Covid-19 in Romania/US Police Shootings

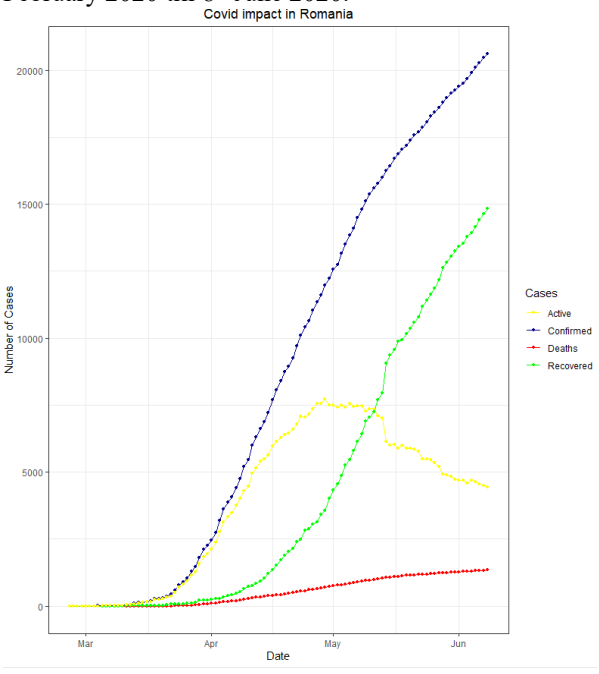
Mihai Matei
Statistics for Data Science Course
Data Science Master, 2020
University of Bucharest, Faculty of Mathematics and Computer Science

Abstract

This paper tries to apply statistical methods, namely time series analysis, to predict future confirmed and death cases of Covid-19 in Romania as well as to devise a statistical experiment to see what are the factors that impact the age of death of a individual by police shootings. Finally I will use the linear regression to predict the age using prior data while analyzing the goodness of fit for each model.

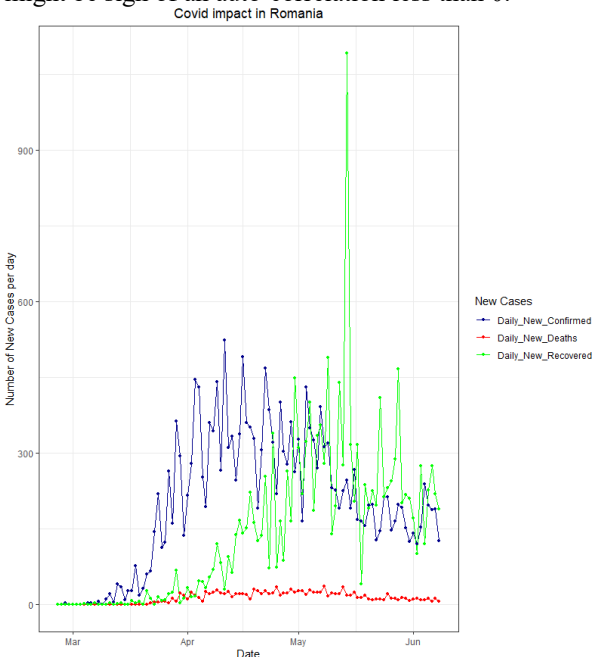
1. Covid-19 data set

Covid-19 pandemic has been the main topic in the news for several months to date. Most aggregated information online contain the number of reported cases, the deaths and several other numeric metrics. A good global source of information can be found on kaggle [1] and contains per date, country and region Covid-19 information such as the total number of confirmed cases, recovered, active and deaths. Since I am only interested in analyzing Romania I filter the data set for this region and take the 104 entries from the first case was confirmed in Romania on 06 February 2020 till 8th June 2020.



As seen in the plot the total number of confirmed cases is increasing linearly not really exponential and the number of death shows a growth trend. The number of active cases (yellow) shows the sign of the bell curve that everyone has been talking about with the top at the start of May.

Probably a better view is when we consider the number of new cases per day (total number of cases today minus the total number of cases in the previous day). We can see that the trend is pretty spiky which might be sign of an auto-correlation less than 0.



The death trend (red) shows signs of stationarity.

2. ARIMA time series analysis

Because the data is highly serial in nature (per day) I will use Arima time series methods to analyze the trends.

I will analyze 8 different time series: Confirmed, Deaths, New Confirmed Cases, New Deaths Cases and

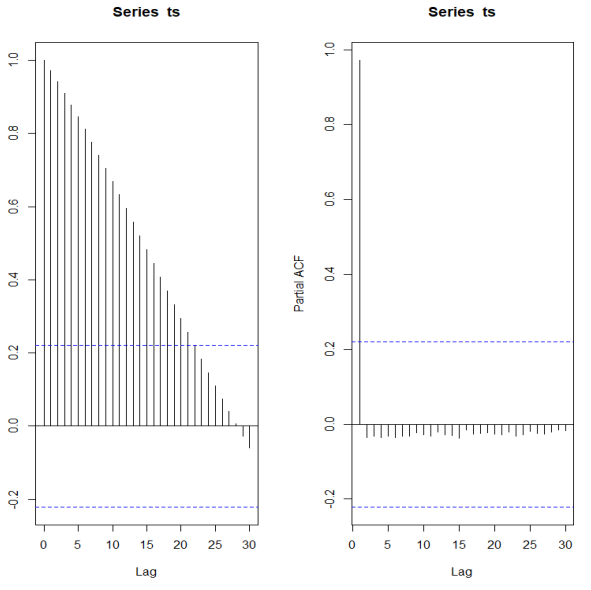
for each of these 4 I will also take their log data for improving their stationarity.

