The Growing Toll: Analyzing Drug Overdose Deaths in U.S America.

Group Members: Allison Asbock, Elizabeth Binkina, Hersh Doshi

Contributions:
Elizabeth - Wrote intro, preliminary analysis, implemented forecasting models, residual diagnostics, and some results.
Allison - Fit regression models and wrote the discussion section.
Hersh - Wrote the abstract, results, and code to format output.

Abstract

Due to the rising rates of drug overdose deaths in the United States, our goal is to find how different factors such as age, sex, and drug types can affect overdose rates. Initially, we analyzed the trends of the time series data for overall overdose deaths for all subgroups and drug types. To find the influence of these factors on overdose rates, we fit two ordinary least squares regressions that use age, sex, and drug type. The first model is a main effects model, as the second model considers all the pairwise interactions for every combination of age/sex group by drug category. We will be using the coefficients of the main effects model to find out how age and sex influence the rate of drug overdose deaths, and if there are specific age groups or sexes that show significantly higher overdose rates since the coefficients will tell us how the rates differ from the baseline group of 15 to 24 years old. Furthermore, the main effects model will also tell us which opioid categories carry the highest overdose rates. Lastly, the interaction model can be utilized to find if specific combinations of drug type and groups account for extra risk in overdose rates.

Introduction

Motivation: The rising rates of drug overdose deaths in the United States represent a growing public health crisis with widespread societal consequences. Understanding the features driving these growing trends is essential for shaping potential intervention and prevention strategies. This project is motivated by the urgency of analyzing how drug overdose death rates have changed over time and identifying what the main contributing factors are. By examining data from the CDC's National Center for Health Statistics, we aim to uncover critical insights as to how age, sex, race, and drug types influence overdose death rates. With regression analysis, we are able to quantify and observe these relationships and assess significant contributing features. Furthermore, by modeling these trends, we can forecast future overdose death rates and potentially inform policy-making decisions and public health responses aiming to mitigate this epidemic.

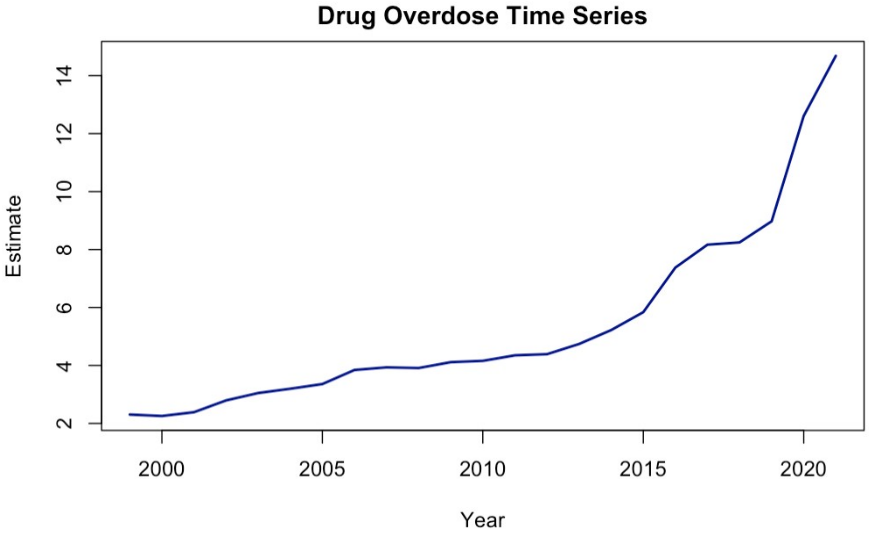
Goal: The goal of this project is to analyze trends in drug overdose death rates in the United States from 1999 to 2021. Specifically, we aim to address the following key research questions:

1. **Trend Analysis:** How the rate of drug overdose deaths per 100,000 residents changed over time from 1999 to 2021, and what type of trend it follows.
2. **Demographic Impact**: How do age and sex influence the rate of drug overdose deaths, and if there are specific age groups or sexes that show significantly higher overdose rates?
3. **Drug Type Influence:** Which drug types are most strongly associated with increases in overdose death rates, and how has their impact changed over the duration of the study?
4. **Interaction Effects:** Is there significant interaction effects between age, sex, and drug type that contribute to higher overdose death rates, for example, if young males are more likely to overdose on opioids compared to other groups?
5. **Forecasting:** Based on our model, what are the forecasted drug overdose death rates for the next five years (2022–2026)?

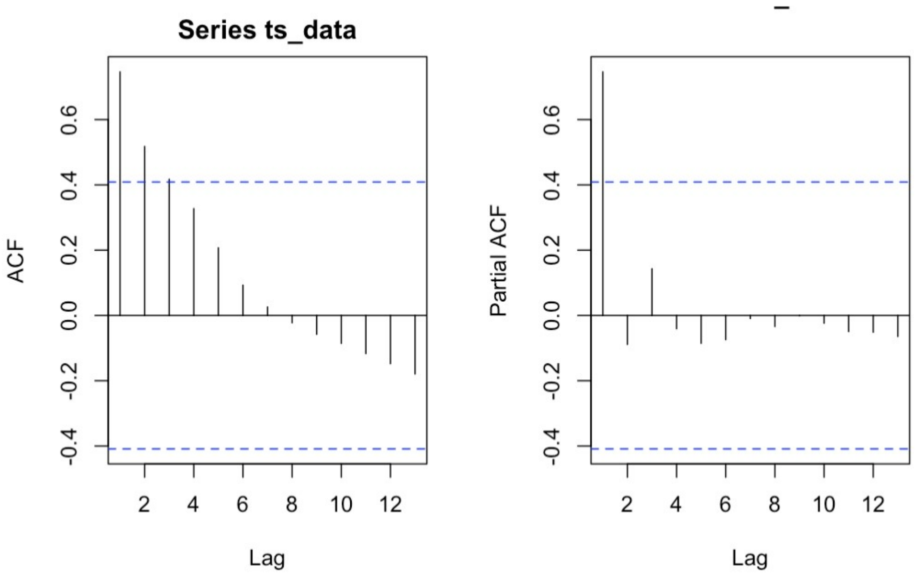
Data: The data was sourced by the National Center for Health Statistics (NCHS) division of Analysis and Epidemiology and is publicly available on the CDC website. The dataset we used consists of annual estimates of drug overdose records from 1999 to 2021, tracking drug overdose deaths per 100,000 residents across the United States. This data includes a variety of variables such as

Year: the year of data collection (1999-2021)
Drug Type: Specific drug responsible for the overdose
Age: The age group of individuals who experienced overdose deaths.
Sex: The sex of individuals who experienced overdose deaths.

