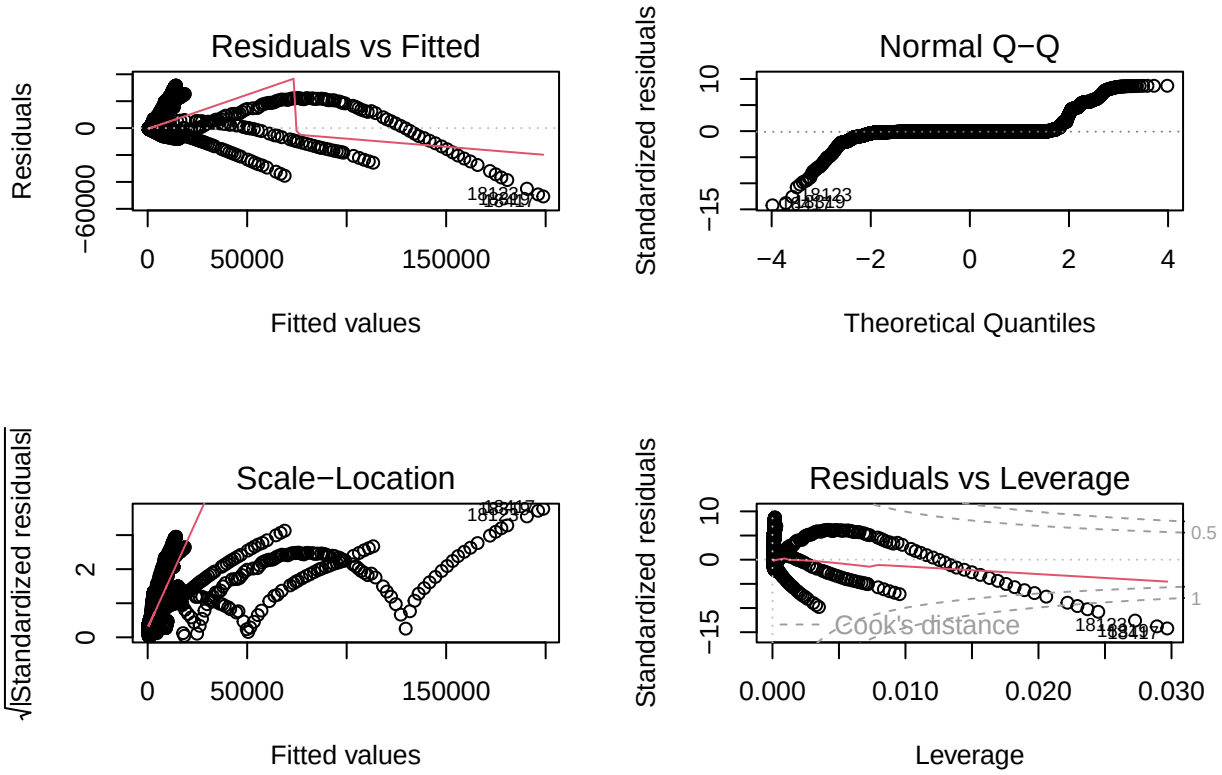
```
lm1 <- lm(Deaths~Confirmed, data=train)
summary(lm1)
```

```
##
## Call:
## lm(formula = Deaths ~ Confirmed, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50795   -424   -405   -402  31523
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.021e+02  3.022e+01    13.3   <2e-16 ***
## Confirmed   4.625e-02  1.465e-04   315.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3625 on 14737 degrees of freedom
## Multiple R-squared:  0.8712, Adjusted R-squared:  0.8712
## F-statistic: 9.965e+04 on 1 and 14737 DF,  p-value: < 2.2e-16
```

Residuals for the linear model of deaths and confirmed

We can see that there is good indication the model was good, as we can see the pattern in the data. Residuals are also lined up well in the normal Q-Q.

```
par(mfrow=c(2,2))
plot(lm1)
```



Evaluate the test set on the model

We evaluate the test data unto the model created and we see that the model gives a correlation of 0.93 on the test data, which is really good.

```
pred1 <- predict(lm1, newdata=test)
cor1 <- cor(pred1, test$Deaths)
mse1 <- mean((pred1-test$Deaths)^2)
rmse1 <- sqrt(mse1)

print(paste('correlation:', cor1))
```

```
## [1] "correlation: 0.931126863751436"
```

```
print(paste('mse:', mse1))
```

```
## [1] "mse: 10548474.4274879"
```

```
print(paste('rmse:', rmse1))
```

```
## [1] "rmse: 3247.84150282737"
```

Multiple Linear Regression

For this model, we both the confirmed and active cases from the data set and create the model.

```
lm2 <- lm(Deaths~Confirmed+Active, data=train)
summary(lm2)
```

```
##
## Call:
## lm(formula = Deaths ~ Confirmed + Active, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -58477   -464   -450   -447  29733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.474e+02  2.935e+01   15.25   <2e-16 ***
## Confirmed   3.207e-02  4.859e-04   66.00   <2e-16 ***
## Active      2.430e-02  7.966e-04   30.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3516 on 14736 degrees of freedom
## Multiple R-squared:  0.8788, Adjusted R-squared:  0.8788
## F-statistic: 5.343e+04 on 2 and 14736 DF,  p-value: < 2.2e-16
```

