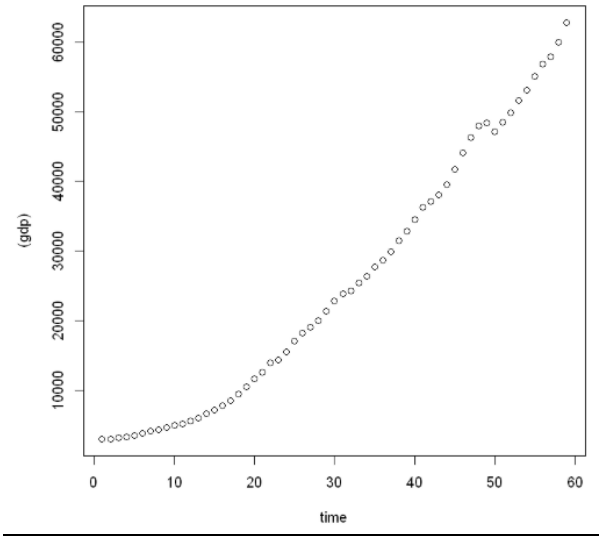
Hursh Desai

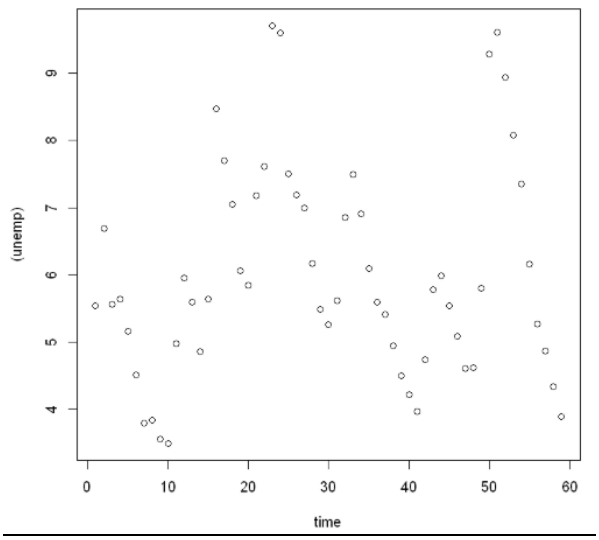
Introduction

I wanted to see what of the few economic/social factors I picked which ones have the strongest effect on immigration to the US over time. I know that there are push factors that push people out of their country and pull factors that pull people towards certain countries when it comes to the reasons that people immigrate to certain countries. In this analysis I wanted to focus on the pull factors in work for the US that cause people to come to the US. The factors that I chose are GDP per capita in the current dollar[1], the unemployment rate[2], and the death rate in the US[3] and their effect on the number of immigrants that naturalize in the US every year[4], in the time period of 1960-2018.

