MATH 6364 Statistical Methods

Gaukhar Nurbek

Final project

12/4/2020

I Introduction

Air pollution is one of the biggest environmental problems to health and the key death factor in low-, middle-, and high-income countries. According to the World Health Organization outdoor air pollution has caused 4.2 million premature deaths around the world in 2016. Along with outdoor air pollution indoor air pollution like indoor smoke is a serious health risk for 3 billion people who cook and heat their homes using biomass, kerosene fuels and coal [1]. The purpose of this study is to examine potential factors associated with number of deaths from the air pollution in low-, middle- and high-income countries for five years between 2012 and 2016. Hypothesis of this study is to examine if there is any connection between number of the deaths for each country where risk factor is air pollution and proportion of the population that has access to clean fuels, access to electricity and the GDP per capita.

II Methods

II.a Study population and Measures

The study is longitudinal, since the data that is going to be used was collected repeatedly for 5 years from 2012 to 2016. The study population of interest are low-, middle- and high-income countries (177 different countries). Single-staged cluster sampling of the low-, middle- and high-income countries was used in the current study. The outcome variable of interest is the number of death in each country where risk factor is air pollution. Potential explanatory variables of interest are: time, access to clean fuels and technologies for cooking (% of population), access to electricity (% of population), GDP per capita (US dollars). Time, access to clean fuels and technologies for cooking, access to electricity, GDP per capita are continuous variables. Time is used as continuous variables and has values between 1 and 5, computed as Year-2011. Access to clean fuels and technologies for cooking is the proportion of total population primarily using clean cooking fuels and technologies for cooking. Data for access to clean fuels and technologies for cooking are based on the World Health Organization's (WHO) Global Household Energy Database. Access to electricity is the percentage of population with access to electricity. GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. Data are in current U.S. dollars.

II.b Statistical Methods

Descriptive statistics methods such as proc freq and proc mean were used to summarize the dataset and to plot number of deaths from air pollution for each country for 2012-2016 years in order to study the distribution of the output variable. It was decided to use the general estimation equation (GEE) model for longitudinal count data clustered by each country. Since general estimation equation model requires link function, distribution and correlation structure, there were done several regression analysis for GEE with different distributions such as Poisson, negative-binomial and geometric and for different correlation structures such as exchangeable, autoregressive and unstructured. The best correlation structure and distribution was chosen based on the QIC statistics. To check the GEE assumption about responses being correlated and clustered working correlation matrix was built. To identify independent

estimators univariable GEE analysis was used. Assessment for interaction effect was also provided by using the GEE for longitudinal count data. Residuals, influential points and outliers were analyzed graphically. All analyses were conducted using SAS 9.4.

III. Results

From the Table 1 below it can be seen mean analysis of the output and possible estimators mean, variance values of our sample data with a size of 885 clustered for the first 5 countries among the 177 countries. Mean value of the proportion of population among the countries that has access to clean energy (CFT) in 2012-2016 years for cooking is 28% in Afghanistan, 73% in Albania, 46% in Angola. It also can be seen that on average 79% of the population in Afghanistan, 46% of the population in Angola has access to the electricity(ELC) along with the mean GDP per capita value being 603 USD, 4687 USD accordingly.

The MEANS Procedure

Country	N Obs	Variable	N	Mean	Variance	Minimum	Maximum
Afghanistan	5	Num_deaths_AP	5	26212.40	96463.30	25780.00	26529.00
		CFT	5	28.1560000	10.6793300	24.0800000	32.4400000
		ELC	5	79.3460000	178.7515800	68.9300000	97.7000000
		GDP	5	603.7176309	1626.49	547.2281102	641.8714792
Albania	5	Num_deaths_AP	5	1570.00	1793.50	1514.00	1611.00
		CFT	5	73.7020000	8.5949200	69.9600000	77.4200000
		ELC	5	99.9800000	0.0020000	99.9000000	100.0000000
		GDP	5	4263.24	59476.64	3952.80	4578.63
Algeria	5	Num_deaths_AP	5	12093.20	419519.20	11304.00	12972.00
		CFT	5	92.7440000	0.0566300	92.4700000	93.1000000
		ELC	5	99.6120000	0.2567700	98.7600000	99.9900000
		GDP	5	4941.84	653088.13	3946.44	5592.26
Andorra	5	Num_deaths_AP	5	22.6000000	0.3000000	22.0000000	23.0000000
		CFT	5	100.0000000	0	100.0000000	100.0000000
		ELC	5	100.0000000	0	100.0000000	100.0000000
		GDP	5	38553.27	4376681.74	35762.52	41303.93
Angola	5	Num_deaths_AP	5	11310.40	279373.30	10758.00	12025.00
		CFT	5	46.7180000	1.0888700	45.3600000	48.0500000
		ELC	5	37.4420000	16.0694700	32.0000000	42.0000000
		GDP	5	4687.29	669645.29	3506.07	5408.41

