Motor Vehicle Fatalities in the U.S.

Jennifer Eberling

November 25, 2020

Abstract

This project's purpose is to model and predict motor vehicle fatalities in the United States of America. Multiple time series and indicators of historical events are used to estimate fatalities. Approaches using ARIMA, SARIMA, and regression models are used. The model that most accurately predicts fatalities in 2019 is an ARIMA model created with data from years after 1945. The average of the predicted values for the next five years is 36,270.888 fatalities. More data needs to be collected to compare the models' success.

Introduction

My primary time series is the number of motor vehicle fatalities in the United States of America. This is yearly data published by the National Highway Traffic Safety Administration (NHTSA) and can be found on Wikipedia, which is where I pulled it from [Motor Vehicle]. The NHTSA defines a motor vehicle as a vehicle "driven or drawn by mechanical power and manufactured primarily for use on public streets, roads, and highways, but does not include a vehicle operated only on a rail line." (Matthew). Therefore cars, busses, motorcycles, and trucks would be examples of vehicles included here, but jet skis, trains, and snowmobiles would not. The time series starts in the year 1899 and is published through 2018 (120 observations).

I also work with several secondary time series that start in various years but run through 2018. They were also collected from Wikipedia but were produced by the NHTSA [Motor Vehicle]. The time series (and year they begin) are US population (1900), vehicle miles traveled (VMT) in billions (1921), fatalities per 100 million VMT (1921), fatalities per 100,000 population (1900), change in per capita fatalities from

year to year (1901), number of motorcycle fatalities (1975), and number of bicycle fatalities (1980). Plots of the raw, original time series will be plotted in blue throughout this paper.

In doing this project, I want to explore how accurately I could model motorized vehicle deaths, determine whether there was any seasonal component to the yearly data, and make predictions for motor vehicle deaths for the next five years.

Looking at the Data

The main time series, motor vehicle fatalities, is shown in Figure 1. It begins with a sweeping incline for 32 years. This was a major growth period for the automobile. The number fluctuates for a few years, but rises again in the 1960's. There are a few notable decreases in fatalities: 1942, 1974, 1983, 1992, and 2010.

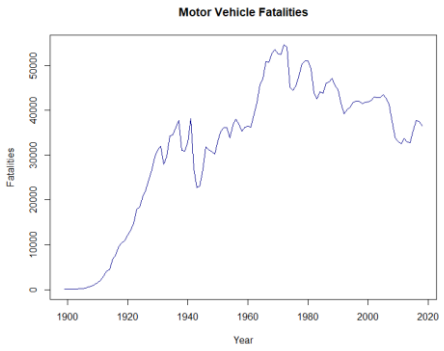
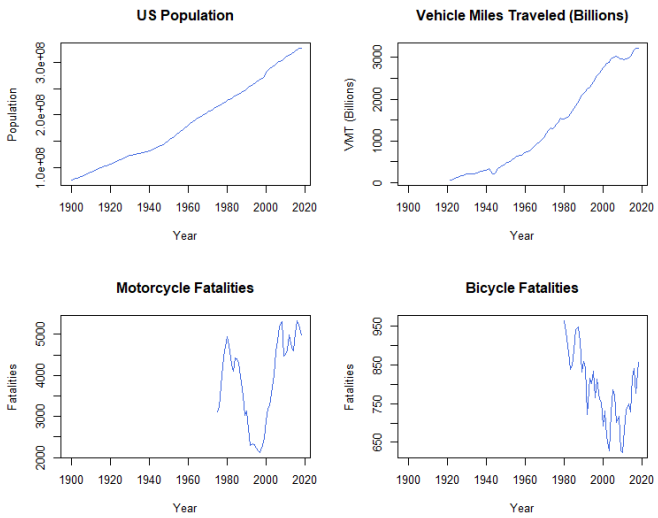


Figure 1

The first trough, in 1942, came right on the heels of the US entering WWII. With men sent abroad, the number of vehicle miles dropped, which explains the decrease in deaths. There is also a dip in VMT around 1974 and 1984. While the US oil embargo from 1973-1974 coupled with increased traffic legislation may explain that decrease, even experts have not been able to explain the dip in 1984 [Upi].