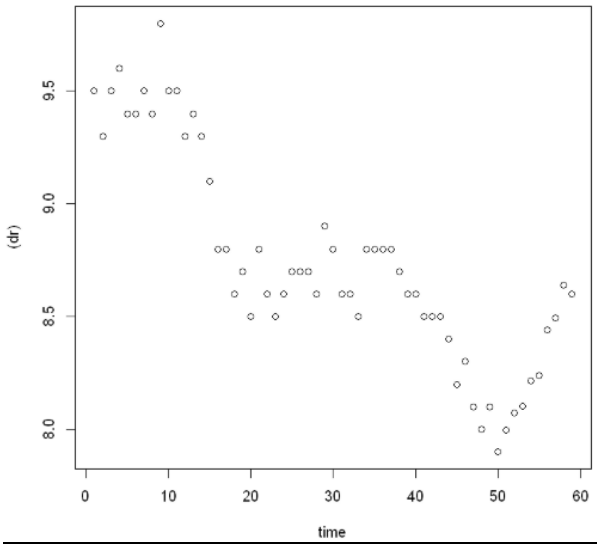
Analysis

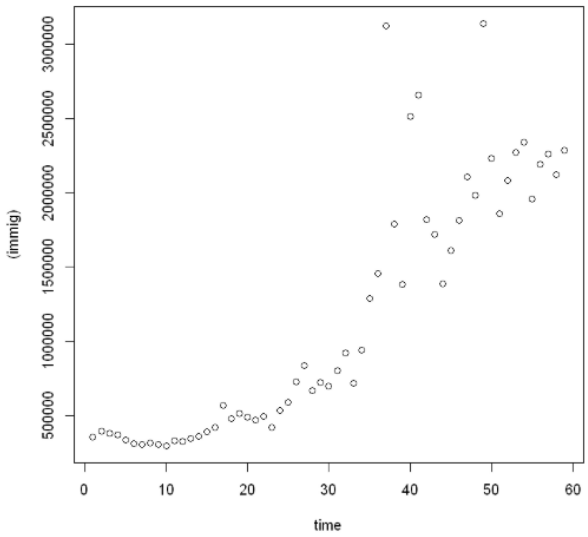
Firstly, there was obviously going to be a relationship with GDP and time. There didn't look to be a relationship between unemployment and time. And, lastly there did seem to be a relationship between death rate and time however it was difficult to see because it wasn't exactly linear or exponential. What did come as a surprise was that there was what seemed to be an exponential relationship between immigration and time. This all made the point that I should include time into my model since there seemed to be some definite trend effects.





[1] https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=US
[2] https://data.bls.gov/pdq/SurveyOutputServlet
[3] https://data.worldbank.org/indicator/SP.DYN.CDRT.IN?end=2018&locations=US&start=1960
[4] https://www.migrationpolicy.org/programs/data-hub/us-immigration-trends





However, before doing that I wanted to compare the summary statistics between the two models including and not including time and also run a best subsets on the model including time to see if time was truly necessary or if it was already taken into account by other factors.

```
Call:
lm(formula = immig ~ dr + gdp + unemp, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-444205 -197002  -58922  108294 1762087

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.645e+06  2.406e+06   1.099   0.2764
dr          -2.215e+05  2.404e+05  -0.922   0.3608
gdp          3.626e+01  5.543e+00   6.541 2.12e-08 ***
unemp       -7.782e+04  4.279e+04  -1.819   0.0744 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 381500 on 55 degrees of freedom
Multiple R-squared:  0.8067,    Adjusted R-squared:  0.7961
F-statistic: 76.51 on 3 and 55 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = immig ~ dr + gdp + unemp + time, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-444630 -197189  -59740  108615 1759230

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2590234.38 2857470.70   0.906   0.3687
dr          -216260.66  282937.20  -0.764   0.4480
gdp             35.68      16.81   2.123   0.0384 *
unemp        -77722.06   43273.08  -1.796   0.0781 .
time            746.14   20670.37   0.036   0.9713
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 385000 on 54 degrees of freedom
Multiple R-squared:  0.8067,    Adjusted R-squared:  0.7924
F-statistic: 56.34 on 4 and 54 DF,  p-value: < 2.2e-16
```

```
[55]: leaps(cbind(dr, gdp, time, unemp),immig,nbest=1)
```

$which

A matrix: 4 × 4 of type lgl

	1	2	3	4
1	FALSE	TRUE	FALSE	FALSE
2	FALSE	TRUE	FALSE	TRUE
3	TRUE	TRUE	FALSE	TRUE
4	TRUE	TRUE	TRUE	TRUE

$label

'(Intercept)' · '1' · '2' · '3' · '4'

$size

2 · 3 · 4 · 5

$Cp

2.40336810539101 · 1.83518651399228 · 3.00130301254244 · 4.99999999999999

The Cp between all the models are not that far off from each other. And it shows that in model with only one predicting variable gdp would be the best predictor and in a model with two predicting variables gdp and unemployment would be the best predictors and would also bring down the Cp a bit. Any more predicting variables would increase Cp.