For each of these 8 time series I will do the following:

- Plot the ACF, PACF to check what could be the proper AR(p), MA(q) settings for the ARMA. We know that for AR the ACF decays and PACF cuts of and for MA the ACF cuts off and PACF decays as described in Course 4. These can help use to choose the proper p,q values of the model. In the ACF, PACF plots the values between the dotted blue lines do not represent values that have any statistical meaning so they can be considered as 0.
- To prevent over-fitting I will use the model with the lowest Akaike Information criterion [2] by doing a grid search on different values of p,q as well as using the diff value of 0 or 1 when fitting the ARIMA model.
- To test the best model found I plot the standardised Residuals and their QQ Plot in order to check for signs of a normal random distribution across the errors and also their random spread across time with no apparent correlation. To check the lack of correlation of the residuals I am plotting the ACF of the residuals that, after the 1st entry should be between [-2,2] to have no statistical meaning and hence no correlation. Ljung-Box statistic is also plotted to check that we cannot reject null hypothesis of the residuals being independent (not auto-correlation). So we need for the p-value in this case to be high so that we cannot reject the null hypothesis (which does not imply it is correct). So by using this statistic test we just try to check if we are not wrong, not that we are correct.
- Finally I plot the time series, the predicted values for the next 20 values as well as the 95% confidence intervals. I will notice this intervals are quite large for some predictions.

Below in the next 2 points I will present the results for the total deaths and total number of confirmed cases

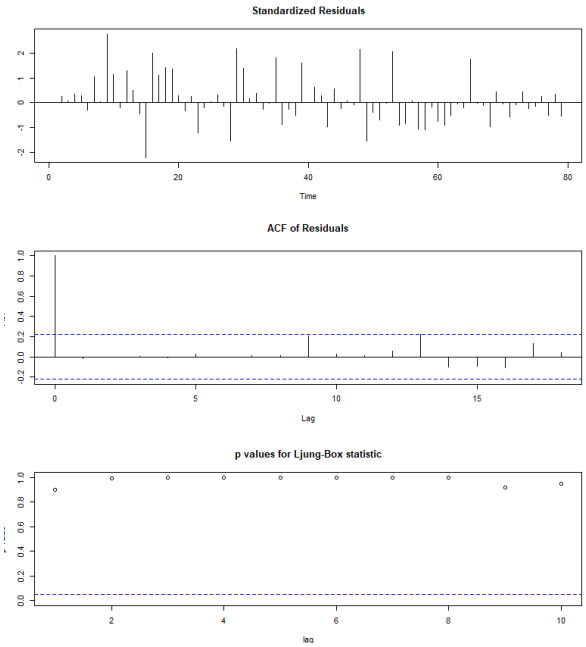
2.1. Romania total confirmed cases prediction



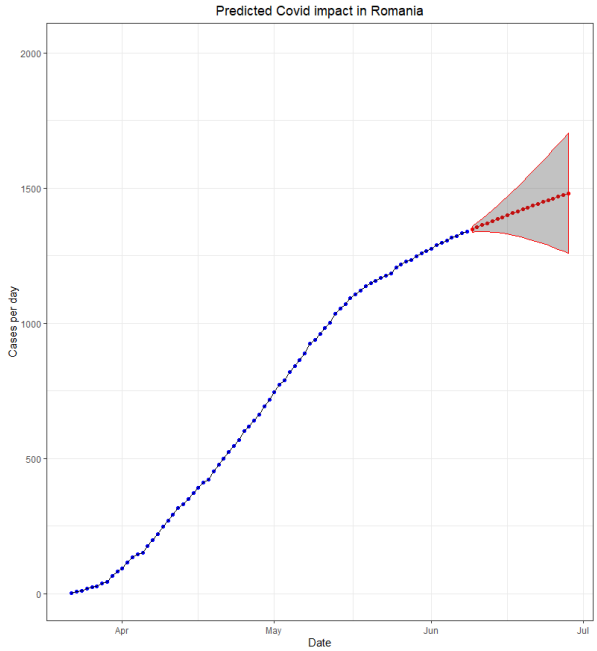
Above the ACF and PACF is clearly decreasing for the first 30 lags.

When running the grid search for p,d,q we find the best mode to be ARIMA(6,1,2) with an AIC value of 509.

The residuals show no signs of auto-correlation as can be seen below, with the p-value of the Ljung-Box statistics to be very high so we cannot reject null hypothesis.