Figures 2 – 5

Events like WWII and the oil embargo had lasting effects. Because there was no clear amount of time for how long its effects were felt, I elected not to use an indicator. 2010 had a similar problem and did not have much data behind it, so I did not create an indicator. Both are also accounted for by VMT, so as long as that is included in the model, they are explained.

I did create two indicators that I thought may be likely to explain fatalities though. One ('seatbelt') is for when the Federal Motor Vehicle Safety Standard 208 was implemented in 1968. This put in place a standard for vehicle safety measures that were to be proved such as seatbelts and airbags [Federal]. The other ('drinking') indicates when the national drinking age was set to 21 years old [The 1984].

ARIMA and SARIMA Models

The first model I tried is an ARIMA model. There was no clear or rational seasonal pattern, so for this model I did not use one. Differencing two and three times made my motor vehicle fatality time series plot look more stationary at a glance, but the ACF and PACF plots showed that differencing once worked best.

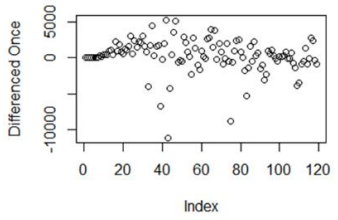


Figure 6

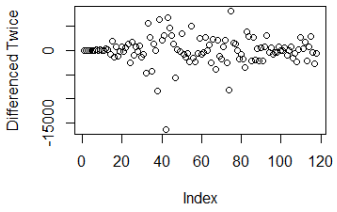


Figure 7

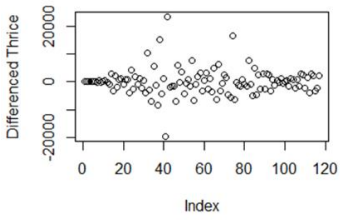


Figure 8

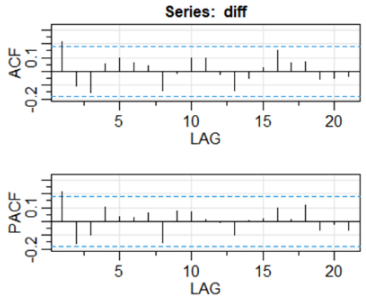


Figure 9

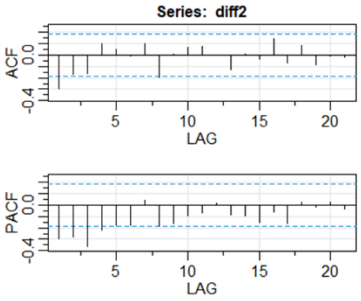


Figure 10

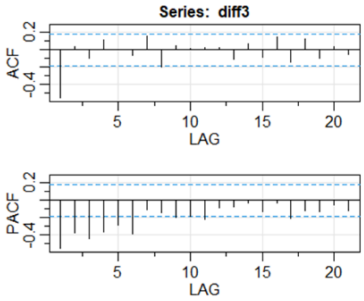


Figure 11

Using the AIC matrix on the differenced data, I chose the lowest AIC value (AIC = 18.33253) as the model to use. This was in cell [1,2], which means p=0 (the AR component) and q=1 (the MA component) is recommended by matrix.

```
         [,1]     [,2]     [,3]     [,4]     [,5]
[1,] 18.37560 18.33253 18.34861 18.33492 18.34778
[2,] 18.34379 18.34907 18.35987 18.34973 18.36036
[3,] 18.33498 18.34787 18.35285 18.36207 18.37664
[4,] 18.34220 18.35397 18.36301 18.37755 18.39344
[5,] 18.34854 18.36342 18.38021 18.38088 18.37889
```

Figure 12

Looking at the model (Figure 13), the standardized residuals are okay except for the large drop in 1943 and the small variance for the first 20 years of the 20th century. The normal qq plot of standardized residuals reflects that as well. The ACF looks like white noise and remains within the bounds as expected. There are a few prominent drops, but nothing too extreme, and they're not at consistent

intervals. Finally, the p-values for the Ljung-Box statistic are not significant. I concluded that the model SARIMA(deaths, 0,1,1,0,0,0,0) fits the data reasonably well.