This is supported by the summary statistics because in the model without time gdp has the strongest evidence of a relationship. However, when you include time it does not increase the strength of the relationship and in fact it brings down the p-value for gdp. This is probably because of the high correlation between gdp and time so it already includes most of the predicting power that time would add. However, unemployment is not affected nearly as much as gdp is which means that it does not hold that relationship with time that gdp does, which was also evidenced by the scatter plot and yet it still holds quite a bit of predicting power for immigration.

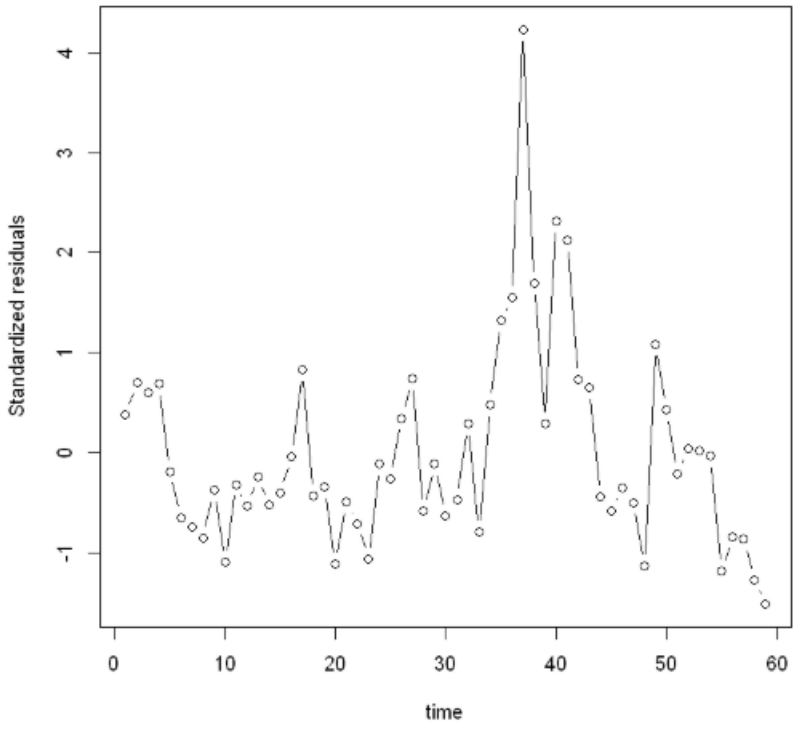
dr: 6.97735432369784 gdp: 38.0874329400764 unemp: 1.81948560910966 time: 49.3256332683859

Looking at the VIF scores of each variable further cements this notion because only the correlation between gdp and time is high.

However, clearly due to the long right-tail of immigration against time as well as gdp immigration must be logged.

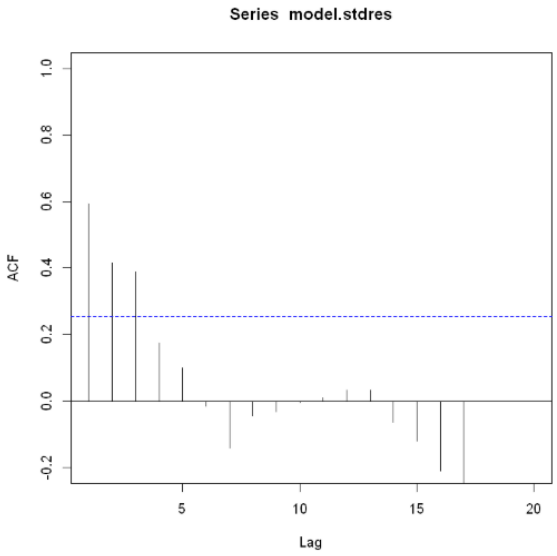
```
Call:
lm(formula = log(immig) ~ dr + gdp + unemp, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37113 -0.16035 -0.06864  0.11752  1.12410

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.715e+01  1.695e+00  10.115 3.70e-14 ***
dr          -4.424e-01  1.694e-01  -2.612   0.0116 *
gdp          3.047e-05  3.906e-06   7.802 1.83e-10 ***
unemp       -6.215e-02  3.015e-02  -2.061   0.0440 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2688 on 55 degrees of freedom
Multiple R-squared:  0.8878,    Adjusted R-squared:  0.8817
F-statistic: 145.1 on 3 and 55 DF,  p-value: < 2.2e-16
```