Table 1

Figure 1a below displays the relationship between number of deaths and each year from 2012 to 2016 for each country among 177 and the red line displays the average. It can be seen that values of the number of deaths from the air pollution didn't change dramatically in 2012-2016 years among each country, but still has slightly changes. And for two countries number of deaths from the air pollution is extremely high (around 1.2 million) in comparison with the rest of the countries. From this point it's been decided to conduct analysis for those values where average number of deaths from the air pollution was higher than 1000000, because of the possible bias it could cause for the model. In figure 1b below relationship between year and countries where number of deaths was less than a million per year is displayed

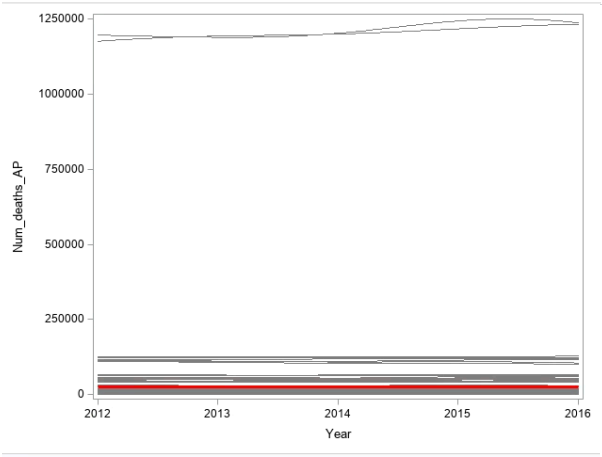

Figure 1a

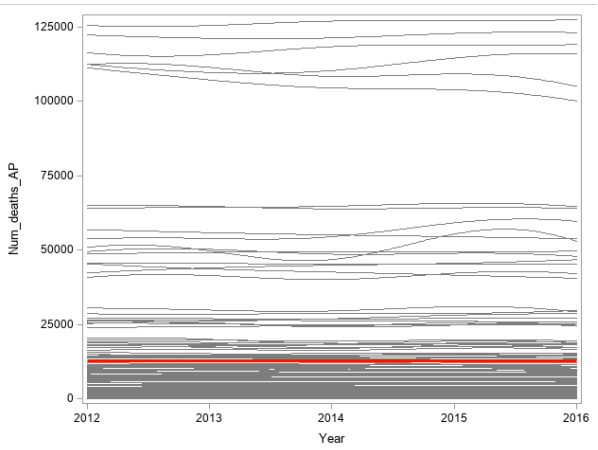

Figure1b

Figure 2 displays the distribution of the number of deaths from the air pollution values for the all dataset.