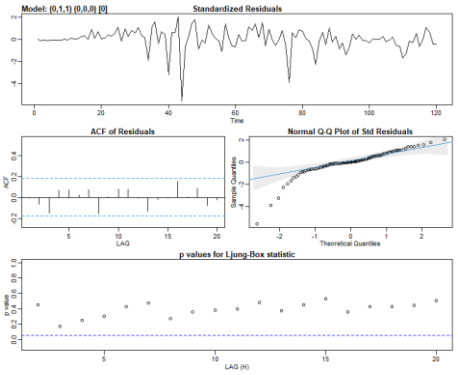


Figure 13

To try to make the residuals look better, I wanted to look at a shortened data set that excluded years where cars were not as popular, and variances fluctuated a lot. Looking at the standardized residuals above and the change in per capita fatalities from previous years (Figure 14), I decided to only look at years after 1945.

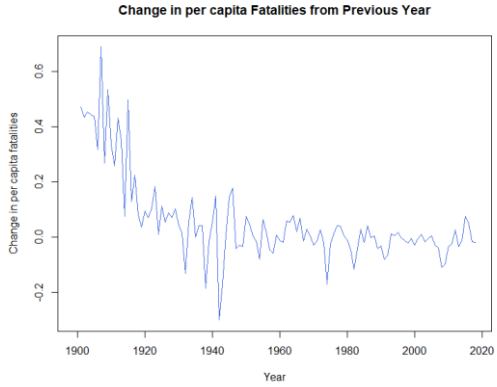


Figure 14

The new time series also requires differencing. The vector ‘newd’ (short for new deaths) contains the number of motor vehicle fatalities from 1946 – 2018 (n = 73). Differencing works extremely well and looks like white noise.

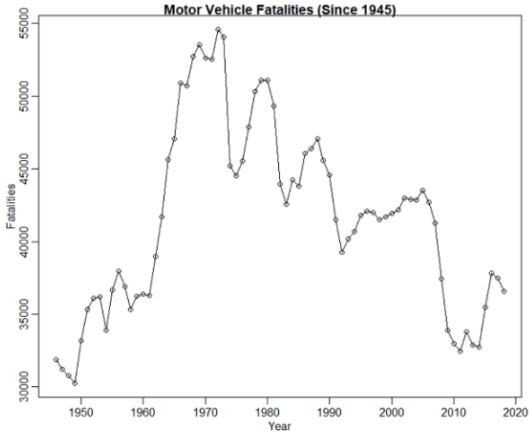


Figure 15

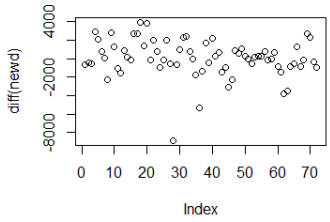


Figure 16

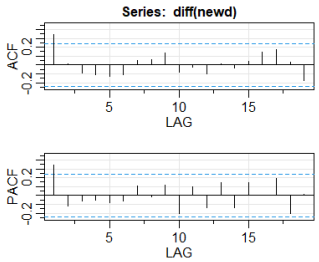


Figure 17

I again ran an AIC matrix, which recommended the same SARIMA(newd, 0,1,1,0,0,0,0) model. This AIC = 18.01259, which is lower than the AIC found using the full data set, suggesting that this model may be a better fit.

```
          [,1]     [,2]     [,3]     [,4]     [,5]
[1,] 18.11588 18.01259 18.03633 18.06162 18.08479
[2,] 18.01906 18.03738 18.06336 18.08167 18.10839
[3,] 18.03385 18.05075 18.07837 18.07351 18.05454
[4,] 18.05737 18.07842 18.03793 18.05220 18.05645
[5,] 18.08125 18.10463 18.06094 18.05756 18.13707
```

Figure 18

Looking at the output for this model (Figure 19), the standardized residuals look a lot better. The ACF of the residuals looks even more like white noise, which is great. The plot of standardized residuals shows that they're not quite normal, but they are very close.

