Fitting a change point model for circular vascular mortality in LA County

Raymond Dueñas
California State University Stanislaus
Rduenas2@csustan.edu

Yangong Wu
California State University Stanislaus
ywu1@csustan.edu

Abstract

In this project, we analyze the weekly mortality of LA county from 1970 to 1979 due to circular vascular complications. Our contributions are in two aspects. First, we find a better model fitting by using the change-point model instead of the linear trend model by treating the temperature and pollution as covariates, and the improvement is quite significant. Second, the time series analysis of the residuals shows that AR(2) gives a quite satisfactory fitting for the errors, and it can give better predictions. All the computations are carried in R, and R-codes and outputs are given.

Introduction

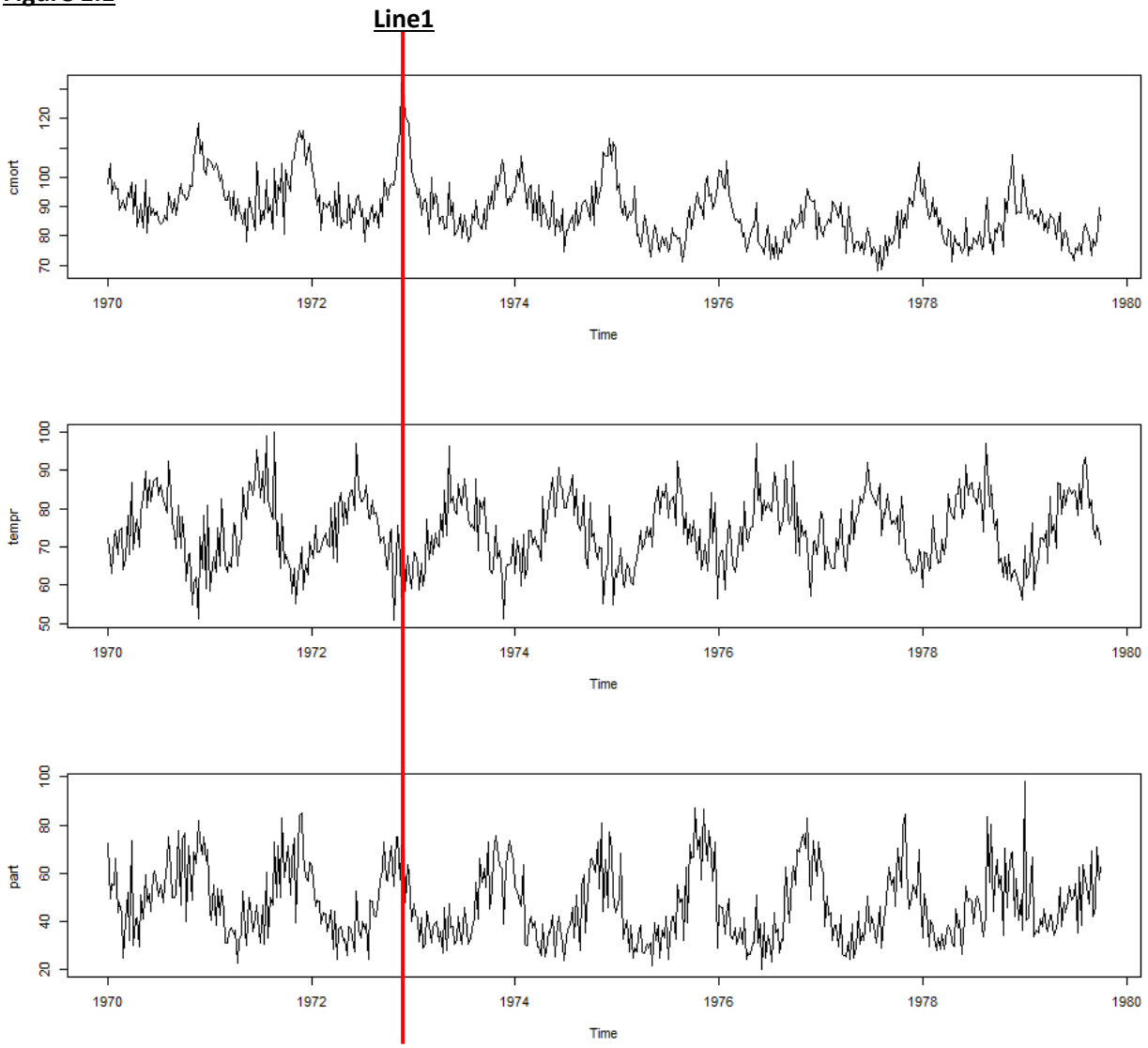
The following reading is broken up into four sections discussing and demonstrating the methods of the project, followed by our conclusion, which will some up the work performed. An index containing the project's source code without any outputs has been included for accessibility. The source code used to execute the computations of this project is also included throughout the methods portion of the reading. These sections will include sections describing the methods of our project, R code followed by their outputs, and have been formatted so that the reading follows the process used to execute our project. This structure allows the reader to read the report and review the work performed linearly from start to finish.