We use anova to compare the two different models that we have.

```
anova(lm1, lm2)
```

```
## Analysis of Variance Table
##
## Model 1: Deaths ~ Confirmed
## Model 2: Deaths ~ Confirmed + Active
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1  14737 1.9368e+11
## 2  14736 1.8217e+11  1 1.1507e+10 930.81 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Evaluate and predict the multiple linear regression model

We run the test data set on the multiple linear regression model and we see that it has a slightly higher correlation that the previous model created.

```
pred2 <- predict(lm2, newdata=test)
cor2 <- cor(pred2, test$Deaths)
mse2 <- mean((pred2-test$Deaths)^2)
rmse2 <- sqrt(mse2)

print(paste('correlation:', cor2))
```

```
## [1] "correlation: 0.932688954916072"
```

```
print(paste('mse:', mse2))
```

```
## [1] "mse: 10375435.9282872"
```

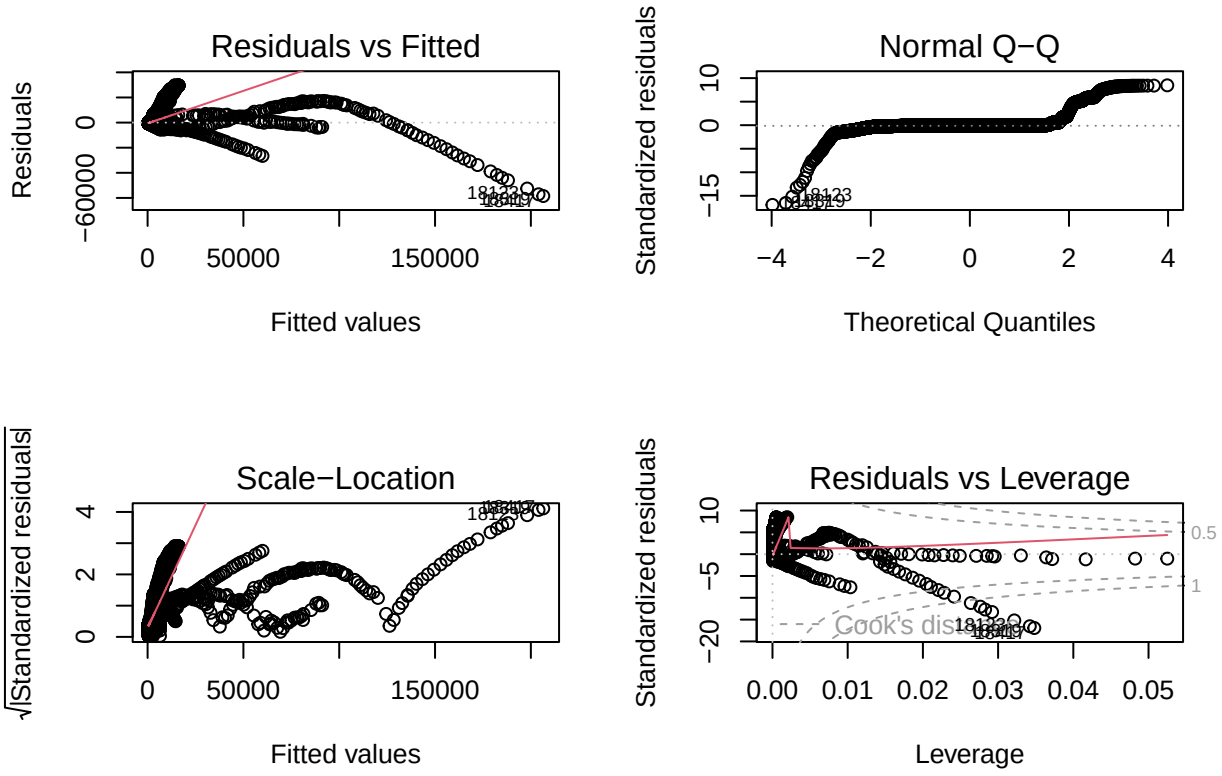
```
print(paste('rmse:', rmse2))
```

```
## [1] "rmse: 3221.09235016434"
```

Residuals for the multiple linear model of deaths and confirmed + active

We can see that there is good indication the model was appropriate and good, as we can see the pattern in the data. Residuals are also lined up well in the normal Q-Q and in the residuals vs leverage.

```
par(mfrow=c(2,2))
plot(lm2)
```



Third linear model

For the third linear model, we use deaths and active data. I tried doing a logarithmic run with deaths but ran into several issues.