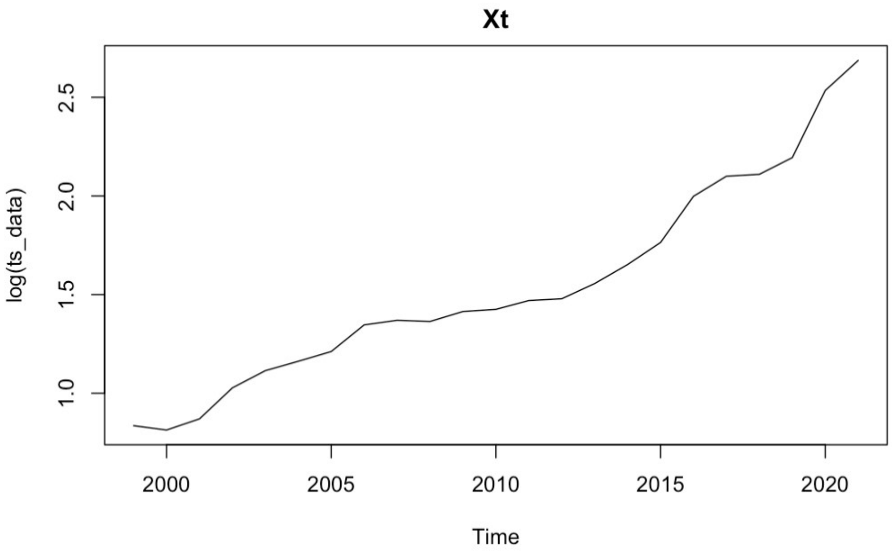
Preliminary Analysis



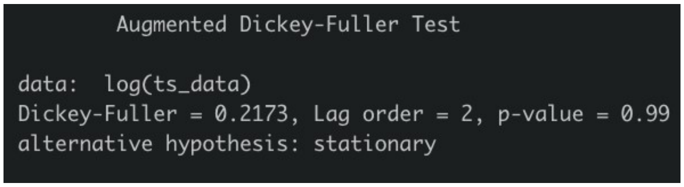
The nature of the data suggests that with each year, drug overdose only increases gradually from 2000-2010, and increases exponentially after 2015.



ACF plot tails off and PACF lag cuts off at 1, suggesting an ARMA(0,1) or MA(1) model. However, the time series plot shows that differencing is required. Therefore, the ARIMA model could be a better fit for the data.

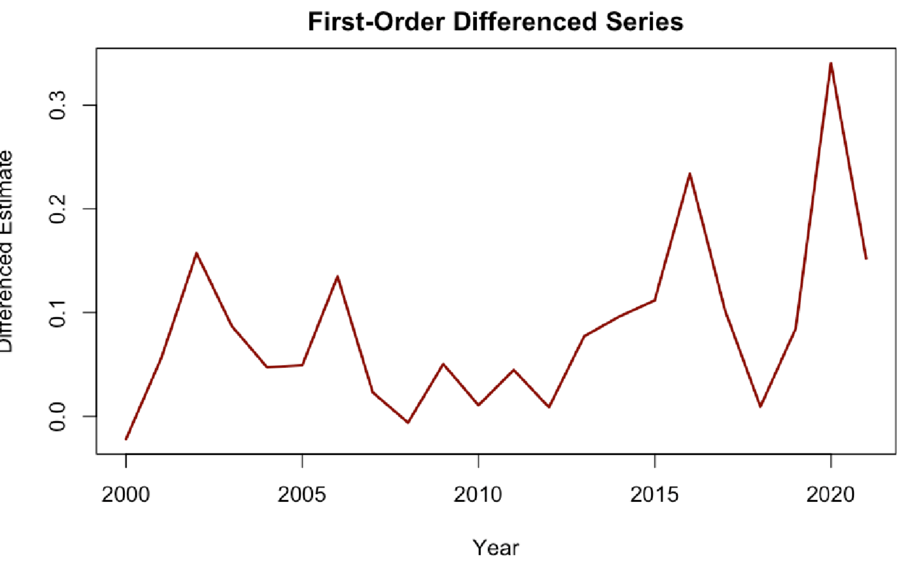


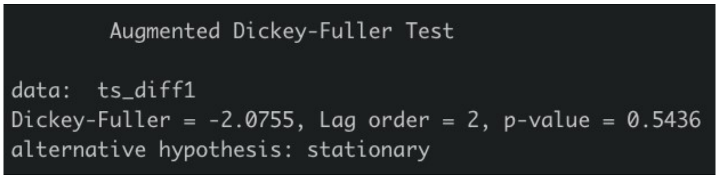
After **log transformation**, differencing is still required because it is not stationary. There is still an upward trend, however, variance seems to not be increasing.



```
        Augmented Dickey-Fuller Test

data:  log(ts_data)
Dickey-Fuller = 0.2173, Lag order = 2, p-value = 0.99
alternative hypothesis: stationary
```

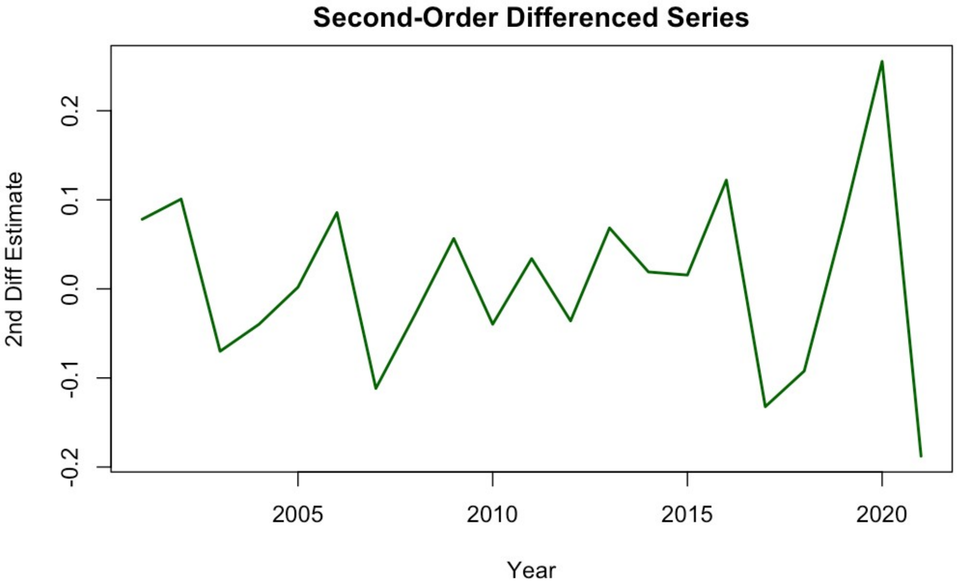
ADF test shows log model is not stationary ($p > 0.05$), and therefore further differencing methods are required.