Section 1 Data set and Variables

The project was executed using weekly records from 1970 to 1979 with respect to three variables mortality count (cmort), average high temperature (tempr), and particulates as the measurement for pollution(part). A plot of the data sets can be found in Figure 1.1, plotting the data sets allows for visualization of the correlation. An observation of Figure1.1 leads to the recognition of a possible relationship between the increased mortality rate with cold temperatures and increased amounts of pollution particles, as demonstrated by Line1. The line highlights one instant of a trend that may indicate the relation between the variables. In the cmort plot, we see the large spike in mortality at this same point in time; the tempr plot records low temperatures, and the part plot indicates increased pollution particles.

```
>library(astsa)
>par(mfrow=c(3,1))
>plot(cmort)
>plot(tempr)
>plot(part)
```

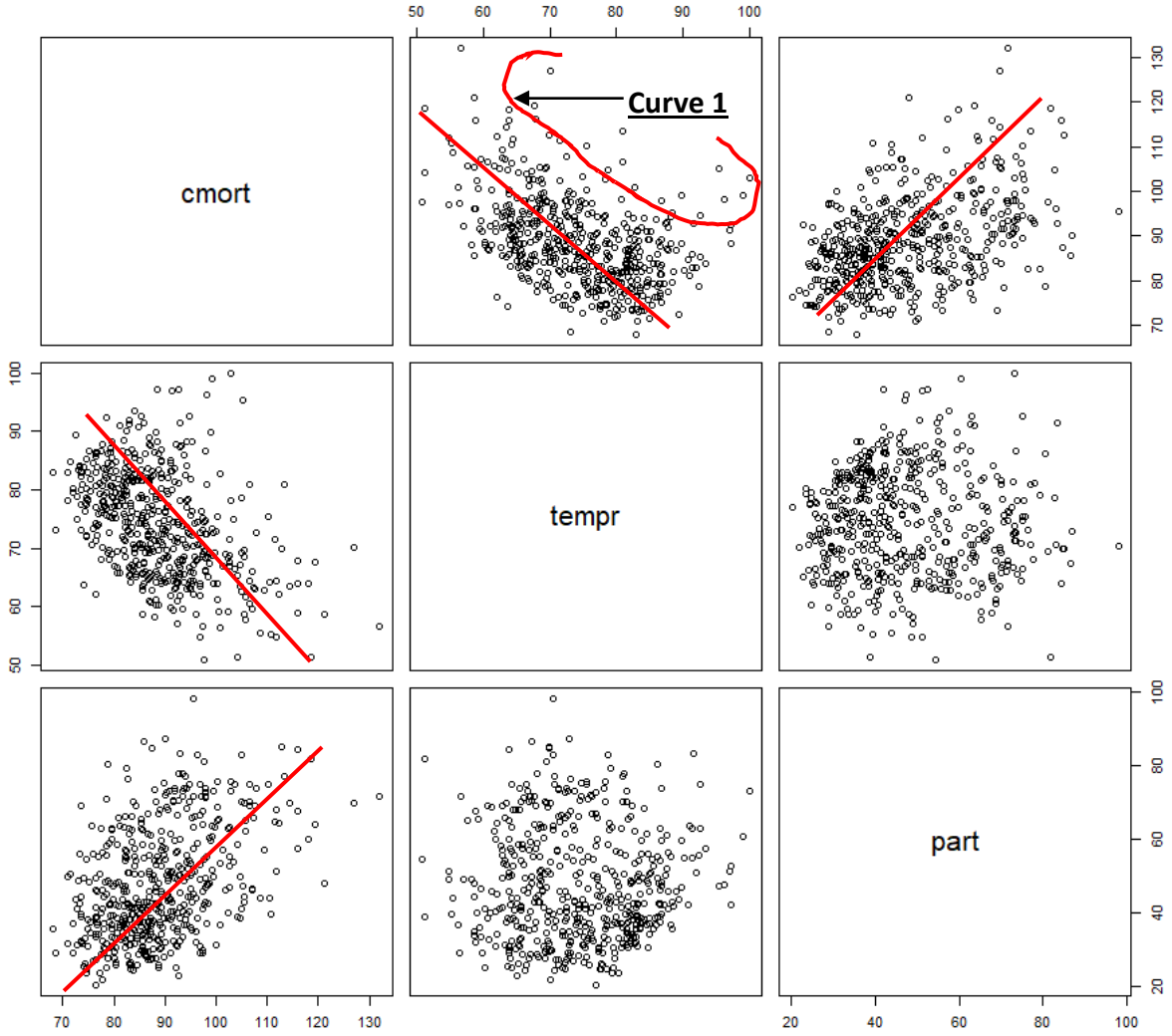
Figure 1.1



Section 2 Regression model (trend fitting)

We will de-trend the data set by fitting a linear regression model in order to perform a time series analysis. First, we visualize the data by combining the three data sets on the table found in Figure 2.1 and viewing a summary of the tree data sets in Table 2.1 found below. We start this process by standardizing temperature by subtracting its mean; temp=tempr-mean(tempr). We have outliers in the temperature plot, which can be seen encapsulated in Curve 1 in Figure 2.1. To account for this in our model, we will add squared temp to the model. We also add a pollution variable to our model so that our current model fitting and an estimator of it can be found in Table 2.2. The last column of the table contains the P values of strength of evidence for each perimeter in our model and indicates our current model fitting is strong. Table 2.2 also holds the residual standards error for our fitting as well as the F-statistic, note that when comparing F-statistics, the larger value is better. The residuals are plotted on the following graph in Figure 2.2. Included is also the AIC (Akaike Information Criterion) that measures how good the model fitting is. The AIC is defined as AIC= -2log-likelihood +2(number of parameters). In the linear model, -log-likelihood is the same as the total sum of squared error. When comparing two model fittings, the model with a smaller AIC has a better fitting>pairs(cbind(cmort,tempr,part))

Figure 2.1 (>pairs(cbind(cmort,tempr,part)))