Once logged the strength of the relation does go up because R-squared is now .8878 but the p-value stays the same. However, now the interpretations of the regression coefficients is different. When the response variable is logged it means, all else constant that increasing x by one is associated with multiplying y by 10^{β}. So using gdp as an example increasing gdp by 1 is associated with a .007% increase in immigration. We can also see that once immigration is logged the other two predicting variables also become much more significant evidence of a strong relationship.



Plotting the standardized residuals over time of first glance seems like there could be some cyclical effect that could be autocorrelation.

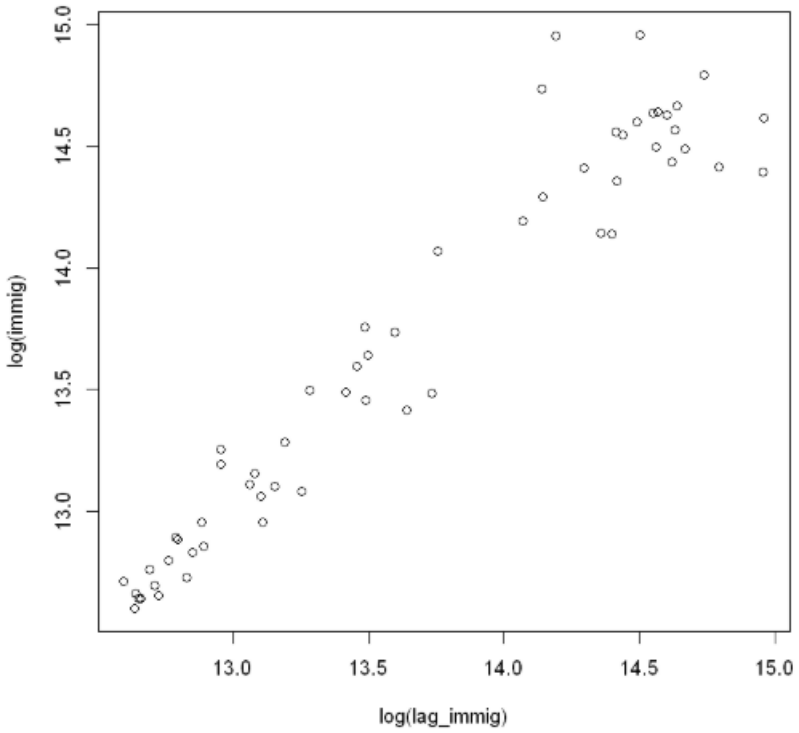
```
        Runs Test

data:  model.stdres
statistic = -4.1028, runs = 14, n1 = 22, n2 = 37, n = 59, p-value =
4.082e-05
alternative hypothesis: nonrandomness
```




```
        Durbin-Watson test

data:  log(immig) ~ dr + gdp + unemp
DW = 0.77434, p-value = 7.009e-09
alternative hypothesis: true autocorrelation is greater than 0
```

All three of the tests point to there being autocorrelation present.

So how do we tackle the autocorrelation. We could try adding time however, we know that because it is highly correlated with gdp that wouldn't be adding much predicting power. We can see if there is some lagging effect that is caused by one of these variables. That is also a fair assumption to make since people would not be able to already know the gdp, death rate, or uemployment rate of the year and decide to move there that year. They would have to be making that decision looking at stats about a year or two in advance. This is worsened by the fact that it is very hard to become a citizen in the United States. It could take 10-20 years for the demographic data of a certain year to convert into an immigration statistic.