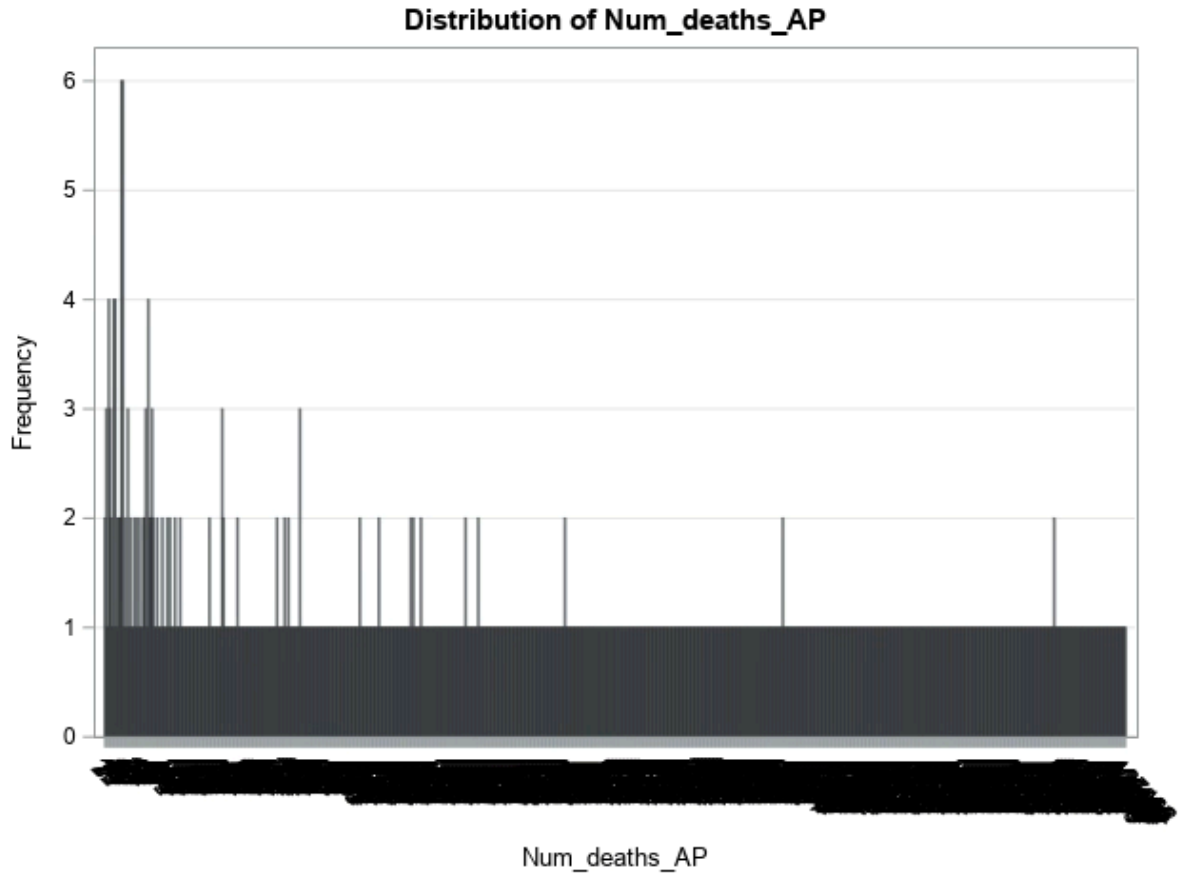

Figure 2

As it was mentioned in II.b in order to find the best fit correlation structure and distribution setting for the dataset, there were made experiments by fitting different GEE models with different correlation structure (exchangeable, unstructured, autoregressive) and for different distributions (Poisson, negative-binomial, geometric). GEE fit criteria QIC was

used in order to choose best fit correlation structure. In the cases when fitting model with exchangeable type correlation matrix for Poisson distribution QIC = -4390.7326 and unstructured type of the correlation matrix for Poisson distribution QIC = -4281.4337, when fitting a model with autoregressive matrix correlation type for Poisson distribution model converged and QIC = -4416.0608. Further there were done analysis in order to find the best fit distribution setting for our data, as a result the model with negative-binomial distribution showed the smallest QIC criteria= -139757706.2 in comparison with the rest Poisson, geometric distributions, zero inflated distribution wasn't included in comparison because there weren't many zeros in the output count variable. Thus, GEE model with negative binomial distribution and autoregressive correlation matrix type was further used in the analysis. After choosing the best settings of the model univariable analysis were conducted for each possible estimator after which they were included into multivariable model. As a result of this step following estimators were included into the model CFT (proportion of population with access to the clean energy for cooking, p-value = 0.2759 SE = 0.0014), ELC (proportion of population with access to electricity) (p-value = 0.5160 SE = 0.0003), time (p-value 0.5268 SE=0.0021), GDP per capita (*p-value <0.0001 SE =0.0000) was significantly associated with difference in the logs of expected number of the death from the air pollution for each country, but the parameter estimate and standard error were equal to 0, in order to reduce bias in the model GDP per capita was excluded from the model. After choosing the estimator for the multivariable regression model, it was fitted and showed the QIC criteria = -138321655.2. Estimates for preliminary GEE can be seen from the Table 2 below.