>summary(cmort); summary(tempr); summary(part) > temp<-tempr-mean(tempr)

Table 2.1 (>summary(cmort); summary(tempr); summary(part) > temp<-tempr-mean(tempr))

Variable	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Cmort	68.11	81.90	87.33	88.70	94.36	132.04
Tempr	50.91	67.23	74.06	74.26	81.49	99.88
Part	20.25	35.85	44.25	47.41	57.54	97.94

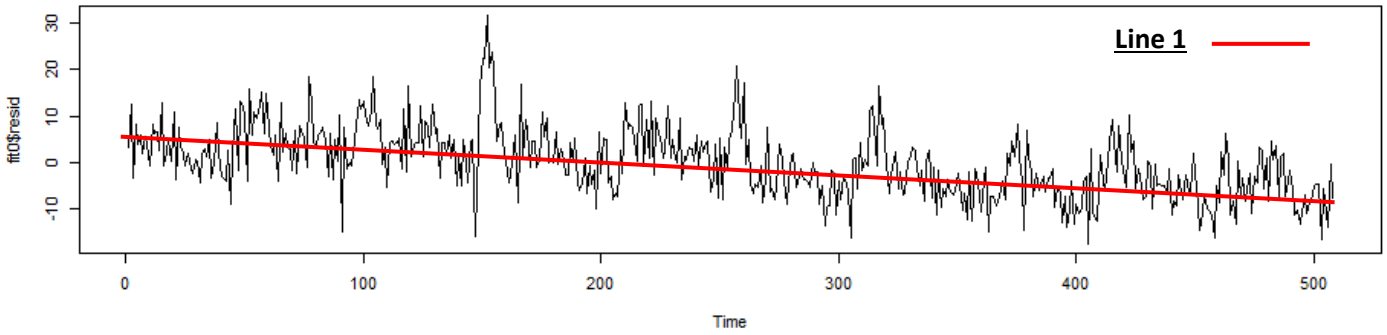
> temp2<-temp^2
> fit0<-lm(cmort~temp+temp2+part)
>summary(fit0); AIC(fit0))

Table 2.2(>summary(fit0); AIC(fit0))

Model: $Y_t = \beta_0 + \beta_1 temp + \beta_2 temp^2 + \beta_3 part + error$				
Residuals				
Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-17.6217	-5.2045	-0.3203	4.7150	31.5653
Coefficients:	**Estimate**	**Std Error**	**T value**	**Pr(>|t|)**
(Intercept)	73.742033	1.108285	66.537	< 2e-16 ***
Temp	-0.499301	0.037036	-13.481	< 2e-16 ***
Temp2	0.024177	0.003314	7.294	1.17e-12 ***
part	0.274110	0.022070	12.420	< 2e-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 7.493 on 504 degrees of freedom				
Multiple R-squared: 0.4418, **Adjusted R-squared:** 0.4384				
F-statistic: 132.9 on 3 and 504 DF, **p-value:** < 2.2e-16				
AIC: [1] 3493.782				

```
>fit0<-lm(cmort~temp+temp2+part)
>summary(fit0)
> plot.ts(fit0$resid)
```

Figure 2.2(> plot.ts(fit0$resid)) **Residuals**



The residual plot in Figure 2.2 shows that the mean decreases, as demonstrated by Line 1. This motivates us to find better model fitting. One possible way to find the desired model is to add a variable of time, defined as trend below. After defining trend and revising the linear model (fit1), we use the summary to view the trend variable's results on the overall fitting of our model, see Table 2.3. Notice that the AIC has decreased from 3493.782 to 3332.282 indicating that fit1 is significantly better than the original linear model fitting (fit0). In observing the residuals plotted on the graph in Figure 2.3, Line1 demonstrates the successful elimination of the decreasing trend that was characteristic of the mean in fit0, fit1 moves the mean closer to zero. In R, ANOVA( ) is used to compare two model fittings, where the second model is larger or an extension of the first. To further compare fit1 to fit0, we use the ANOVA function, the results of which are found in Table 2.4. From the above Table 2.4, the F-value is increased by 190.98 with p-value<2.2e-16, which is very strong, further verifying that fit1 is superior to fit0.

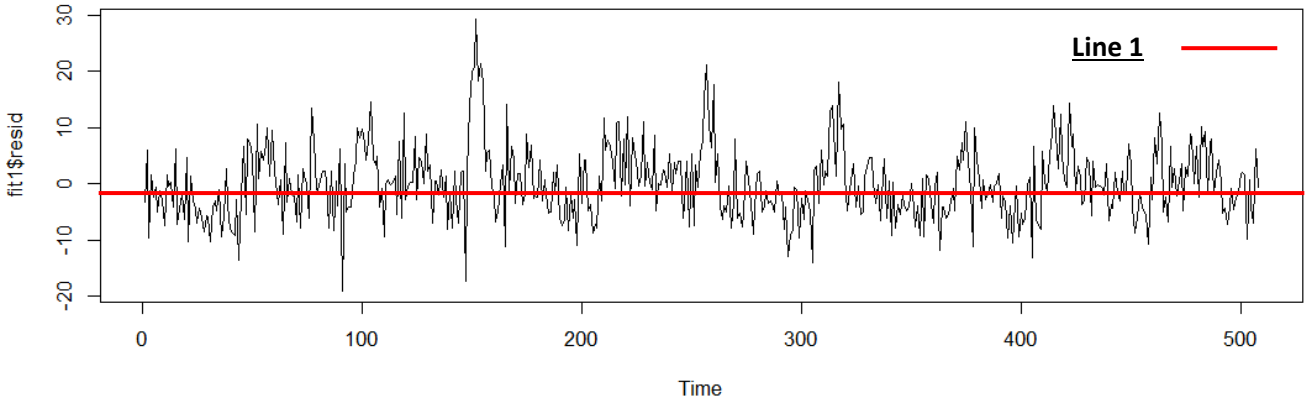
> trend<-time(cmort)
> fit1<-lm(cmort~trend +temp+temp2+part)
>summary(fit1); AIC(fit1)