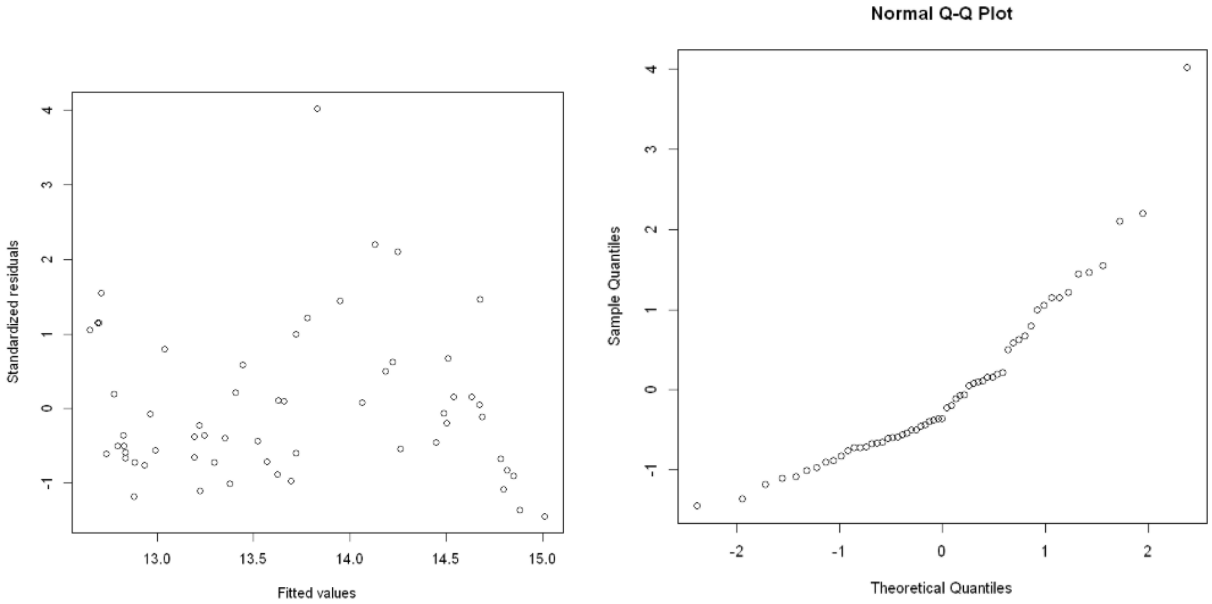
It seems as though the there is definitely a linear relationship between just lagging immigration by one year and immigration so that is definitely a way we could address autocorrelation.

```
Call:
lm(formula = log(immig) ~ dr + gdp + unemp + lag_immig, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3213 -0.1327 -0.0534  0.1022  1.0811

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.637e+01  1.583e+00  10.344 2.04e-14 ***
dr          -3.712e-01  1.579e-01  -2.351  0.02239 *
gdp          2.070e-05  4.695e-06   4.408 5.00e-05 ***
unemp       -4.718e-02  2.821e-02  -1.672  0.10023
lag_immig    2.768e-07  8.516e-08   3.251  0.00199 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2481 on 54 degrees of freedom
Multiple R-squared:  0.9062,    Adjusted R-squared:  0.8992
F-statistic: 130.4 on 4 and 54 DF,  p-value: < 2.2e-16
```

Adding in the lagged immigration also supports that hypothesis since it is significant.

The differences plot however does not provide the same results and so difference does not seem like it could help in this situation.

Just to check that there are not outliers, leverage values, or influential points we should check the fitted values vs. standardized residuals and the Q-Q plot.



These plots indicate that there isn't any nonconstant variance at least the kind that I have seen. And the Q-Q plot seems point out a right tail skew.