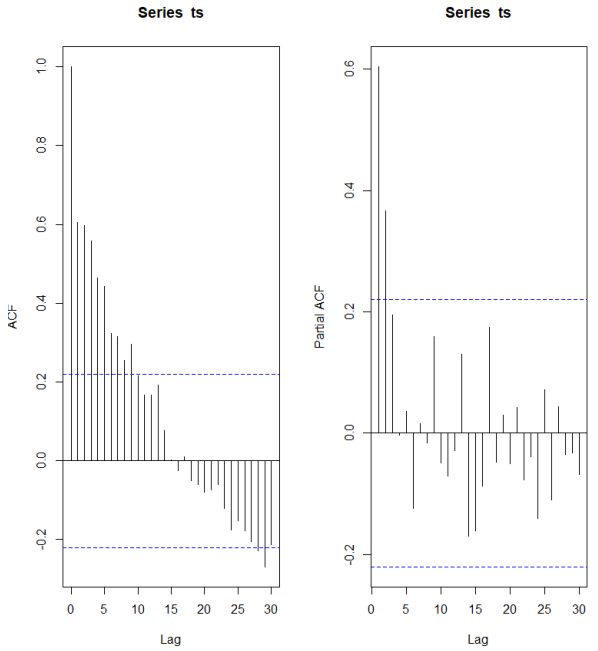
Finally lets see the predicted plot and the confidence intervals:



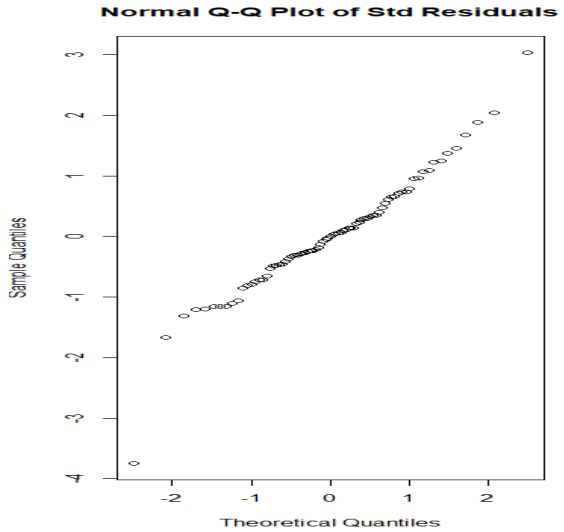
You can see that the prediction indicates that the total number of confirmed cases will also increase so the total number of new cases per day will be greater than 0.

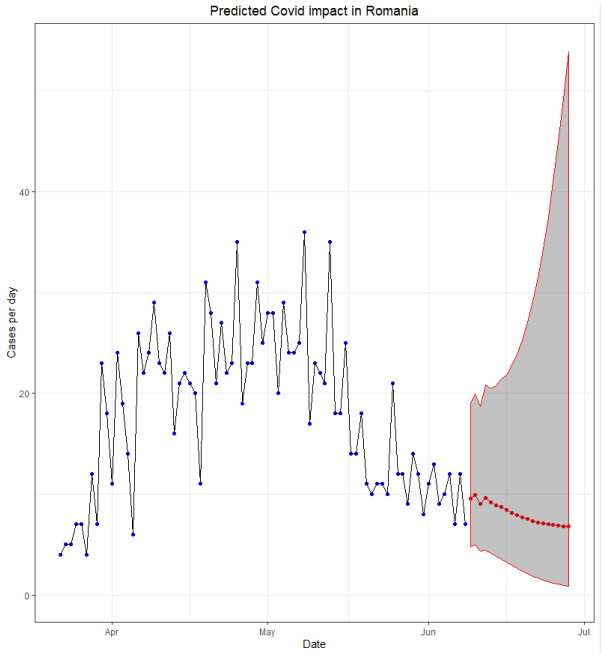
2.2 Romania new deaths per day predictions

I do the same thing for the number of new deaths per day and check the ACF and PACF. Since the plots are not really that decreasing I smooth the data by applying log on it. Now I have the following results:



After doing the hyper-parameter grid search on p,q I obtained the best model to be ARIMA(6,1,2) with an AIC value of 73.37. The residuals statistical tests are also good and show no auto-correlation and independence and normality. Let's check just the QQ plot for them:

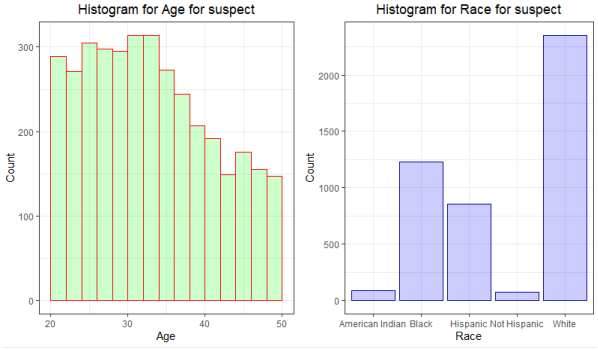




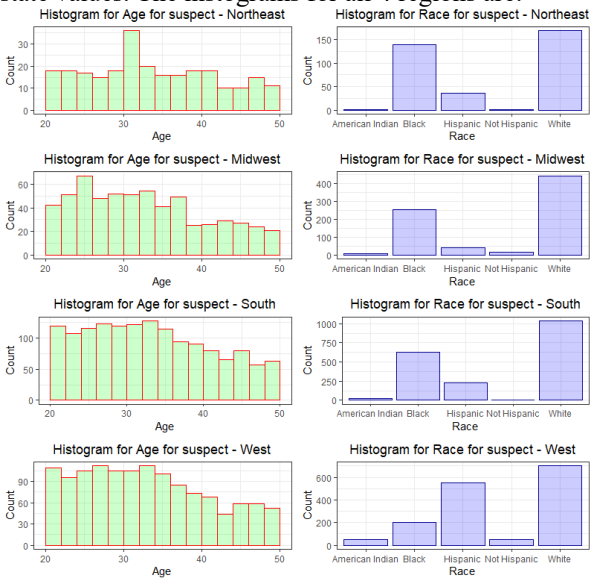
The prediction for new daily deaths shows a decreasing trend but check the larger 95% confidence interval.

3. US Police Shootings Data Set

The dataset taken from [3] contains the police arrests that lead to the death of the suspect in United States. The data contains 5338 entries from January 2015 till June 2020. It contains information about the name of the suspect, the date, the location of the arrest, the threat level the suspect and most importantly the race, gender and age. First I filter the data set and take all rows with non-empty values for data, age, gender, race, state and threat level. It leads to 400 rows where the race must be known and coerce gender, race, state and thread_level columns to be factors with the number of levels of 2,3,51,3. I am interested in mainly in 2 factors: the race and the age of the suspect and hence I plot the histogram of the 2 values:



As one can see the age is maximum between 20-40 but with no significant differences between them. Also the Black, Hispanic and White population seems to be the largest one involved. But please keep in mind that this histogram/density estimator does not take into account the overall proportions of the ethnic groups in the total US population – for example black people represent ~16% of the total population. To be able to apply 2 factor ANOVA on the above metrics we separate the data into the 4 US regions – Northeast, Midwest, South, West by grouping by their respective state values. The histograms for all 4 regions are:



One can see that in Northeast the Black population is more impacted, even more than in South, and in the West the Hispanic population is very high. Whites are the largest one in all regions but one must keep in mind the global population distribution.

4. Two-way ANOVA

I will test using ANOVA two-way statistical test to check if age is influenced by race, region or threat_level. Then I will compute the average age and number of deaths per race for each state and check influences there as well.

First ANOVA analysis requires homogeneity of variance of the sample data. For the 4 Regions described above we apply the Bartlett test where the null hypothesis is that all group variances are equal against the alternate hypothesis that at least 2 of the are different. So I am interested in a larger p-value so that I cannot contradict the null hypothesis. For the entire filtered data set, grouped into 4 regions/groups the Bartlett test is:

```
Bartlett test of homogeneity of variances
data:  list(group1$age, group2$age, group3$age, group4$age)
Bartlett's K-squared = 26.691, df = 3, p-value = 6.835e-06
```

This is not good as p-value is low so we filter the data to take **only ages between 18 and 40** yielding 2989 observations. This time I cannot reject homogeneity of variances as Bartlett's:

```
Bartlett test of homogeneity of variances
data:  list(group1$age, group2$age, group3$age, group4$age)
Bartlett's K-squared = 0.70764, df = 3, p-value = 0.8714
```

In the Bartlett test the df - degree of freedoms parameter of the statistic is the number of groups - 1

For the data represented by the selected groups I apply 2 two-way ANOVA for:

- age~race * Region

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
race	**4**	**2108**	**527.0**	**15.855**	**7.66e-13**	*******
Region	3	192	64.1	1.927	0.123	
race:Region	12	146	12.2	0.367	0.975	
Residuals	2869	95358	33.2			

Notice that race does influence the data but the Region is not significant. Also between race and Region there is no significant influence. Notice that the DF parameter – degree of freedoms is the number of factors - 1 for race and Region alone and 12=3*4 for the race:Regions interaction. For the Residuals DF=no_observations -1 – no_df_params = 2889 -1 – (3+4+12).

- age~race * threat_level

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
race	**4**	**2108**	**527.0**	**15.931**	**6.62e-13**	*******
threat_level	2	46	23.2	0.702	0.496	
race:threat_level	**8**	**581**	**72.7**	**2.197**	**0.025**	*****
Residuals	2874	95068	33.1			

In this case I can also notice that we can reject Ho that race does not have a significant influence (or race might influence the age as above). One interesting aspect is that race and threat_level might be influenced one another.