Table 2.3(>summary(fit1); AIC(fit1))

Model: $Y_t = \beta_0 + \beta_1 trend + \beta_2 temp + \beta_3 temp^2 + \beta_4 part$ +error				
Residuals				
Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-19.0760	-4.2153	-0.4878	3.7435	29.2448
Coefficients	**Estimate**	**Std Error**	**T value**	**Pr(>|t|)**
(Intercept)	2.831e+03	1.996e+02	14.19	<2e-16***
Trend	-1.396e+00	1.010e-01	-13.82	<2e-16***
Temp	-4.725e-01	3.162e-02	-14.94	<2e-16***
Temp2	2.259+e-02	2.827e-03	7.99	9.26e-15***
part	2.554e-01	1.886e-02	13.54	<2e-16
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 6.385 on 503 degrees of freedom				
Multiple R-squared: 0.5954, **Adjusted R-squared:** 0.5922				
F-statistic: 185 on 4 and 503 DF, **p-value:** < 2.2e-16				
AIC: [1] 3332.282				

>plot.ts(fit1$resid)

Figure 2.3(>plot.ts(fit1$resid)) **Residuals**



>anova(fit0,fit1)

Table2.4 (>anova(fit0,fit1))

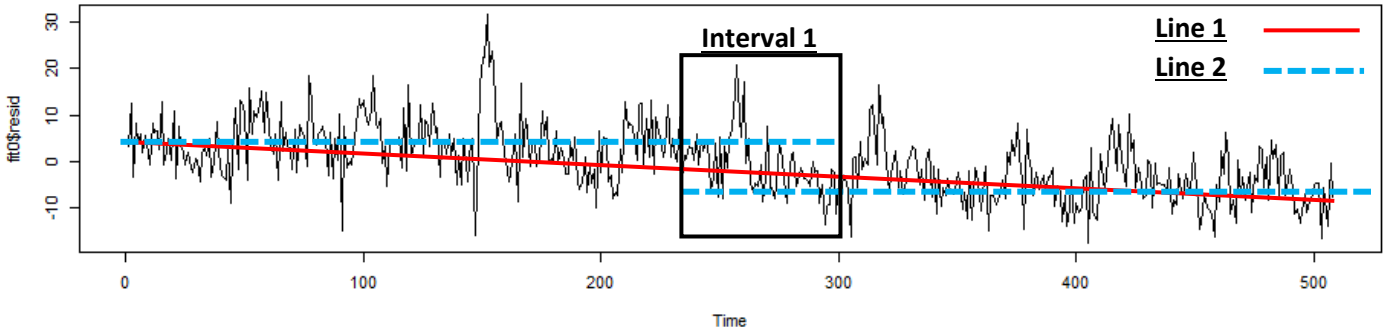
Analysis of Variance Table						
Model 1 (fit0): $Y_t = \beta_0 + \beta_1 temp + \beta_2 temp^2 + \beta_3 part + error$ (cmort ~ temp + temp2 + part)						
Model 2 (fit1): $Y_t = \beta_0 + \beta_1 trend + \beta_2 temp + \beta_3 temp^2 + \beta_4 part + error$ (cmort~trend+temp+temp2+part)						
Model:	**Res. DF**	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
1	504	28295				
2	503	20508	1	7786.6	190.98	<2.2e-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Section 2.1 A Change-Point Model: Our Contribution

The contribution of this project is to consider a change-point model, meaning the mean has a sudden drop instead of a gradual decrease. We will show that the change-point model is a better fitting then the linear model. In Figure2.4, we revisit the graph containing the plotted residuals of fit0 to demonstrate the difference in the way the two types of model fitting the data. Line 1 is the same line seen in Figure 2.1 and represents the linear model it treats the data as plot points fluctuating about a mean with a constant slope in the negative direction. Line 2 represents the change-point model and treats the data as a plot with points fluctuating about two different relative means. The point in time at which the points begin to fluctuate about different means is what we call the change point. Interval 1 identifies the relative area in which we can visually identify a change point is taking place. However, this visual is not enough to determine where exactly the change point is.

>plot.ts(fit0$resid)