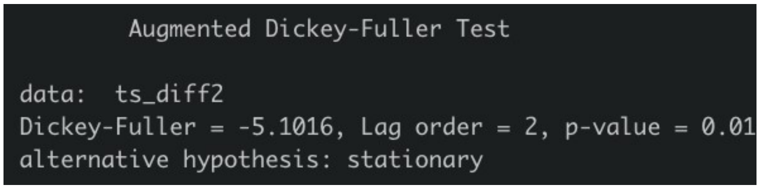




```
        Augmented Dickey-Fuller Test

data:  ts_diff1
Dickey-Fuller = -2.0755, Lag order = 2, p-value = 0.5436
alternative hypothesis: stationary
```

First-order differencing is still not stationary, and the ADF test shows a p-value < 0.05, and therefore requires differencing a second time.





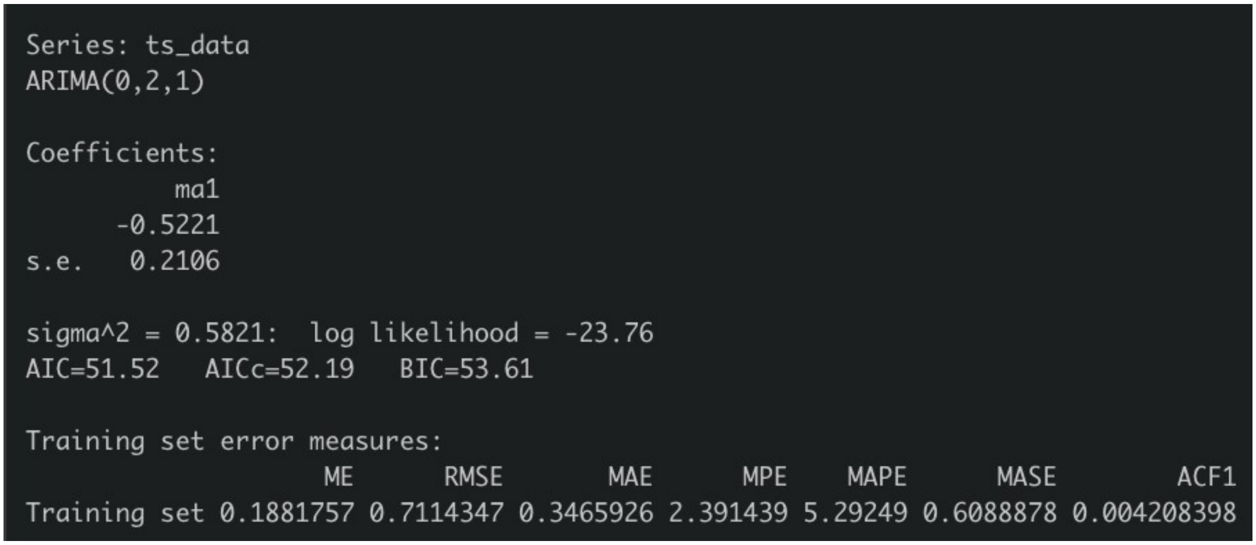
```
        Augmented Dickey-Fuller Test

data:  ts_diff2
Dickey-Fuller = -5.1016, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

The **Second-Order differenced** model is now stationary because the ADF test shows a p-value < 0.05.

Original Time Series Model: ACF plot tails off gradually, and the PACF plot cuts off sharply after lag 1. This pattern suggests an AR(1) process, or more generally, an ARMA(1,0) model would be most appropriate. However, the time series is non-stationary and requires second-order differencing to achieve stationarity. After differencing twice, the appropriate model becomes an ARIMA(0,2,1), which includes no autoregressive terms, two differences to induce stationarity, and one moving average term to capture short-term correlation in the differenced series.

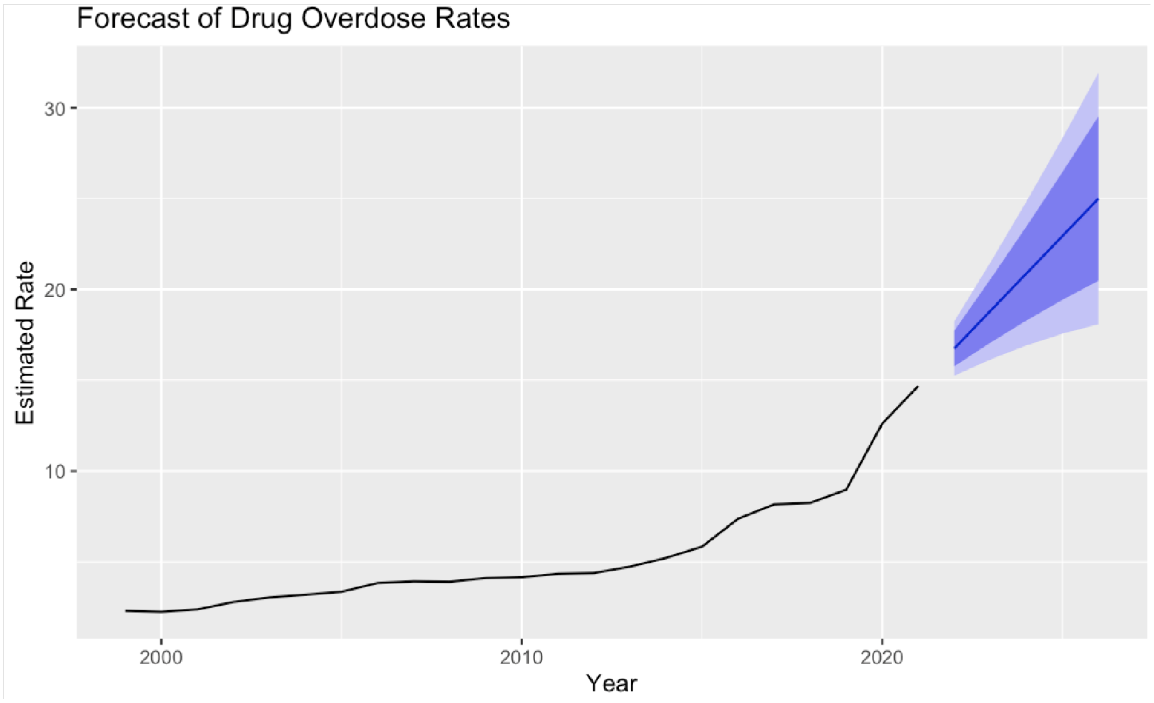
Statistical Methods

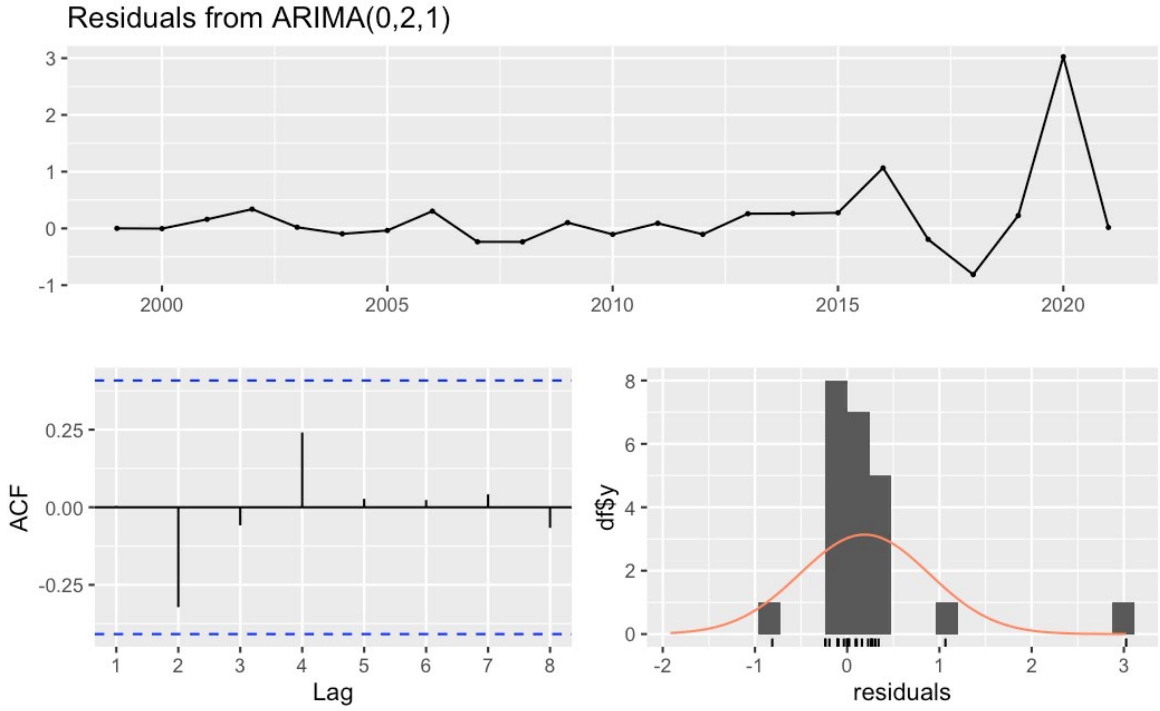
Forecasting Next 5 Years of Drug Overdose Rates with ARIMA(0,2,1) Model