I looked at models for various starting years (1920, 1935, 1950), and while all were very similar, this model has the highest p-values for the Ljung-Box statistic and the best looking ACF of residuals.

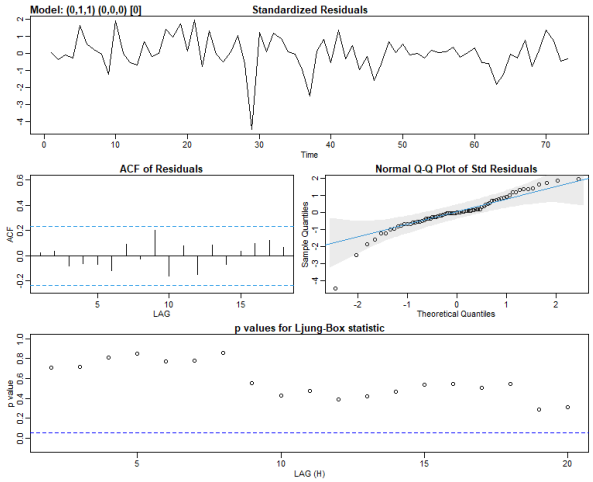


Figure 19

With a lower AIC and slightly better plots, I concluded that the model using only data from after 1945 fits a little better than the model fitted with the entire data set.

Next, I wanted to know if accounting for some form of seasonality could improve my full-data model. I began by creating a scaled periodogram that gave me a period of 1/120 (Figure 20). Because I used the full data set, I suspect the small numbers at the turn of the 20th century, when cars were not yet popular, coupled with a drop in 2010 (towards the end of the data), lead R to think that the cycle will be sinusoidal. I do not think that is reasonable, so I rejected the idea of a period so large.

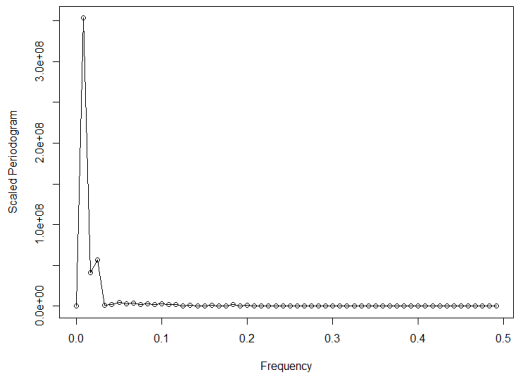


Figure 20

Based on the prominent low spikes in the ACF of the differenced data at lags 2, 8, and 13, I looked at all similar lags in the non-differenced data. After looking at ACF and PACFS with seasonal lags two through fourteen, the one that most looked like white noise was differenced at lag 7. Differencing as I did in the first model, with d = 1, was then appropriate to make the plots look even more like white noise.

I first ran the AIC matrix to find the seasonal AR and seasonal MA components. While the matrix could not complete its calculations, it did complete all of the low ordered models. Those are what I likely would have chosen anyway. It shows that the lowest AIC is with [1,2] means P = 0 (the seasonal AR component), Q = 1 (the seasonal MA component) is recommended by AIC.

```
         [,1]     [,2]     [,3]     [,4]     [,5]
[1,] 17.62726 17.57777 17.59472 17.57911 17.59460
[2,] 17.58852 17.59472 17.60781 17.59532 17.60392
[3,] 17.58138 17.59322 17.59232 17.61012 17.62084
[4,] 17.58458 17.59656 17.59661 17.62696 17.62524
[5,] 17.59129 17.60301 17.61996  0.00000  0.00000
```