Figure 2.4(> plot.ts(fit0$resid)) **Residuals**



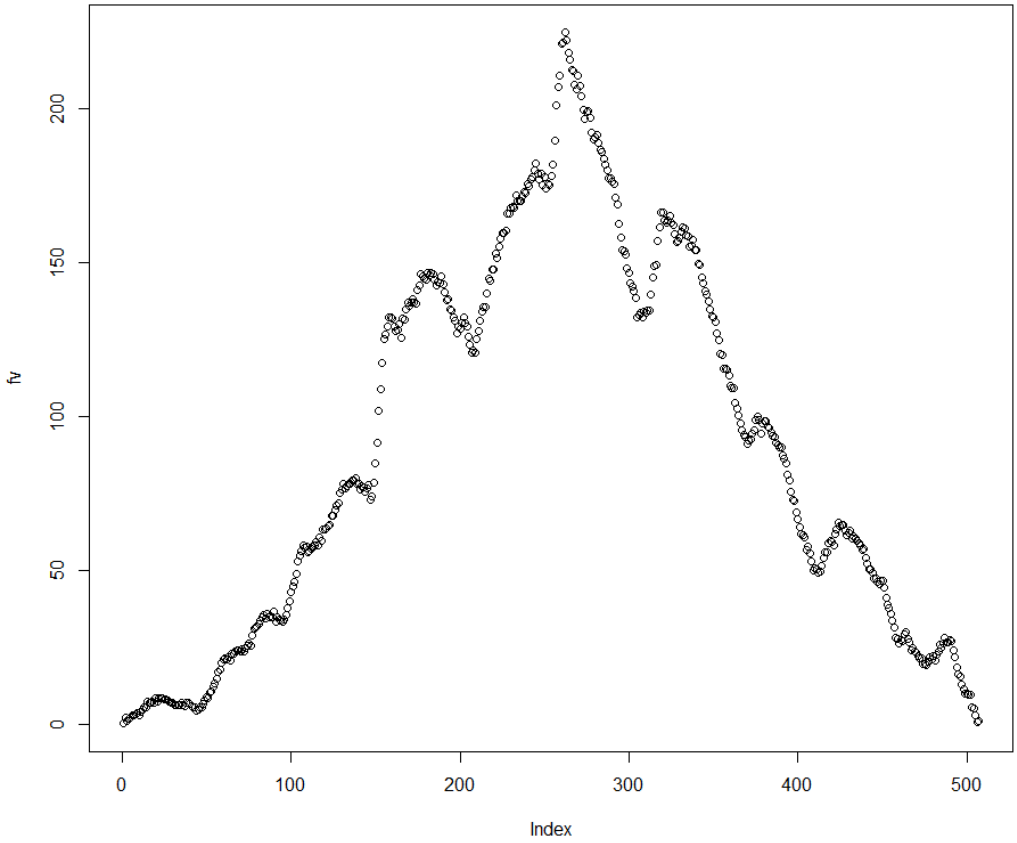
Since we do not know where the change point is, we search by considering all possible values of change time and select the one that has the largest F-value. The point in time with the largest F-value represents the best fitting. The process of finding this value is handled by running a loop in R, the code for which is found below. We define the trend of the F-value into variable fv[i], in which we define our model (fit2), which will be adjusted with respect to time to yield 507 models, and 507 F-values. We adjust time in fit2 using variable i int the interval [1,507]. After running the loop we visualize the trend by plotting each fv[i] value on a graph, as seen in Figure 2.5.

Change point Identifying loop code

```
> fv <-numeric(507)
>  for(i in 1:507){                    #(the change-point is at t=i with largest F-value)
+ time<-c(rep(1,i),rep(0,508-i))       #(changes the value of time with every iteration of the loop)
+ fit2<-lm(cmort~time+temp+temp2+part)  #(defines our model fit2 which is adjusted by i with every loop)
+ fv[i]<-anova(fit0,fit2)$F[2] }        #(assigns the increased F-value, resulting from comparison of
                                          fit2's F-value with fit0's, into fv[i] for i in [1,507])
```

> plot(fv)

Figure 2.5(> plot(fv)) Plot of all F-Values



Now we have visualized the data and can see that there is a distinct point with the greatest F-value. To identify what that point is we use an R program max function (c(1:507)[fv==max(fv)]). We find that the greatest increase in F-value is at 262 weeks. Using this information, we define our time variable as the interval for fit2. With the change point identified and time variable defined, we can express the change-point model fitting fit2, which is found below. After defining the fitting, we must compare it with our previous fittings to determine which is the best fitting. The qualities used to compare fittings include the improvement of the F-value found in the Analysis of Variance Table in Table 2.5, the value of the AIC, and information from the summary of fit2, which can be found below in Table 2.6. Notice that the increase in the F-value for fit2 is significantly larger than the increase in F-value for fit1, previously observed in Table2.4 and that the AIC value found in Table 2.5 is significantly less than the AIC value for fit1 previously observed in Table 2.3. That fit2 has a significantly larger F-value and smaller AIC value indicates fit2 is a significantly better fitting than fit1 and fit0. We also observed in Table 2.6. that the p values of the individual variables selected for the fitting fit2 all have strong evidence of a correlation with cmort. From Table 2.6, we can determine the parameters for our model, yielding us the complete fitting fit2, which is listed below. This model means there is a drop in mortality of 8.364 after week 262. The model fit2 gets its parameters from the estimate column in the table. It also includes an identifier function to handle the adjustment of the mean once the change point is reached. Also included below are the residuals from fit2 plotted on the graph in Figure 2.6.

The largest F-value is at time

> c(1:507)[fv==max(fv)]

[1] 262

Defining the time variable as

> time<-c(rep(1,262),rep(0,508-262))

The change-point model fitting (before defining parameters)

$Y_t = \beta_0 + \beta_1 I[t \leq t_0] + \beta_2 temp + \beta_3 temp^2 + \beta_4 part$

> fit2<-lm(cmort~time+temp+temp2+part)

> anova(fit0,fit2)

Table 2.5(> anova(fit0,fit2))

Analysis of Variance Table						
Model 1 (fit0): $Y_t = \beta_0 + \beta_1 temp + \beta_2 temp^2 + \beta_3 part + error$ (cmort ~ temp + temp2 + part)						
Model 2 (fit2) $Y_t = \beta_0 + \beta_1 I[t \leq t_0] + \beta_2 temp + \beta_3 temp^2 + \beta_4 part$ (cmort~time+temp+temp2+part)						
Model:	**Res. DF**	**RSS**	**Df**	**Sum of Sq**	**F**	**Pr(>F)**
1	504	28295				
2	503	19562	1	87.331.6	224.55	<2.2e-16 ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