```
Series: ts_data
ARIMA(0,2,1)

Coefficients:
          ma1
      -0.5221
s.e.   0.2106

sigma^2 = 0.5821:  log likelihood = -23.76
AIC=51.52   AICc=52.19   BIC=53.61

Training set error measures:
                    ME      RMSE       MAE      MPE    MAPE      MASE        ACF1
Training set 0.1881757 0.7114347 0.3465926 2.391439 5.29249 0.6088878 0.004208398
```

Model Summary: The MA(1) coefficient (-0.5221 with standard error 0.21) is significant, implying the model captures short-term shocks well. The log-likelihood, AIC, and BIC are all acceptable. Residual ACF1 ≈ 0.0042, meaning that the residuals are essentially uncorrelated and the model fits well.





Top Plot: Residuals Over Time

Residuals appear randomly scattered around 0, with no clear pattern. Some higher residuals post-2020, but overall randomness suggests the model captures the structure well.

Bottom Left: ACF of Residuals

All autocorrelations are within the confidence bounds, meaning no significant autocorrelation is left. This confirms the residuals are white noise, which is a good sign.

Bottom Right: Histogram of Residuals

Light skew, but residuals roughly follow a normal distribution. The red density line floors the bars fairly well, and there are no extreme outliers. Residuals are approximately uncorrelated and normally distributed, meeting assumptions for ARIMA modeling.

Regressions

It is clear from the preliminary analysis of the data that the amount of death from overdoses does not follow a simple linear trend. In order to fit the data better a quadratic time component was added to the regressions that were run on the data. A cubic regression was also run, but the coefficient was found to be non-significant due to collinearity, therefore the quadratic model was chosen.

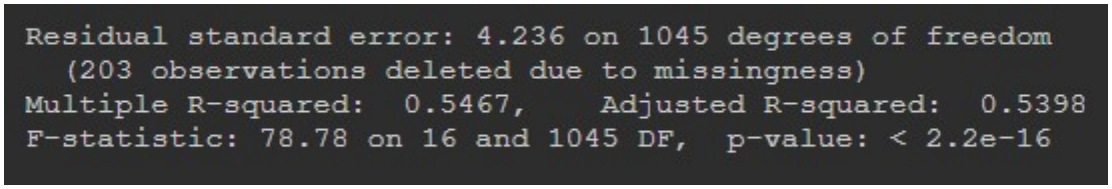
Model 1:

The first regression is a quadratic model with three predictor variables: age group, sex, and drug type. This simple model looks at the three categorical variables and shows their individual effects on the number of overdose deaths per 100,000.

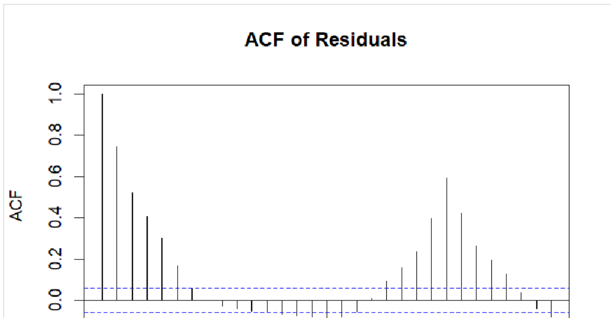
$$Y_t = \beta_0 + \beta_1 Time + \beta_2 Time^2 + \beta_3 AgeGroup + \beta_4 Sex + \beta_5 DrugType$$

All of the coefficients in this model are significant other than the coefficient for the female dummy variable. A possible explanation for this is that female is used as a reference category and is therefore explained in the value of the intercept. The other references, which are explained by the intercept are the 15–24-year-old age group and any death involving opioids.

It is unsurprising that the three predictors chosen for this model have a significant effect on the number of overdoses. The age of an individual is closely tied to health and also how likely they are to access certain types of drugs. In addition, drug type is unsurprisingly significant. Certain types of drugs pose more danger to an individual and can also be more prevalent.