Figure 21

Then I ran another AIC matrix to find the nonseasonal AR and MA components now accounting for the seasonal components. The lowest AIC is for the model correlating to [1,2], so p = 0, q = 1 is recommended by AIC.

```
         [,1]     [,2]     [,3]     [,4]     [,5]
[1,] 17.62726 17.57777 17.59472 17.57911 17.59460
[2,] 17.58852 17.59472 17.60781 17.59532 17.60392
[3,] 17.58138 17.59322 17.59232 17.61012 17.62084
[4,] 17.58458 17.59656 17.59661 17.62696 17.62524
[5,] 17.59129 17.60301 17.61996  0.00000  0.00000
```

Figure 22

Putting it all together results in the model SARIMA(deaths,0,1,1,0,1,1,7). However, the p-values for the Ljung-Box statistic are relatively low, and the ACF of residuals does not look any better. Adding a seasonal component did not eliminate the big drops and jumps every few lags. The standardized residuals and their normal qq plot are the same as the initial model. Seasonal differencing does not improve the model even though it gives a lower AIC value (AIC = 17.57777), so I did not proceed any further with it.

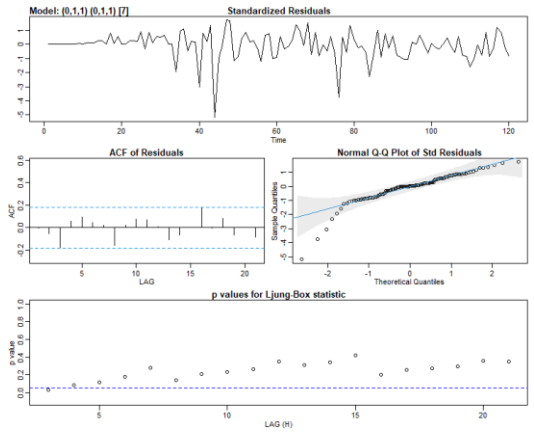


Figure 23

Cross-Correlation Functions

The primary use for the cross-correlation functions (CCFs) here is to give insight into how the time series relate to one another so that lagged variables can be chosen for a regression model. Extremely high correlations would indicate a lagged variable is appropriate for a regression model. Unchanged, nonstationary time series are used. The titles reflect the order variables are entered into the ccf(x,y) function. Most of the CCFs with high values at many lags have two highly correlated time series, as is the

case for population and VMT, or have a predictable relationship, like how an increase in Years after 1983 (drinking indicator) is most likely showing how the general trend is positive since 1983.

No lags were found to be useful. Using just one lagged variable would not allow a regression model to make predictions anyway, as it will be used alongside non-lagged variables, so there is no incentive to push for unnecessarily lagged variables.