Let's see the same thing for ages between 40 and 60 which also have homogeneity according to Bartlett test. The same 2 combinations of factors are studied:

- age~race * Region

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
race	**4**	**1176**	**294.04**	**9.869**	**7.27e-08**	*******
Region	3	46	15.21	0.510	0.675	
race:Region	10	328	32.83	1.102	0.357	
Residuals	1283	38227	29.79			

The same thing only race has a significant factor.

- age~race * threat_level

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
race	**4**	**1176**	**294.04**	**9.876**	**7.17e-08**	*******
threat_level	2	8	3.90	0.131	0.877	
race:threat_level	6	244	40.63	1.364	0.226	
Residuals	1288	38349	29.77			

As opposed to the 18-40 category here only race influences and there is no longer a correlation between race and thread level.

To check further influences I group the data per region and state and count the number of kills per race and the average age for each race. So this is done for each state in US:

```
  Region state race             no_kills_per_race avg_age
1 West   AK    American Indian                  2      27
2 West   AK    Black                            3      32
```

I also take the 4 groups of states grouped by US regions and do the Bartlett's homogeneity test for the average age and number of death per race as computed above and check that both have equal variances.

I can the do the two-way ANOVA test for the following:

- no_kill_per_race~ state * avg_age

```
              Df Sum Sq Mean Sq F value   Pr(>F)
state         50 107855  2157.1   1.404 0.0843 .
avg_age        1   1192  1192.0   0.776 0.3809
state:avg_age 49  34463   703.3   0.458 0.9982
Residuals     84 129030  1536.1
```

- avg_age~ state * no_kill_per_race

```
                     Df Sum Sq Mean Sq F value   Pr(>F)
state                50   2784   55.68   0.951  0.570
no_kills_per_race     1     58   57.88   0.988  0.323
state:no_kills_per_race 48   2962   61.70   1.054  0.410
Residuals            85   4977   58.56
```

So we can see that the state does not significantly influence the number of killings per race or the average age per race.

5. Linear regression

To use the linear regression we need the predictor to be a continuous variable. I will use the same police shooting data set discussed above, filtered for empty values and with age – the target variable and the features to be the date, gender, race, treat_level factors. The linear regression model assumes the errors is independent and the errors has constant variance and are normally distributed and there is no correlation between successive values. For this checking the model's residuals (the difference between the true and predicted value) is used.

First I check the correlation between age on one side and gender, race and state on the other using the Pearson's product momentum correlation and all 3 statistical tests issue the alternate hypothesis: true correlation is not equal to 0. This means that age is influence by at least the above 3 features.

I also do random a 0.9/0.1 train/test split of the data to asses the performance of the models against unseen data yielding a train set size of 4140/460.

I train 4 linear regression models using the age as the target/response value against the train set:

- age~date+gender+race+state+threat_level-1

Linear regression model with all features with no intercept. GenderM, raceBlack and stateNY are among the most significant features. Adjusted R2 is above 0.8:

```
Residual standard error: 12.23 on 4081 degrees of freedom
Multiple R-squared:  0.9013,    Adjusted R-squared:  0.8998
```

- AIC(age~date+gender+race+state+threat_level-1)

The above model after applying the Akaike Information Criteria step to filter out some features. I obtained a model with only 8 features with genderF, genderM, raceBlack the most important. Adjsuted R-2 should be very similar to the original model:

```
Residual standard error: 12.27 on 4131 degrees of freedom
Multiple R-squared:  0.8994,    Adjusted R-squared:  0.8992
```

The larger degrees of freedom seen there is because date is also present in the AIC model.

- BIC(age~date+gender+race+state+threat_level-1)

The above model after applying the Bayesian Information Criteria step to filter out some features. I obtained a model with only 6 features with genderF, genderM, raceBlack the most important. Adjsuted R-2 should be very similar to the original model:

```
Residual standard error: 12.28 on 4134 degrees of freedom
Multiple R-squared:  0.8991,    Adjusted R-squared:  0.899
```

- age~state+gender+race-1

Linear regression model with only the 3 most important predictors and no intercept. The state feature has the most importance in all 51 factors. Adjusted R2 is also above 0.8:

```
Residual standard error: 12.24 on 4084 degrees of freedom
Multiple R-squared:  0.9011,    Adjusted R-squared: 0.8997
```

5.1 Fitness of the models

For each of the 4 models we check the goodness of fit by doing the following:

- use plotres library to plot some statistic of the residuals including the QQ plot
- do a ANOVA test of the model to check the importance of each predictors
- do a Shapiro-Wilk test to test the normality of the residuals (this test has the highest power for normality testing). The null hypothesis is that residuals are normally distributed so we need a higher p-value so that we are not able to reject the hypothesis.
- check the homoscedascity of the residuals by plotting them against the model response to check that they are equally distributed.