```
Residual standard error: 4.236 on 1045 degrees of freedom
  (203 observations deleted due to missingness)
Multiple R-squared:  0.5467,    Adjusted R-squared:  0.5398
F-statistic: 78.78 on 16 and 1045 DF,  p-value: < 2.2e-16
```



The R-squared value for the first model is 0.5467, which means that it explains most of the variance in the data. This model, although not perfect, is a good starting point to explain the amount of drug overdoses per 100,000 people.

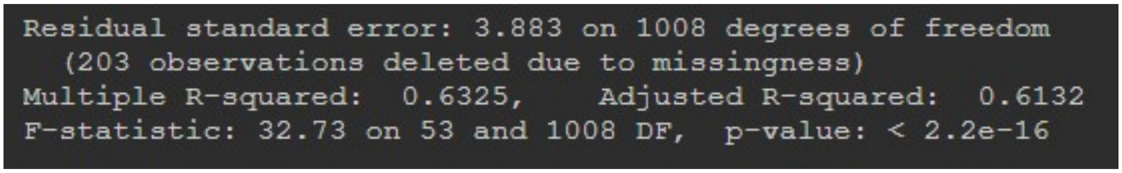
This ACF plot here shows that the residuals do seem to gradually decline, however they rise again. This could mean that the error terms are autocorrelated.

Model 2:

The second regression fit to the data uses the same predictors as the first but looks at the interaction between drug type and different characteristics of an individual. This adds two additional coefficients that predict the number of deaths per 100,000 residents from drug overdoses.

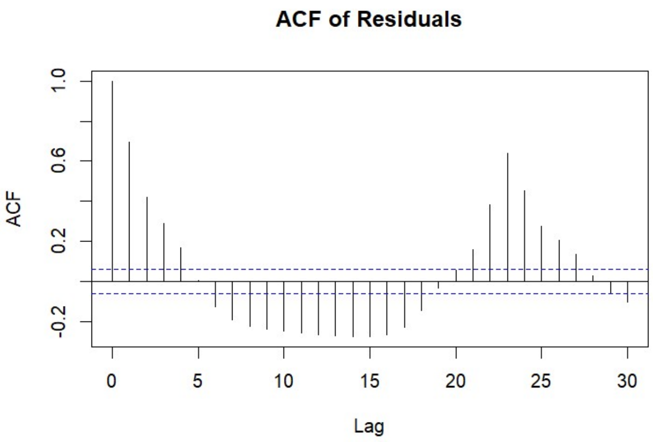
$$Y_t = \beta_0+\beta_1 Time+\beta_2 Time^2+\beta_3 AgeGroup+\beta_4 Sex+\beta_5 DrugType+\beta_6 AgeGroup : DrugType+\beta_7 Sex : DrugType$$

This regression model, although it only has 7 predictor coefficients in total, has many dummy variables due to the number of categorical variables in the data set. This more complicated model gives better insight into the groups of individuals who die from overdoses at higher rates. More specifically it shows whether specific ages or sexes are more likely to overdose on specific drug types. These interaction terms can be helpful in seeing which drugs are particularly deadly to certain age groups.