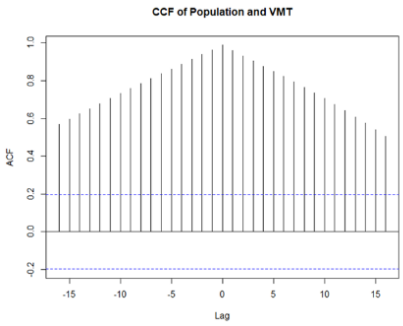


Figure 24

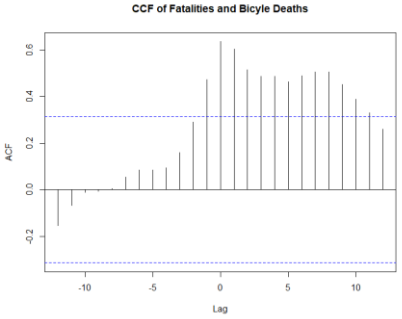


Figure 25

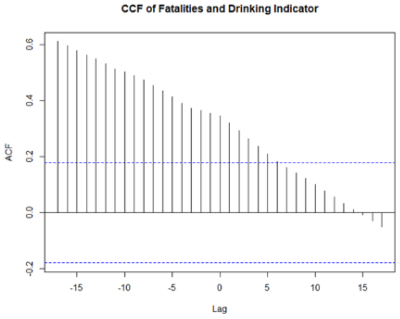


Figure 26

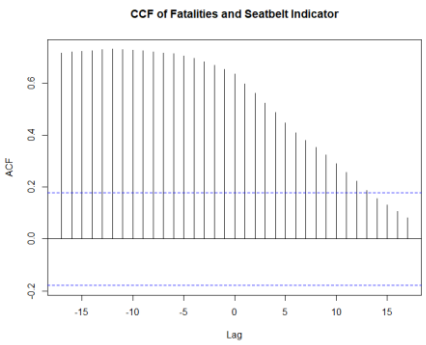


Figure 27

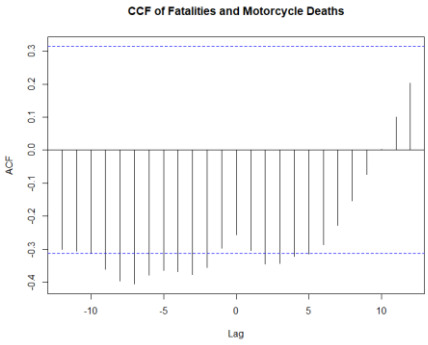


Figure 28

Regression Model with ARMA Errors

The final model I looked at was a regression model with ARMA errors. I examined numerous combinations of variables and settled on the model with year, VMT, motorcycle deaths, and bicycle deaths. The adjusted R squared value (0.9153) was very high, and all the variables are very significant.