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	9.4854	0.1536	9.1844	9.7864	61.76	<.0001
CFT	-0.0004	0.0012	-0.0028	0.0021	-0.28	0.7759
ELC	-0.0001	0.0003	-0.0007	0.0005	-0.24	0.8141
time	-0.0011	0.0025	-0.0060	0.0038	-0.43	0.6682

Table 2

After fitting the initial model, there were done tests for interaction effect as a result no interaction effect was added to the model, since none of them was significantly associated to with a output. Fitting the final model with following estimators: CFT, ELC, time produced following tables 3,4,5:

Working Correlation Matrix					
	Col1	Col2	Col3	Col4	Col5
Row1	1.0000	0.9995	0.9990	0.9984	0.9979
Row2	0.9995	1.0000	0.9995	0.9990	0.9984
Row3	0.9990	0.9995	1.0000	0.9995	0.9990
Row4	0.9984	0.9990	0.9995	1.0000	0.9995
Row5	0.9979	0.9984	0.9990	0.9995	1.0000

Table 3

GEE Fit Criteria	
QIC	-138321655.2
QICu	-138321658.5

Table 4

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	9.4854	0.1536	9.1844	9.7864	61.76	<.0001
CFT	-0.0004	0.0012	-0.0028	0.0021	-0.28	0.7759
ELC	-0.0001	0.0003	-0.0007	0.0005	-0.24	0.8141
time	-0.0011	0.0025	-0.0060	0.0038	-0.43	0.6682

Table 5

Table 3 shows the working correlation matrix for the proposed model and it can be seen that response variable is highly correlated for each clustered country level for 5 years which satisfies the assumption of the GEE about the high correlations of the output variables. It can be seen from the table 4 that QIC and QICu values are almost similar for the current model, that shows that the model is a good fit for the data. Also, from the table 5 it's clear that for all chosen estimators significant error is below 0.003, even though none of them has a p-value < 0.05, but still was included to the model, in order to find relationship between explanatory and output variables. For one-unit growth in CFT the difference in the logs of expected average number of the deaths from the air pollution for each country would be expected to decrease by 0.0004 unit per country population, while holding the other variables in the model constant. For one-unit growth in ELC and time estimator the difference in the logs of expected average number of the death from the air pollution would be expected to decrease by 0.0001 and 0.0011 unit per country population accordingly, while holding the other variables in the model constant. Which shows that access to clean energy and to the electricity as time goes by may affect in reducing the difference in the logs of expected average number of the deaths from the air pollution for each country. Along with fitting final model, there were done graphical analysis for residuals, outliers and influential points. Residuals of the model can be seen from figure 3, taking into consideration that the dataset wasn't simulated and contains real data the model seems to predict most of the values well. Besides few values clustered at the top of the plot with the highest residuals.

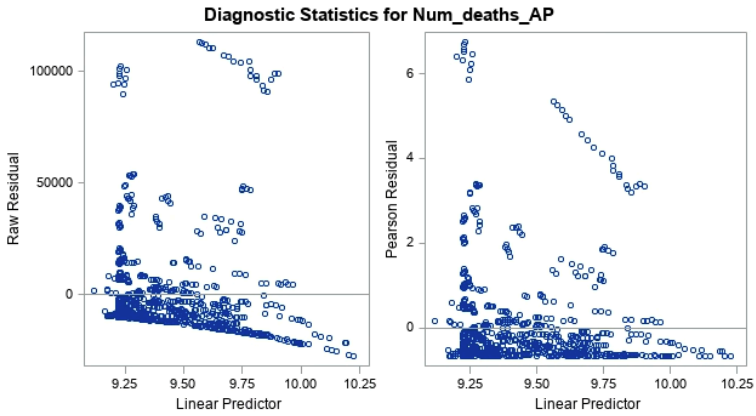

Figure 3

Figure 4 below shows that according to Cook's D vs Observation plot there are only few values deleting of which will affect the output value and Leverage Vs Observation plot, there are not much of the outliers or influential points that could affect the model performance.