```
Residual standard error: 3.883 on 1008 degrees of freedom
  (203 observations deleted due to missingness)
Multiple R-squared:  0.6325,    Adjusted R-squared:  0.6132
F-statistic: 32.73 on 53 and 1008 DF,  p-value: < 2.2e-16
```

Since there are many different age groups and drug types of the output of this regression is large and it is clear that there are some unnecessary predictors. Many of the interaction terms do not seem to be significant, although some are. However, this more complicated model is a better fit for the data than the simpler model above, and it explains more of the variance in the model than the one that does contain the interactions. This can be seen from the higher R-squared value in this model. However, there are more non-significant predictors in the second model. Despite having some unnecessary predictors, it is still a better model than the simple one.



This ACF plot shows that the residuals are likely autocorrelated, since the ACF seems to gradually decline. Fitting an auto-regressive model to this data could potentially solve this problem. However, the quadratic model is still an overall good choice to explain the data, and the high R-squared value shows that it fits the data well.

Results

Trend Analysis: How the rate of drug overdose deaths per 100,000 residents changed over time from 1999 to 2021, and what type of trend it follows.

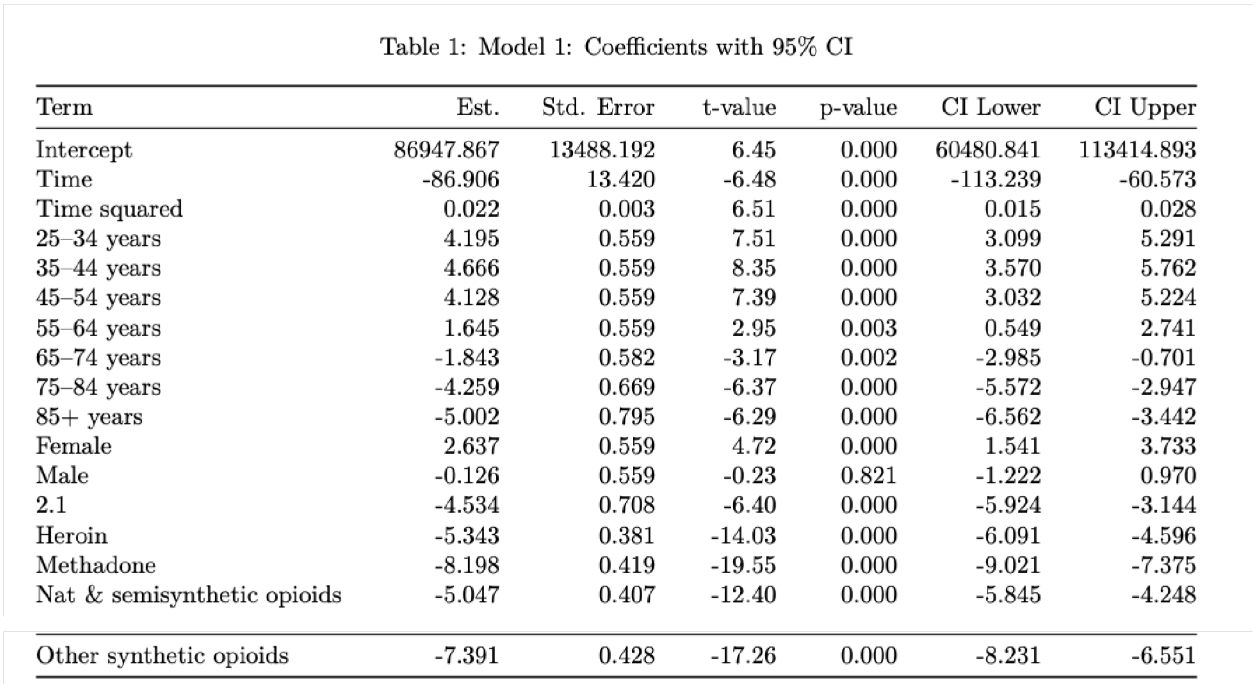


Table 1: Model 1: Coefficients with 95% CI

Term	Est.	Std. Error	t-value	p-value	CI Lower	CI Upper
Intercept	86947.867	13488.192	6.45	0.000	60480.841	113414.893
Time	-86.906	13.420	-6.48	0.000	-113.239	-60.573
Time squared	0.022	0.003	6.51	0.000	0.015	0.028
25–34 years	4.195	0.559	7.51	0.000	3.099	5.291
35–44 years	4.666	0.559	8.35	0.000	3.570	5.762
45–54 years	4.128	0.559	7.39	0.000	3.032	5.224
55–64 years	1.645	0.559	2.95	0.003	0.549	2.741
65–74 years	-1.843	0.582	-3.17	0.002	-2.985	-0.701
75–84 years	-4.259	0.669	-6.37	0.000	-5.572	-2.947
85+ years	-5.002	0.795	-6.29	0.000	-6.562	-3.442
Female	2.637	0.559	4.72	0.000	1.541	3.733
Male	-0.126	0.559	-0.23	0.821	-1.222	0.970
2.1	-4.534	0.708	-6.40	0.000	-5.924	-3.144
Heroin	-5.343	0.381	-14.03	0.000	-6.091	-4.596
Methadone	-8.198	0.419	-19.55	0.000	-9.021	-7.375
Nat & semisynthetic opioids	-5.047	0.407	-12.40	0.000	-5.845	-4.248
Other synthetic opioids	-7.391	0.428	-17.26	0.000	-8.231	-6.551

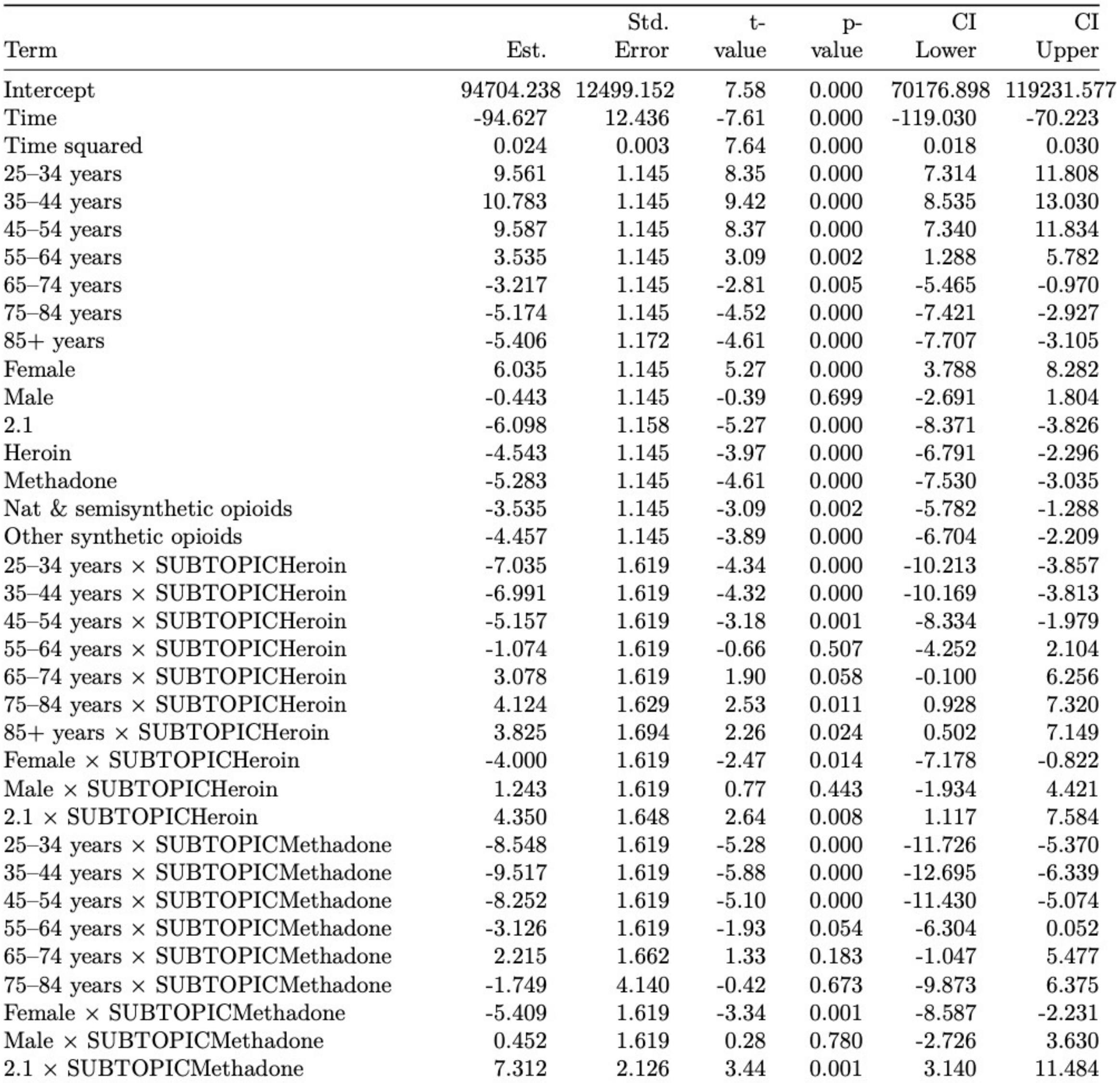
Demographic Impact: How do age and sex influence the rate of drug overdose deaths, and if there are specific age groups or sexes that show significantly higher overdose rates?

Looking at the coefficient estimates for the main effects model fitted for age, sex, and drug type, we can see that middle-aged adults have significantly higher overdose death rates. From the table, age groups 25 to 34, 34 to 44, and 45 to 54 all have significantly higher estimated rates than the baseline age group of 15 to 24 years old, with estimates of 4.195, 4.666, and 4.128, respectively. Furthermore, older age groups like seniors, specifically 65 + have much lower overdose death rates, as we can see the estimates are negative, such as -1.843, -4.259, and -5.002. We can also see that females have a significantly higher.

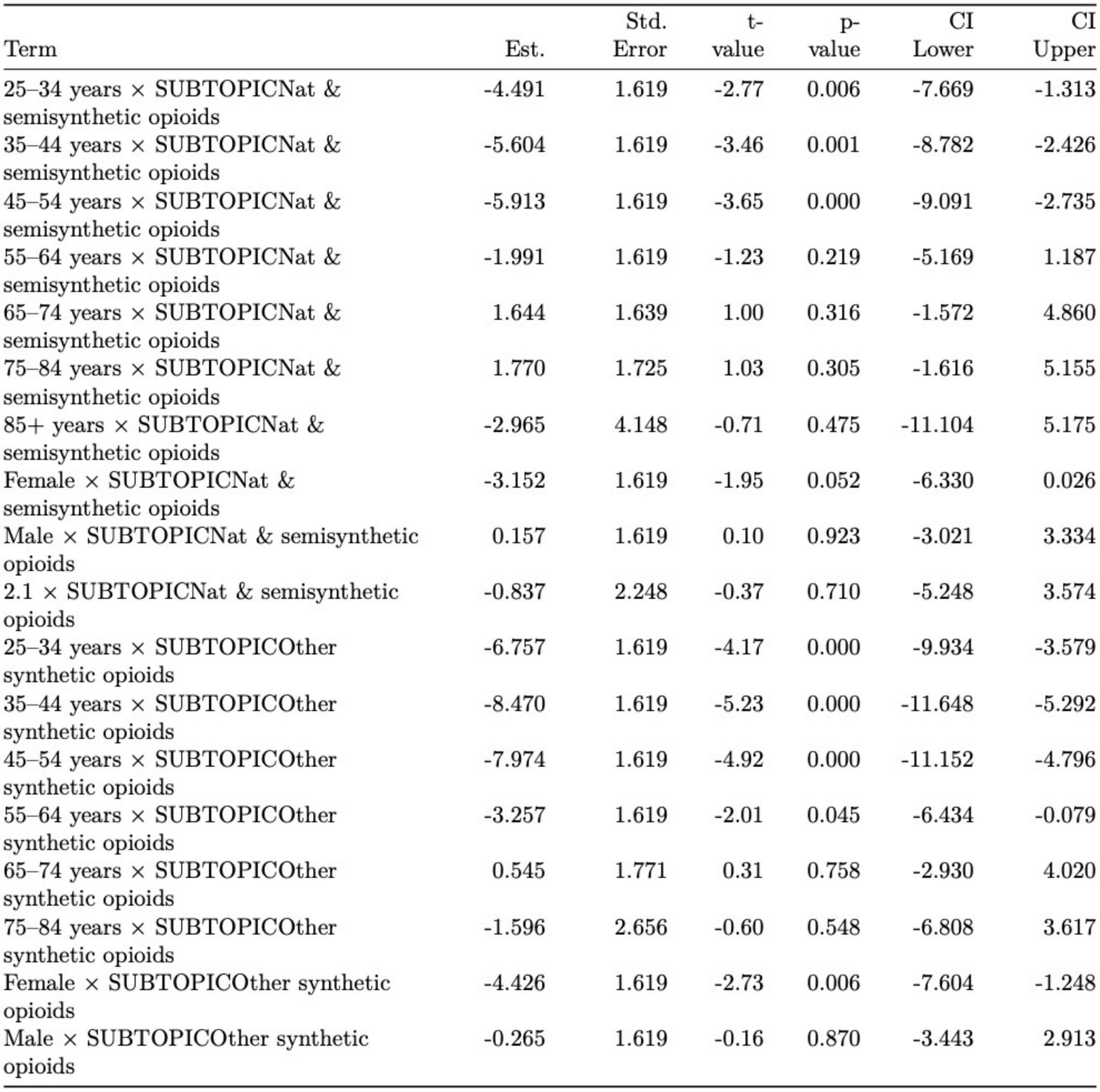
Drug Type Influence: Which drug types are most strongly associated with increases in overdose death rates, and how has their impact changed over the duration of the study?

Because all four subclass coefficients are negative, technically, “All opioids” has the highest baseline rate. However, among the subclasses, heroin shows the smallest reduction (−5.343), so it’s the highest‑risk subclass, followed closely by nat & semisynthetic. Furthermore, in both the models, there is a common Time + Time squared trend across all drug types, so these rank‑order gaps will stay constant over the study period.

Table 2: Model 2: All Coefficients (incl. interactions) with 95% CI



Term	Est.	Std. Error	t-value	p-value	CI Lower	CI Upper
Intercept	94704.238	12499.152	7.58	0.000	70176.898	119231.577