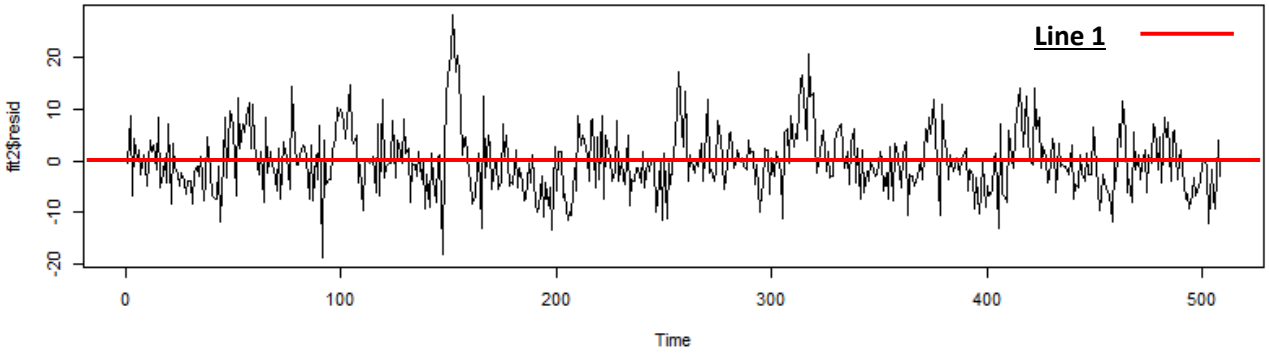
>summary(fit2); AIC(fit2)

Table 2.6(>summary(fit2); AIC(fit2))

Model: $Y_t = \beta_0 + \beta_1 I[t \leq t_0] + \beta_2 temp + \beta_3 temp^2 + \beta_4 part$				
Residuals				
Minimum	**1st Quartile**	**Median**	**3rd Quartile**	**Maximum**
-18.6411	-3.9988	-0.3711	3.4749	28.0561
Coefficients	**Estimate**	**Std Error**	**T value**	**Pr(>|t|)**
(Intercept)	70.791516	.0943212	75.054	<2e-16***
Time	8.364444	0.558183	-14.985	<2e-16***
Temp	-0.468028	0.030896	-15.148	<2e-16***
Temp2	0.021889	0.002769	7.923	1.51e-14***
part	0.249267	0.018444	13.515	<2e-16***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 6.236 on 503 degrees of freedom				
Multiple R-squared: 0.6141, **Adjusted R-squared:** 0.611				
F-statistic: 2001 on 4 and 503 DF, **p-value:** < 2.2e-16				
AIC: [1] 3308.28				

>plot.ts(fit2$resid)

Figure 2.6(>plot.ts(fit2$resid)) **Residuals**



Fitting fit2 with parameters

$Y_t = 70.79 + 8.364I[t \leq 262] - 0.4680 \times temp + 0.022 \times temp^2 + 0.2493 \times part$

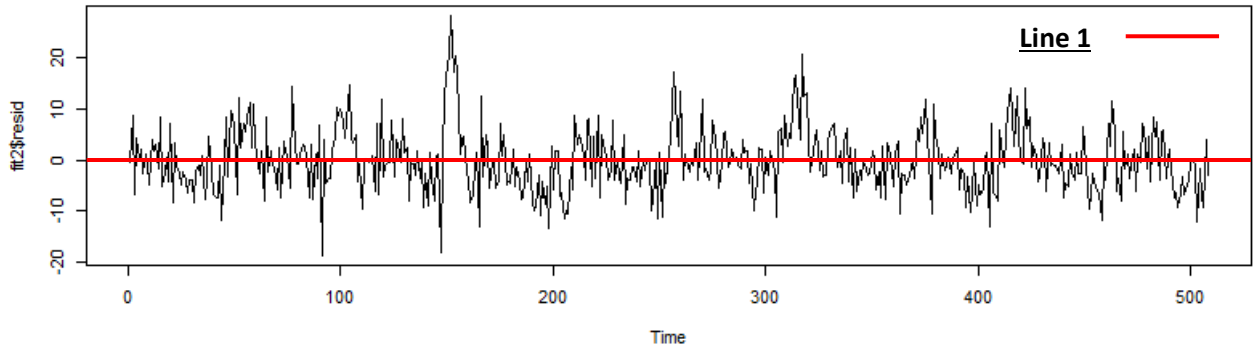
Section 3 Error Analysis as a Time Series

In order to develop a better prediction, we need to study the correlation of the residual process given by fit2, which can be seen in Figure 3.1 below. We use the Auto-correlation function (ACF) available in R to study this correlation, which measures the relationship between a variable's current values and its historical values over successive time intervals3. The ACF will allow us to assess whether or not the time series is dependent on its past. Figure 3.2 shows the ACF of the residuals of fit2. The trend given by the correlation almost matches the trend that would prompt the use of the Auto-Regressive 1 (AR1) model. Autoregressive (AR) refers to a model that shows a changing variable that regresses on its prior values. AR1 is AR parametrizes so that the current observation is only dependent on the previous observation. Line 1 in Figure 3.2 models the mentioned trend, while Point 1 highlights the anomaly that distorts the match of the trend given by our correlation and the trend represented by Line1. While the ACF does not imply that AR1 is an appropriate model for fit2 it does hint that AR2 may be an appropriate model. AR2 is AR parametrizes so that the current observation is dependent on the previous two observations. Due to the near match, we first use the AR1 model and analyze the prediction strength of the results given by its ACF. The R function Arima() is used to gain access to the AR model. In using Arima() we fit an AR1 model with the residuals of fit2 and call this new model ts1. The ACF in Figure 3.3 displays a correlation of the ts1 model residuals. We observe that the trend given by the correlation is outside of the tolerance (-0.1,0.1), of a robust prediction model. We also observe that the trend tends to change from high to low with each step. The results of ACF of ts1 and our assumption gained from analyzing Figure3.2 motivates the fitting of an AR(2) model.