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              2.589e+06  2.027e+05  12.773 1.57e-14 ***
fatalities$Year                         -1.317e+03  1.048e+02 -12.571 2.46e-14 ***
fatalities$`Vehicle miles traveled (billions)` 2.322e+01  2.259e+00  10.280 5.73e-12 ***
fatalities$Bicycle                       2.578e+01  3.716e+00   6.939 5.34e-08 ***
fatalities$Motorcycle                    1.524e+00  2.813e-01   5.416 4.95e-06 ***
```

Figure 29

'Year' is the most significant variable, and I found that it makes all the other variables much more significant when it's added. Population was significant on its own, but it is not in this model because it is extremely highly correlated with VMT (correlation = 0.9882541). The indicators 'seatbelt' and 'drinking' were significant until put into a model alongside bicycle and motorcycle.

The errors, as well as the squared errors and their absolute values, already looked like white noise. This suggests the errors are independent of each other, so it was not necessary to fit an ARMA model to them. The final fitted model is pictured in Figure 30. It begins in 1980, the first year that data for bicycle deaths was collected. The AIC for the model is 679.4463, which is much higher than previous models.

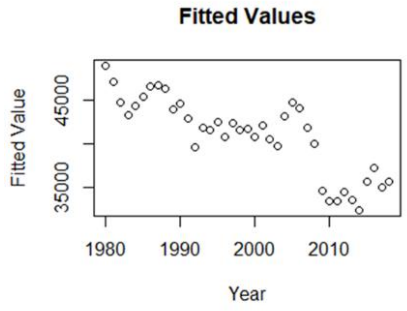


Figure 30

The model does not include any lagged variables, so it cannot make predictions into the future. For this reason, the model is not particularly useful. I did not proceed with it in my analysis.

Forecasting

I decided to move forward with the ARIMA models for the full data set as well as the one for data since 1945. Each model was used to estimate motor vehicle fatalities for the next five years. The only year that has complete data for so far is 2019 [Media]. 2020 still has another month of data to collect, but the numbers will be impacted by the coronavirus pandemic, as is noted in Table 1.

Table 1

Predicted Motor Vehicle Fatalities			
Year	Full Model	Shortened Model	Actual
2019	36,599.30	36,283.82	36,096
2020	36,845.39	36,224.60	COVID
2021	37,136.39	36,239.28	
2022	37,437.14	36,279.15	
2023	37,740.00	36,327.59	
Mean	37,151.644	36,270.888	

Figures 31 and 32 show the predictions and confidence intervals for both models.

Model with full data

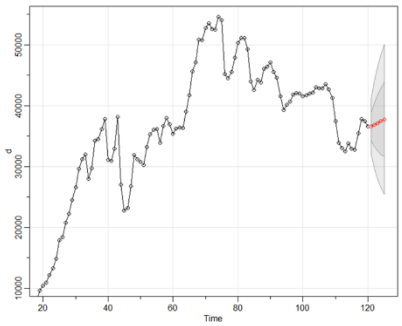


Figure 31

Model with shortened data

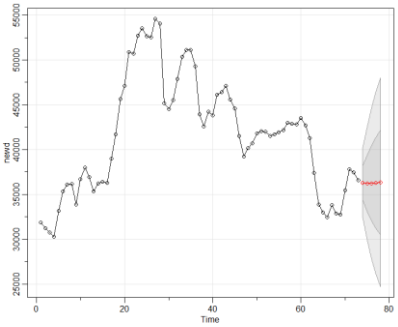


Figure 32

Accounting for the different y-axis scales pictured, the confidence intervals are about the same range for each model. The model from the full data predicts an upwards trend, while the model from the

shortened data predicts a decrease the first two years followed by three years of increasing fatalities. The shortened model is 315.48 fatalities closer to the actual number of motor vehicle fatalities in 2019.

The shortened model predicts another decrease in fatalities in 2020 though, and while the year is not over, experts estimate that the number will increase [Qureshi]. The coronavirus pandemic has decreased traffic and therefore decreased the overall number of accidents, but open roads lead to high speeds, meaning that the accidents that do happen are more serious.

Conclusions

The model based on the data after 1945 predicted motor vehicle fatalities in 2019 more accurately than the model created using the full data set, but the general positive trend predicted in the later model could still prove to be true. In five years, we can look back and compare the predictions to the actual numbers to gauge which model was ultimately more accurate. As it stands, the shortened model is the better choice.

Both ARIMA models had a lower AIC than the regression model, which is not useful without lagged variables. A SARIMA model differenced at lag 7 has the lowest AIC of the models looked at by a small margin, but the associated plots look worse. It does not make sense logically either, so the ARIMA models are the better choice.

One of the biggest changes to the world in recent years is the coronavirus pandemic which reached the US in early 2020. The two SARIMA models account for the pandemic's possible influence very well. Large confidence intervals make predictions much less precise, but the estimates for motor vehicle fatalities in 2019 are not very far from the actual values for either model. In a regression model, Year and VMT would account for some of the pandemic's affect, but not all of it.

I recommend future research on motor vehicle fatalities to be broken down by state. Doing so allows more specificity with law indicators, weather impact, and driving habits. Potentially some of the state-specific models may call for lagged variables in the regression model, in which case a regression model may be a better fit than an ARIMA model. This possibility may make some of the time series used, including the motor vehicle fatalities time series, more stationary. If so, it could lead to a more accurate model.

Sources

"The 1984 National Minimum Drinking Age Act." *National Institute on Alcohol Abuse and Alcoholism*, U.S. Department of Health and Human Services, alcoholpolicy.niaaa.nih.gov/the-1984-national-minimum-drinking-age-act.

Federal Motor Vehicle Safety Standards and Regulations, U.S. Department of Transportation, icsw.nhtsa.gov/cars/rules/import/FMVSS/.

Matthew.lynberg.ctr@dot.gov. "07-007541as." NHTSA, National Highway Traffic Safety Administration, 1 Sept. 2018, www.nhtsa.gov/interpretations/07-007541as.

Media, NHTSA. "2019 Fatality Data Show Continued Annual Decline in Traffic Deaths." *NHTSA*, NHTSA, 1 Oct. 2020, www.nhtsa.gov/press-releases/2019-fatality-data-traffic-deaths-2020-q2-projections.

"Motor Vehicle Fatality Rate in U.S. by Year." *Wikipedia*, Wikimedia Foundation, 16 Nov. 2020, en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year.

Qureshi, Adnan I, et al. "Mandated Societal Lockdown and Road Traffic Accidents." *Accident; Analysis and Prevention*, Elsevier Ltd., Oct. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7475733/.

Upi. "Traffic Fatalities Drop 11%; The Experts Don't Know Why." *The New York Times*, The New York Times, 27 Sept. 1983, www.nytimes.com/1983/09/27/us/traffic-fatalities-drop-11-the-experts-don-t-know-why.html.