Time	-94.627	12.436	-7.61	0.000	-119.030	-70.223
Time squared	0.024	0.003	7.64	0.000	0.018	0.030
25–34 years	9.561	1.145	8.35	0.000	7.314	11.808
35–44 years	10.783	1.145	9.42	0.000	8.535	13.030
45–54 years	9.587	1.145	8.37	0.000	7.340	11.834
55–64 years	3.535	1.145	3.09	0.002	1.288	5.782
65–74 years	-3.217	1.145	-2.81	0.005	-5.465	-0.970
75–84 years	-5.174	1.145	-4.52	0.000	-7.421	-2.927
85+ years	-5.406	1.172	-4.61	0.000	-7.707	-3.105
Female	6.035	1.145	5.27	0.000	3.788	8.282
Male	-0.443	1.145	-0.39	0.699	-2.691	1.804
2.1	-6.098	1.158	-5.27	0.000	-8.371	-3.826
Heroin	-4.543	1.145	-3.97	0.000	-6.791	-2.296
Methadone	-5.283	1.145	-4.61	0.000	-7.530	-3.035
Nat & semisynthetic opioids	-3.535	1.145	-3.09	0.002	-5.782	-1.288
Other synthetic opioids	-4.457	1.145	-3.89	0.000	-6.704	-2.209
25–34 years × SUBTOPICHeroin	-7.035	1.619	-4.34	0.000	-10.213	-3.857
35–44 years × SUBTOPICHeroin	-6.991	1.619	-4.32	0.000	-10.169	-3.813
45–54 years × SUBTOPICHeroin	-5.157	1.619	-3.18	0.001	-8.334	-1.979
55–64 years × SUBTOPICHeroin	-1.074	1.619	-0.66	0.507	-4.252	2.104
65–74 years × SUBTOPICHeroin	3.078	1.619	1.90	0.058	-0.100	6.256
75–84 years × SUBTOPICHeroin	4.124	1.629	2.53	0.011	0.928	7.320
85+ years × SUBTOPICHeroin	3.825	1.694	2.26	0.024	0.502	7.149
Female × SUBTOPICHeroin	-4.000	1.619	-2.47	0.014	-7.178	-0.822
Male × SUBTOPICHeroin	1.243	1.619	0.77	0.443	-1.934	4.421
2.1 × SUBTOPICHeroin	4.350	1.648	2.64	0.008	1.117	7.584
25–34 years × SUBTOPICMethadone	-8.548	1.619	-5.28	0.000	-11.726	-5.370
35–44 years × SUBTOPICMethadone	-9.517	1.619	-5.88	0.000	-12.695	-6.339
45–54 years × SUBTOPICMethadone	-8.252	1.619	-5.10	0.000	-11.430	-5.074
55–64 years × SUBTOPICMethadone	-3.126	1.619	-1.93	0.054	-6.304	0.052
65–74 years × SUBTOPICMethadone	2.215	1.662	1.33	0.183	-1.047	5.477
75–84 years × SUBTOPICMethadone	-1.749	4.140	-0.42	0.673	-9.873	6.375
Female × SUBTOPICMethadone	-5.409	1.619	-3.34	0.001	-8.587	-2.231
Male × SUBTOPICMethadone	0.452	1.619	0.28	0.780	-2.726	3.630
2.1 × SUBTOPICMethadone	7.312	2.126	3.44	0.001	3.140	11.484



Term	Est.	Std. Error	t-value	p-value	CI Lower	CI Upper
25–34 years × SUBTOPICNat & semisynthetic opioids	-4.491	1.619	-2.77	0.006	-7.669	-1.313
35–44 years × SUBTOPICNat & semisynthetic opioids	-5.604	1.619	-3.46	0.001	-8.782	-2.426
45–54 years × SUBTOPICNat & semisynthetic opioids	-5.913	1.619	-3.65	0.000	-9.091	-2.735
55–64 years × SUBTOPICNat & semisynthetic opioids	-1.991	1.619	-1.23	0.219	-5.169	1.187
65–74 years × SUBTOPICNat & semisynthetic opioids	1.644	1.639	1.00	0.316	-1.572	4.860
75–84 years × SUBTOPICNat & semisynthetic opioids	1.770	1.725	1.03	0.305	-1.616	5.155
85+ years × SUBTOPICNat & semisynthetic opioids	-2.965	4.148	-0.71	0.475	-11.104	5.175
Female × SUBTOPICNat & semisynthetic opioids	-3.152	1.619	-1.95	0.052	-6.330	0.026
Male × SUBTOPICNat & semisynthetic opioids	0.157	1.619	0.10	0.923	-3.021	3.334
2.1 × SUBTOPICNat & semisynthetic opioids	-0.837	2.248	-0.37	0.710	-5.248	3.574
25–34 years × SUBTOPICOther synthetic opioids	-6.757	1.619	-4.17	0.000	-9.934	-3.579
35–44 years × SUBTOPICOther synthetic opioids	-8.470	1.619	-5.23	0.000	-11.648	-5.292
45–54 years × SUBTOPICOther synthetic opioids	-7.974	1.619	-4.92	0.000	-11.152	-4.796
55–64 years × SUBTOPICOther synthetic opioids	-3.257	1.619	-2.01	0.045	-6.434	-0.079
65–74 years × SUBTOPICOther synthetic opioids	0.545	1.771	0.31	0.758	-2.930	4.020
75–84 years × SUBTOPICOther synthetic opioids	-1.596	2.656	-0.60	0.548	-6.808	3.617
Female × SUBTOPICOther synthetic opioids	-4.426	1.619	-2.73	0.006	-7.604	-1.248
Male × SUBTOPICOther synthetic opioids	-0.265	1.619	-0.16	0.870	-3.443	2.913

Interaction Effects: Is there significant interaction effects between age, sex, and drug type that contribute to higher overdose death rates, for example, if young males are more likely to overdose on opioids compared to other groups?