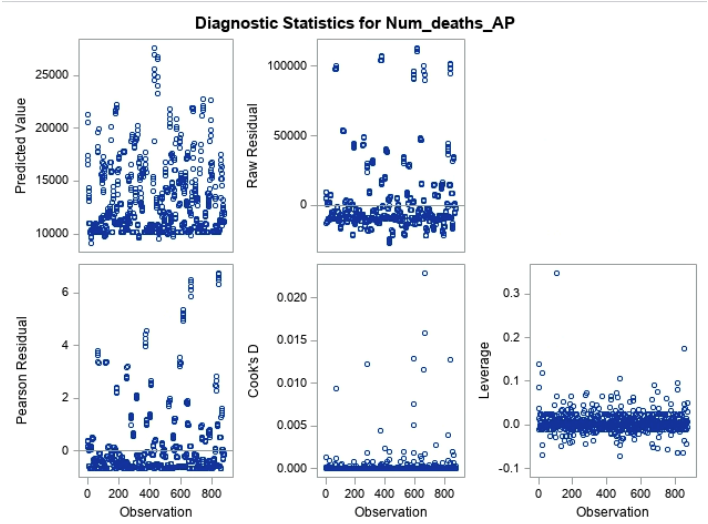

Figure 4

Figure 5 below shows that standardized DFBETA plots of each estimates. DFBETA plots show each estimator has a few influential points, deleting of which will result maximum in -0.2(in case of time and ELC) on the output, overall these points are not significantly influential on the model's output.

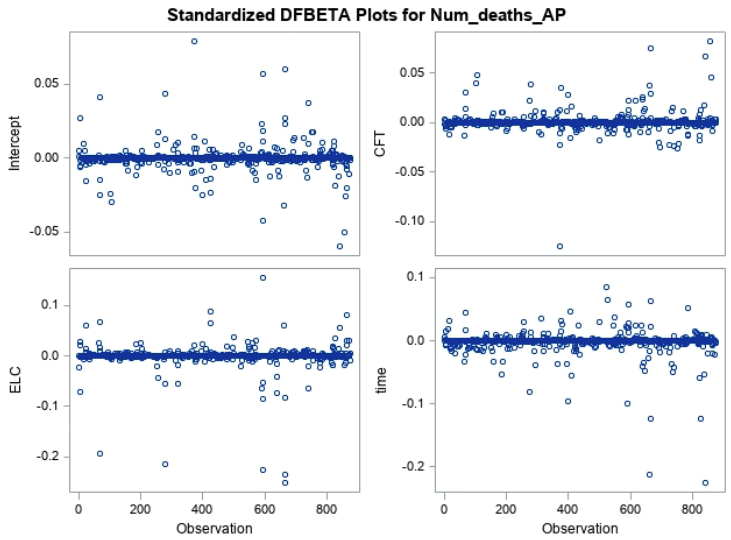

Figure 5

IV. Discussion

The main finding of the model is that the difference in the logs of expected average number of the deaths from the air pollution for each country can be decreased by increasing access to clean energy for cooking and electricity, one of the explanations can be that more people have clean energy for cooking and electricity, the less people will burn coal at home, which will lead to that the level of the air pollution will decrease, thus the number of deaths from the air pollution will also decrease. Although none of the relationship between GDP of the country and the average number of deaths from the air pollution was found. That is why hypothesis of this study was satisfied partially. Strengths of this study is that it provides GEE regression analysis for longitudinal count data wnumber of deaths from air pollution with negative-binomial distribution and builds the model that helps to understand what are the main factors that can help to reduce the number of deaths from air pollution and to make the environment safer place. Weakness of this study is that it was conducted only for 5 years data of the period 2012-2016 and it can contain bias according to the data, also reduced number of estimators could cause bias in the suggested model. Results could be generalizable to the general population since they can be useful for the broader group of countries of any type.