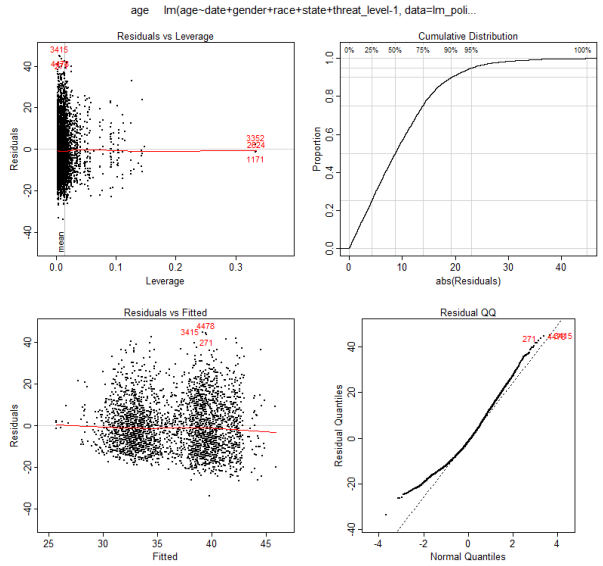
I check the violation of constant variance with homoscedascity check and the violation of normality with QQ plot and Shapiro-Wilk test.

The QQ (quantile-quantile) plot is used in statistics to plot the similarity between different distributions, in our case through a parametric curve – a line. Since the X axis contains the quantiles of the Normal distribution and the Y – the quantile of the Residuals if the two distributions were the same the plot should be the diagonal line with more data around -2,2.

The null hypothesis of the Shapiro-Wilk test is that the data is Normally Distributed so a small p-value means we reject the null hypothesis – which is wrong.

Now lets see how these steps for:

– **age~date+gender+race+state+threat_level-1**



Analysis of Variance Table

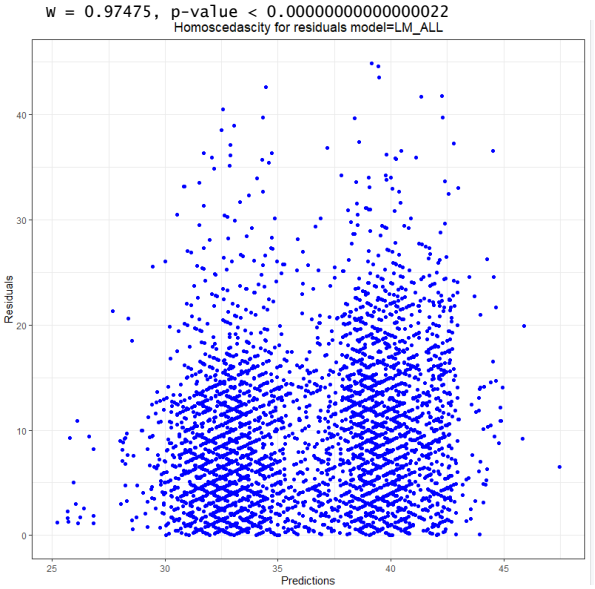
Response: age

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
date	1	5504634	5504634	36810.6165	< 0.00000000000000022 ***
gender	2	4053	2027	13.5528	0.00000136 ***
race	4	49023	12256	81.9559	< 0.00000000000000022 ***
state	50	11846	237	1.5843	0.005648 **
threat_level	2	912	456	3.0487	0.047529 *
Residuals	4081	610270	150		

Shapiro-Wilk normality test

data: residuals

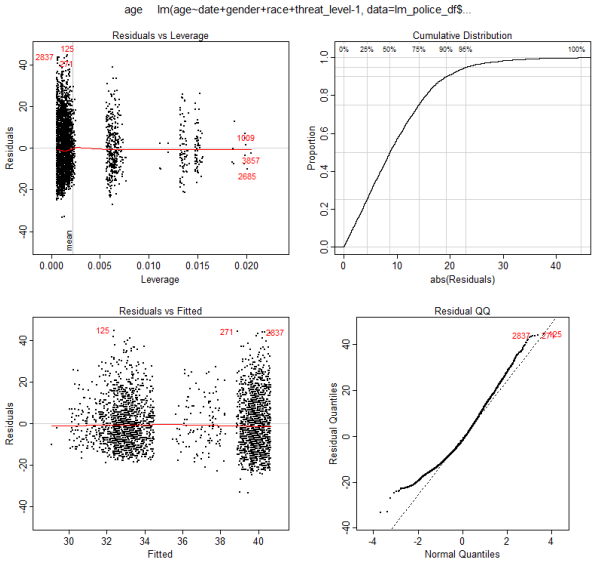
W = 0.97475, p-value < 0.00000000000000022

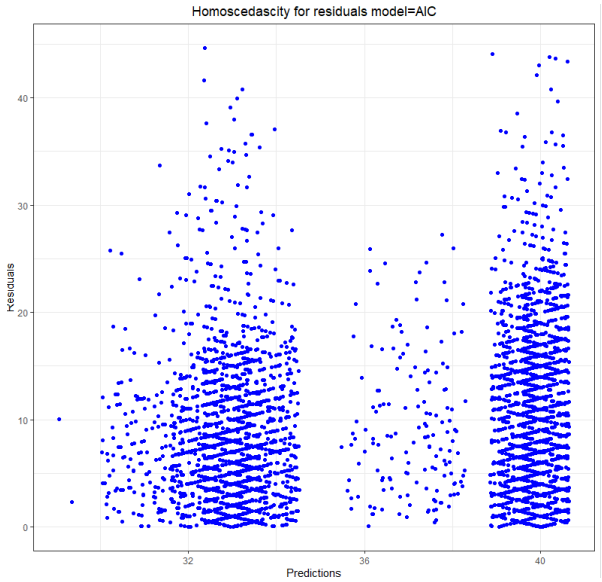


One can see that the normality of errors is rejected by the Shapiro-Wilk test due to the low p-value. Also the errors are centered between 30-45 which shows a tendency of the model to predict data in this range.