From the table for the interaction model, we can see that seniors aged 75–84 see 4.1 more heroin deaths per 100k than you'd predict just from being in the 75–84 group, plus the overall heroin effect. Furthermore, the 85+ group similarly has an extra approximately 3.8 heroin deaths / 100k. There are also many negative interactions, for example, 25 to 34 × Heroin: $\beta = -7.035$, $p < 0.00$, 35 to 44 × Heroin: $\beta = -6.991$, $p < 0.001$ and 45 to 54 × Heroin: $\beta = -5.157$, $p = 0.001$, these groups have substantially fewer heroin deaths than you'd expect from adding their age effect and the heroin effect. The same can also be seen for these methadone or synthetic‑opioid groups: All age groups 25–54 × Methadone ($\beta \approx -8$ to -9, all $p < 0.001$), All age groups 25–54 × Nat & semisynthetic opioids ($\beta \approx -4$ to -6, $p < 0.01$), All age groups 25–54 × Other synthetic opioids ($\beta \approx -6$ to -8, $p < 0.001$). These negative values mean that these groups would have fewer overdose deaths than what the main effects model would predict, meaning the interactions are to be counted for. Lastly, we can also see that women have about 4 fewer heroin deaths than the sum of “Female effect” + “Heroin effect.”

Forecasting: Using an ARIMA (0,2,1) model, we forecasted drug overdose death rates for the next five years (2022–2026). The model included statistically significant coefficients and reasonable AIC/BIC values as well as a low residual autocorrelation (ACF = 0.0042). This supports a good fit. The forecast plot indicates a continuing upward trend in drug overdose rates. By 2026, the estimated rate is projected to exceed 30 deaths per 100,000 individuals. The prediction intervals, shown in increasingly wider shaded bands, reflect growing uncertainty over time by consistently suggesting a rising trajectory. These forecasts imply that, without intervention, the public health burden of drug overdoses will likely intensify.

Discussion

One of the largest limitations in fitting the appropriate model for the dataset was the amount of data collected. There were very few predictor variables collected, the total dataset only included age, sex, race, and drug type. Race was not included in the models fitted above, since there was no data for race as an individual categorical variable; the dataset provided information for the interactions between race and another characteristic, such as age or sex. In further models, looking at these interactions could be beneficial, but it would also lead to much larger outputs than the models chosen above. There are many other factors that could affect drug overdose deaths that were not included in the data. Some other potential predictors could be location, education level, and income level. All of these variables could have a large impact on the number of overdose deaths.

Another limitation is that all of the predictors are categorical. This is a large issue for a variable like age, since there could be different rates of deaths for different ages within the range. This can make the results of the regression harder to interpret than if age was made into a numerical variable. The categorical variables also mean that the regression output will be much longer since it must include each category and the interactions between them. Since there are so many categories, there is a mix of significant and non-significant coefficients. This makes it more difficult to tell if the interaction terms are necessary.

An additional factor that makes our model imperfect is the irregularity of health data. There are so many factors that affect deaths, and it is impossible to be able to fully explain the amount of overdose deaths. The uncertainty in each human’s health and the unobserved variables make it impossible to fit a perfect model. These factors also make it difficult to predict overdoses in the future. Advancement in technology and policy changes could have large effects on the number of deaths, but these types of changes cannot be accounted for in our model.

Citation:

National Center for Health Statistics. (2025, April 21). ***DQS drug overdose death rates, by drug type, sex, age, race, and Hispanic origin: United States*****. Centers for Disease Control and Prevention.** **https://www.cdc.gov/nchs/dataquery/index.htm**