V. Conclusions

During this study there were done several general effects estimating analysis for clustered longitudinal count data to find the best correlation structure type and the best distribution type to fit the data. Assumption of the GEE model for the correlated output was satisfied. Univariable analysis and interaction effect tests were done in order to find the final model. During fitting the final model residuals, outliers and influential points were graphically analyzed. It was found that more people have access to clean energy and electricity the less amount of people is expected to die from the air pollution.

References

1. https://www.who.int/en/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

Appendix

```
* Connect the local folder to sas working directory;
libname gnb "Z:\Downloads\math6364\final";
* Import the dataset from the local folder into sas working directory;
PROC IMPORT OUT= GNB.final
        DATAFILE= "Z:\Downloads\math6364\final\new_dataset_1.csv"
        DBMS=csv REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

* explore data;
proc sgplot data = gnb.final noautolegend ;
        pbspline x = year y = num_deaths_AP
        / group = country nomarkers LINEATTRS = (COLOR= gray PATTERN = 1 THICKNESS
=1);
        pbspline x = year y = num_deaths_AP
        / nomarkers LINEATTRS = (COLOR= red PATTERN = 1 THICKNESS = 3);
run;
quit;

proc sgplot data = gnb.final(where=(num_deaths_ap<1000000)) noautolegend ;
        pbspline x = year y = num_deaths_AP
        / group = country nomarkers LINEATTRS = (COLOR= gray PATTERN = 1 THICKNESS
=1);
        pbspline x = year y = num_deaths_AP
        / nomarkers LINEATTRS = (COLOR= red PATTERN = 1 THICKNESS = 3);
```

```
run;
quit;

proc means data = gnb.final(where=(num_deaths_ap<1000000)) n mean var min max;
  var num_deaths_AP CFT ELC GDP;
  class country;
run;

proc freq data=gnb.final(where=(num_deaths_ap<1000000));
tables num_deaths_AP / plots=freqplot;
run;

* add new column time to store year-2011;
PROC SQL;
ALTER TABLE GNB.final  ADD time NUM (8);
QUIT;

* insert values into column time;
PROC SQL;
UPDATE GNB.final SET time=year-2011;
QUIT;

* choosing correlation structure and distribution ;
* convergenes is questionable;
proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = CFT time/dist=pois link=log;
REPEATED SUBJECT=country / TYPE=EXCH CORRW;
run; quit;
* QIC -4390.7326;
```

```
proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = CFT time/dist=pois link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;
* -4416.0608;

proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = CFT time/dist=pois link=log;
REPEATED SUBJECT=country / TYPE=un CORRW;
run; quit;
* QIC -4281.4337;

proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = CFT time/dist=nb link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;
* QIC -139757706.2 best option;

proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = CFT time/dist=geometric link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;
* QIC -63027593.50;

* bivariate analysis;
proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
```

```
model Num_deaths_AP = CFT/dist=nb link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;*p-value 0.2759 SE 0.0014;
proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = ELC/dist=nb link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;*p-value 0.5160 SE 0.0003;

proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = GDP/dist=nb link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;*p-value <0.0001 SE =0.0000;

proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = time/dist=nb link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;*p-value 0.5268 SE=0.0021;

* fitting model data;
proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = CFT ELC time/dist=nb link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;*QIC -138321655.2;



* Test for interaction effect;
proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
```

```
class country;
model Num_deaths_AP = CFT ELC time CFT*ELC/dist=nb link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;*p -0.00000;

proc genmod data = gnb.final(where=(num_deaths_ap<1000000));
class country;
model Num_deaths_AP = CFT ELC time CFT*time/dist=nb link=log;;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;*p -0.00000;

proc genmod data = gnb.final;
class country;
model Num_deaths_AP = CFT ELC time ELC*time/dist=nb link=log;;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;*p -0.00000;


*Fitting final model along with analyzing residuals, outliers and influential points;

proc genmod data = gnb.final(where=(num_deaths_ap<1000000)) plots=all descending;
class country;
model Num_deaths_AP = CFT ELC time /dist=nb link=log;
REPEATED SUBJECT=country / TYPE=ar CORRW;
run; quit;
```