So this model seems not to be a really good fit even through Adjusted R value is high.

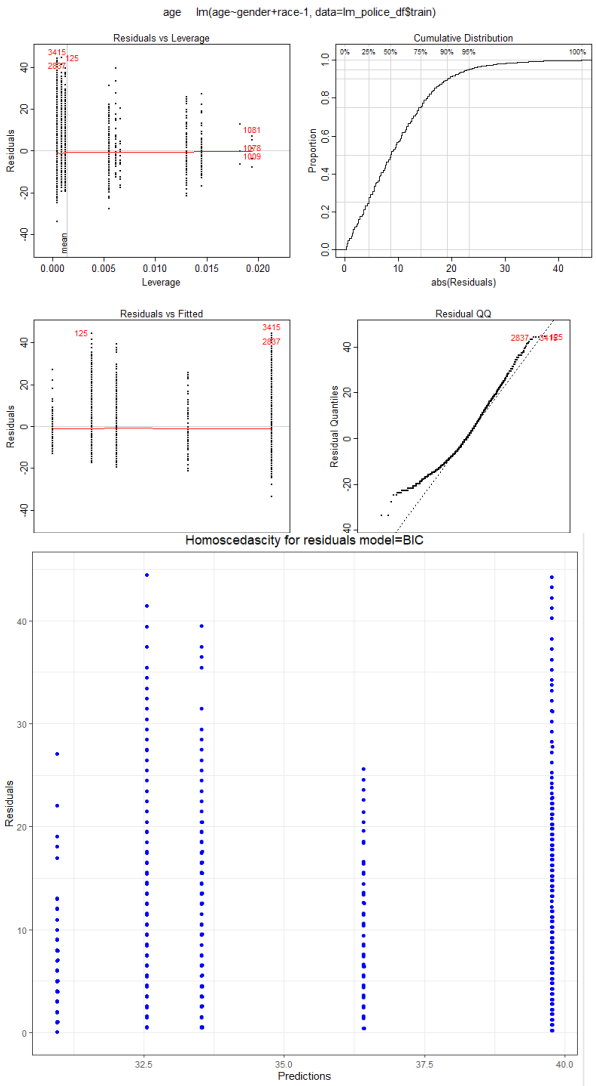
– **AIC(age~date+gender+race+state+threat_level-1)**





The normality of errors is also rejected here and we can see that if we apply the AIC criterion the prediction band is limited to 25-35 and 38-42. So this model is also not a good fit but this was expected due to the fact that we used a parameter filtering method for the original values.

- BIC(age~date+gender+race+state+threat_level-1)



The same thing for BIC filtering as well but in this case the prediction range is even more impacted making this a very biased model with a low capacity.

– age~state+gender+race-1

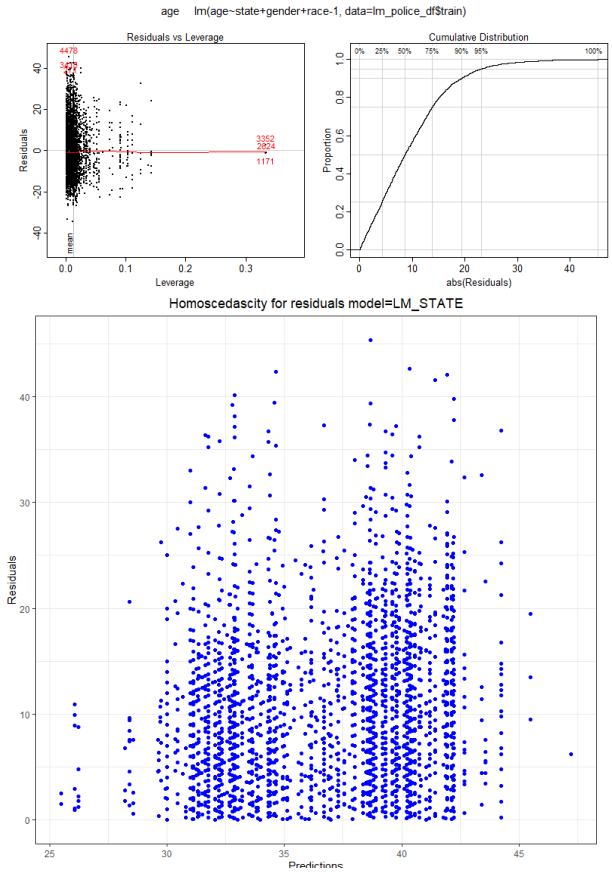
I apply a new model with just the 3 parameters that were seen to have the greatest influence on the data. This can be seen in the ANOVA analysis of the model:

```
Response: age
             Df  Sum Sq Mean Sq  F value              Pr(>F)
state         1 5559060  109001 715.9902 <0.0000000000000002 ***
gender        1      51      51   0.3363               0.562
race          4   48882   12221  80.2724 <0.0000000000000002 ***
Residuals  4084  621742     152
```