We call the AR2 model ts2. After fitting the AR2 model with the residuals of fit2 we examine the prediction strength of the results given by its ACF in Figure 3.4. Examining the correlation demonstrated by ACF of the residuals of ts2 shows that the correlation is less than 0.10 and can be ignored, meaning that the fitting ts2 is a very strong fitting for error. The AR2 model ts2 can be found in Table 3.1. Evaluating Table 3.1, we find that AR2 is larger than AR1. An interpretation of this implies that the mortality rate of a week is highly dependent on the two weeks that preceded it. We can develop our final model by taking the standard error values found in the table as the parameters for model ts2. The final model Y_t takes the change point model fit2, which was shown to be a stronger fitting than the linear model, and adds to it the Auto Regression model fitting for error ts2, which increases the prediction strength of fit2.

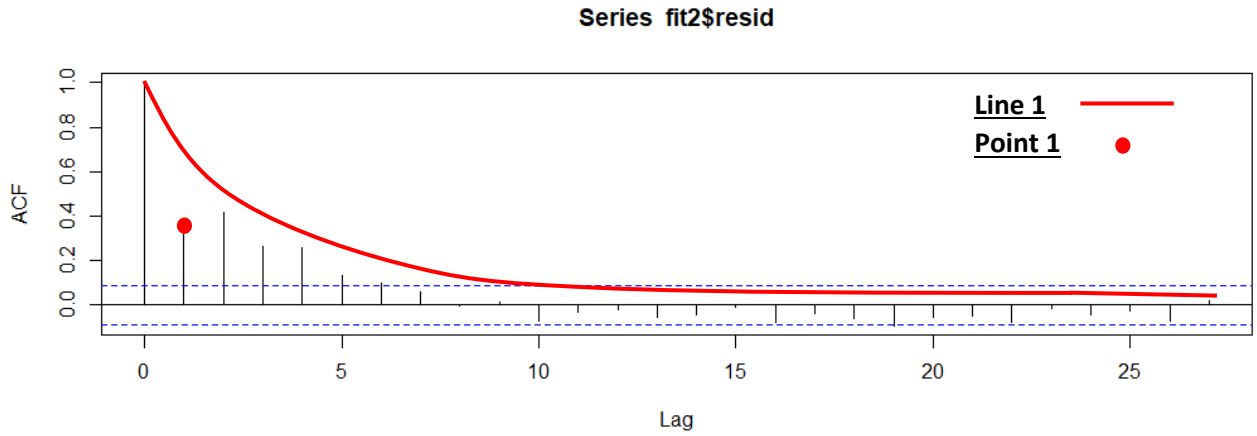
>plot.ts(fit2$resid)

Figure 3.1(>plot.ts(fit2$resid)) **Residuals**



>acf(fit2$resid)

Figure 3.2(>acf(fit2$resid))

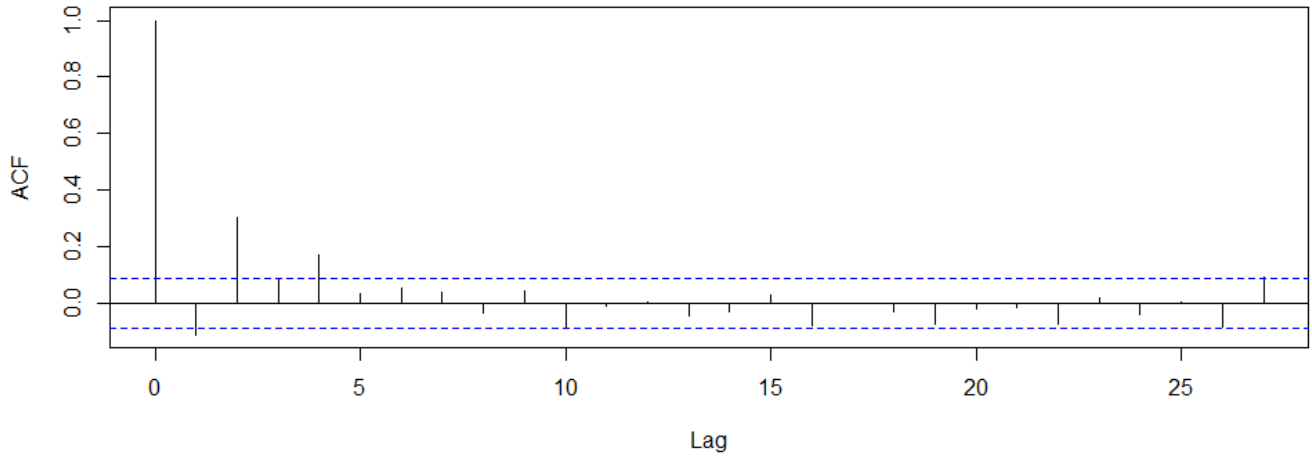


Fitting an AR(1) model

ts1<- arima(fit2$resid,order=c(1,0,0))