```r
lm3 <- lm((Deaths)~Active, data=train)
summary(lm3)
```

```
##
## Call:
## lm(formula = (Deaths) ~ Active, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -62691   -631   -628   -625  49090
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.282e+02  3.326e+01   18.89   <2e-16 ***
## Active      7.459e-02  2.652e-04  281.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4002 on 14737 degrees of freedom
## Multiple R-squared:  0.843,  Adjusted R-squared:  0.843
## F-statistic: 7.913e+04 on 1 and 14737 DF,  p-value: < 2.2e-16
```

Evaluate and predict the third linear model

We see that the model for death and active is lower than the previous two models.

```r
pred3 <- predict(lm3, newdata=test)
summary(pred3)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    627.4    628.2    629.1   1655.6    654.1 205976.7
```

```r
cor3 <- cor(pred3, test$Deaths)
mse3 <- mean((pred3-test$Deaths)^2)
rmse3 <- sqrt(mse3)
print(paste('correlation:', cor3))
```

```
## [1] "correlation: 0.912042452552287"
```

```r
print(paste('mse:', mse3))
```

```
## [1] "mse: 13401349.7763904"
```

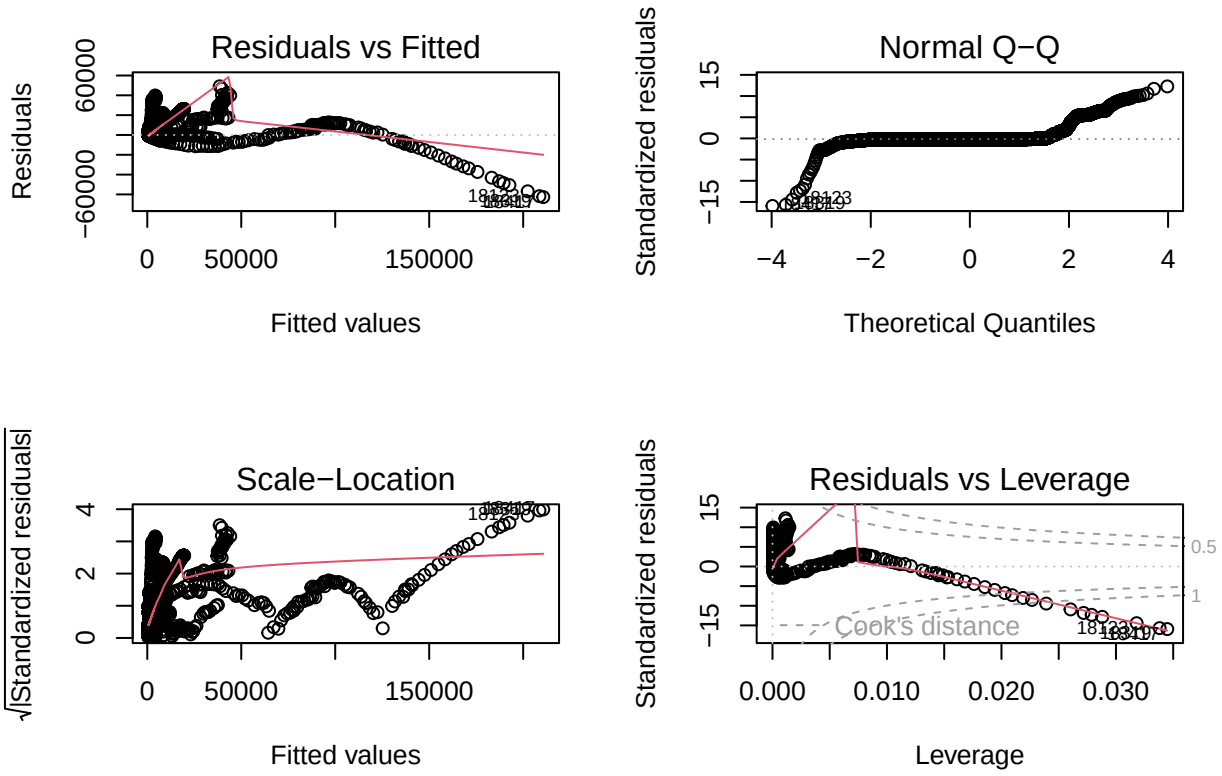
```r
print(paste('rmse:', rmse3))
```

```
## [1] "rmse: 3660.78540430745"
```

Plot the residuals

We see the residuals vs fitted tend to be a better fit than previous models, along the normal Q-Q. The residuals vs leverage has a significant skew which dampens it's usability.

```
par(mfrow=c(2,2))
plot(lm3)
```



Conclusion data

From all the models, the multiple linear regression model had a higher correlation and lower mse, which is good. It was followed by the first model, which was deaths and confirmed cases which had slightly lower and higher correlation and mse. Lastly, the last model, which is deaths and actives. The reason we got these results include the fact that active cases don't necessarily take into account past data and is the most passive data set. The confirmed data includes those that are active and thus may have a better relation with how many have died, which helps in this case.