Normality of residuals also fails:

```
        Shapiro-Wilk normality test

data:  residuals
W = 0.97466, p-value < 0.00000000000000022
```



The prediction is also very close to the full model also centered around the 30-45 values so this model seems also biased as well.

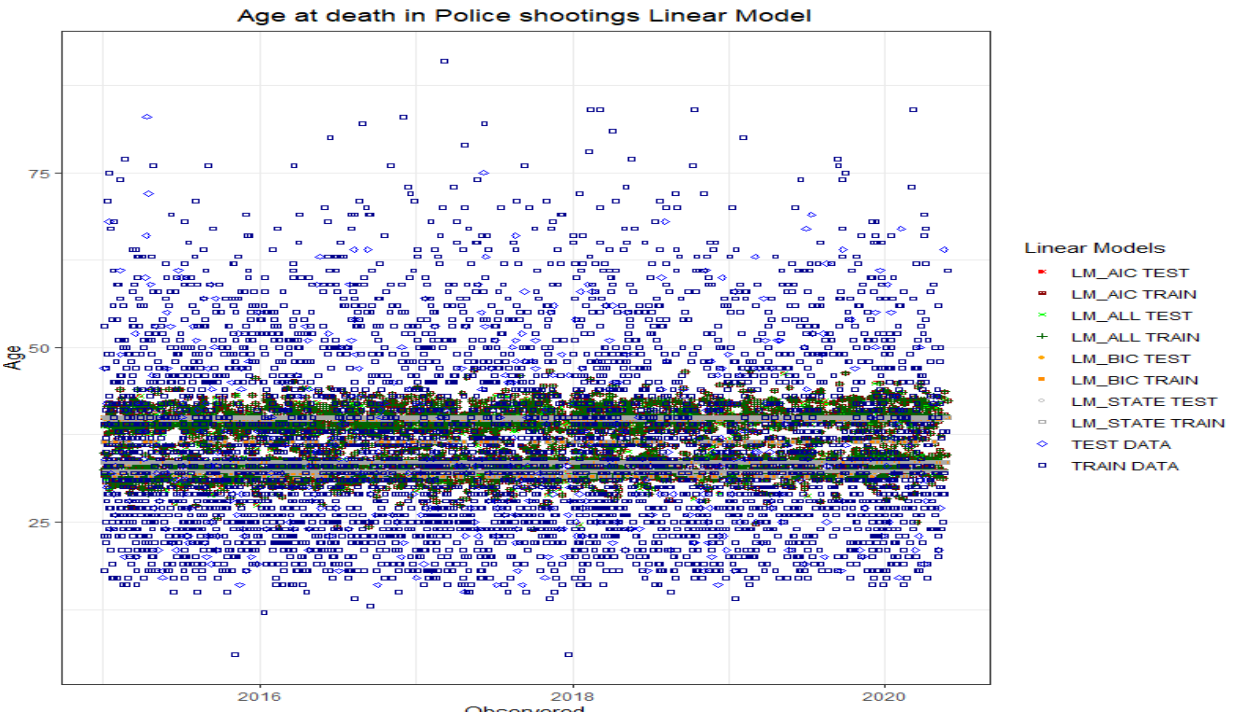
5.2 Model performance

To test the the models performance we asses them against the test data. The following results are obtained:

	R2	MSE
LM_ALL	0.8930270	164.3002
AIC	0.8941613	162.5581
BIC	0.8936993	163.2677
LM_STATE	0.8936993	163.2677

At first glance these seem very good results. But we have seen when doing the goodness of fit for the models that the residuals are not normally distributed an there is no homoscedascity as the residuals seems to be centered around certain prediction values.

Hence we plot the predictions of all models on both train and test data:



The figure shows that the models really predict data in the 27-40 range where most of the data is concentrated. Data outside this age range is not predicted at all and considered as an outlier (which it is not).

One possible cause of this bad fit is the lack of linearity in the data. This can be solved by increasing the models capacity/complexity, by using composed feature in the linear mode (e.g. state*race) by applying kernels or by using models that can learn non-linear data (neural networks, svm with kernels etc). Or perhaps using non-linear regression as discussed in Course 12.

6. References

[1]https://www.kaggle.com/imdevskp/corona-virus-report

[2]https://en.wikipedia.org/wiki/Akaike_information_criterion

[3]https://www.kaggle.com/andrewmvd/police-deadly-force-usage-us