>acf(ts1$resid)

Figure 3.3(>acf(ts1$resid)) **Series ts1$resid**

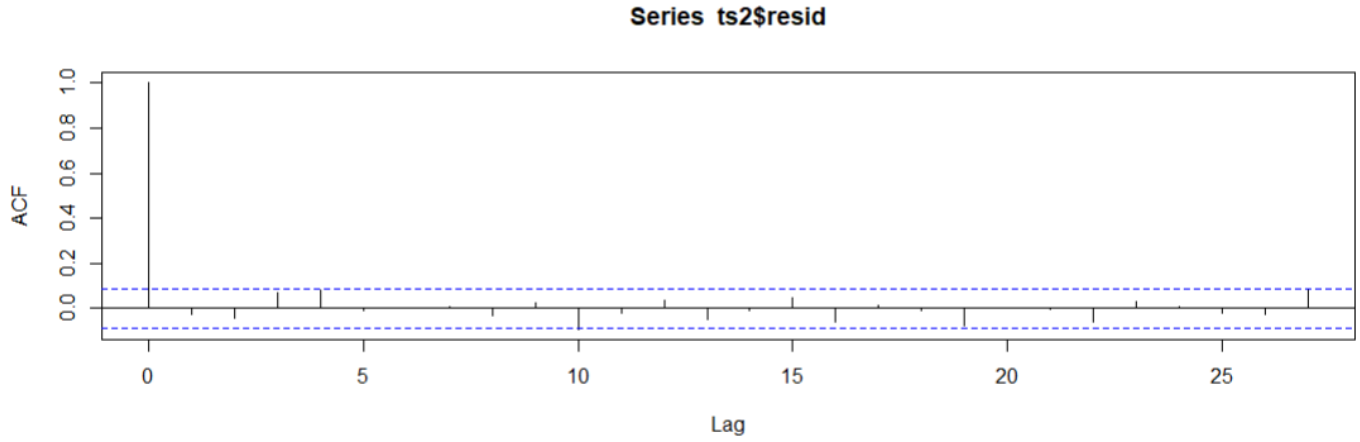


Fitting an AR(2) model

> ts2<-arima(fit2$resid,order=c(2,0,0))

>acf(ts2$resid)

Figure 3.4 (>acf(ts2$resid))



>ts2

Table 3.1 (>ts2)

Call: $Z_t = \alpha_1 Z_{t-1} + \alpha_2 Z_{t-2} + \varepsilon_t$, arima(x=fit2$resid, order = c(2,0,0))				
Coefficients				
	AR1	AR2	Intercept	
	0.2057	0.3504	0.0112	
Standard error.	0.0415	0.0416	0.5485	
Sigma^2 estimated as 30.35: log likelihood = -1587.84, aic = 3183.69				

Final Model

- $Y_t = 70.79 + 8.364I[t \leq 262] - 0.4680 \times temp + 0.022 \times temp^2 + 0.2493 \times part + Z_t$
  a. $Where\ Z_t = 0.0415\alpha_1 Z_{t-1} + 0.0416\alpha_2 Z_{t-2} + 0.5485$

Conclusion

In this project, we analyzed the weekly mortality of LA county from 1970 to 1979 due to circular vascular complications with respect to temperature and pollution variables. Describing the mortality, temperature, and pollution data sets and visualizing them graphically, we identified possible dependencies. We de-trended the data set by fitting a strong linear regression model in order to then perform a time series analysis. Treating the temperature and pollution as covariates, we showed through a comparison of P-values and F-values that the change-point model is a significantly better fitting than the linear model. Lastly, we conducted a time series analysis of the residuals to show that AR(2) gives a quite satisfactory fitting for the errors and allows for better predictions.

Bibliography

1. Matloff, Norman, The Art of R Programming, San Francisco, So Starch Press, 2011
2. R-data: Package 'astsa'
3. R-data: cmort: Cardiovascular Mortality from the LA Pollution study
4. R-data: tempr: Temperatures from the LA pollution study
5. R-data: Part: Particulates from the LA pollution study
6. Zach, How to Calculate Autocorrelation in R – Statology, Statology, https://www.statology.org/autocorrelation-in-r/
7. Top Schools, What Is ARIMA Modeling?, Master's in Data Science, https://www.mastersindatascience.org/learning/what-is-arima-modeling/

Index:

1. Source Code

```
>library(astsa)
>par(mfrow=c(3,1))
>plot(cmort)
>plot(tempr)
>plot(part)

>pairs(cbind(cmort,tempr,part))
>summary(cmort); summary(tempr); summary(part)
>temp<-tempr-mean(tempr)
>temp2<-temp^2
>fit0<-lm(cmort~temp+temp2+part)
>summary(fit0)
>trend<-time(cmort)
>fit1<-lm(cmort~trend +temp+temp2+part)
>summary(fit1)
>AIC(fit1)
> anova(fit0,fit1)

> fv <-numeric(507)
>  for(i in 1:507){
+ time<-c(rep(1,i),rep(0,508-i))
+ fit2<-lm(cmort~time+temp+temp2+part)
+ fv[i]<-anova(fit0,fit2)$F[2] }
> plot(fv)
> c(1:507)[fv==max(fv)]
> time<-c(rep(1,262),rep(0,508-262))
> fit2<-lm(cmort~time+temp+temp2+part)
> anova(fit0,fit2)
> AIC(fit2)
> AIC(fit1)
> summary(fit2)
```