Report on Chicago Data Set

Saikat Bera, Presidency University, KOLKATA

December 4, 2022

Abstract

The data set chicago, in the package gamair, contains data about air pollution and the death rate in Chicago from 1 January 1987 to 31 December 2000. Our response variable of interest is death, the total number of non-accidental deaths each day. The other variables in the data set are time, recorded in days before or after 31 December 1993, and five possible predictor variables:

- pm10median: the median density over the city of large pollutant particles
- pm25median: the median density of smaller pollutant particles
- o3median: the median concentration of ozone (O3) in the air
- so2median: the median concentration of sulfur dioxide (SO2) in the air
- tmpd: the mean daily temperature.

Loading the dataset in R,

```
##      death         pm10median          pm25median         o3median
##  Min.   : 69.0   Min.   :-37.3761   Min.   :-16.426   Min.   :-24.779
##  1st Qu.:105.0   1st Qu.:-13.1082   1st Qu.: -6.588   1st Qu.:-10.232
##  Median :114.0   Median : -3.5391   Median : -1.326   Median : -3.326
##  Mean   :115.4   Mean   : -0.1464   Mean   :  0.243   Mean   : -2.179
##  3rd Qu.:124.0   3rd Qu.:  8.3029   3rd Qu.:  5.344   3rd Qu.:  4.468
##  Max.   :411.0   Max.   :320.7248   Max.   : 38.150   Max.   : 43.688
##                  NA's   :251        NA's   :4387
##    so2median            time            tmpd
##  Min.   :-8.2061   Min.   :-2556   Min.   :-16.00
##  1st Qu.:-2.6894   1st Qu.:-1278   1st Qu.: 35.00
##  Median :-1.2183   Median :    0   Median : 51.00
##  Mean   :-0.6361   Mean   :    0   Mean   : 50.19
##  3rd Qu.: 0.8316   3rd Qu.: 1278   3rd Qu.: 67.00
##  Max.   :28.9034   Max.   : 2556   Max.   : 92.00
##  NA's   :27
```

Examining the variables:

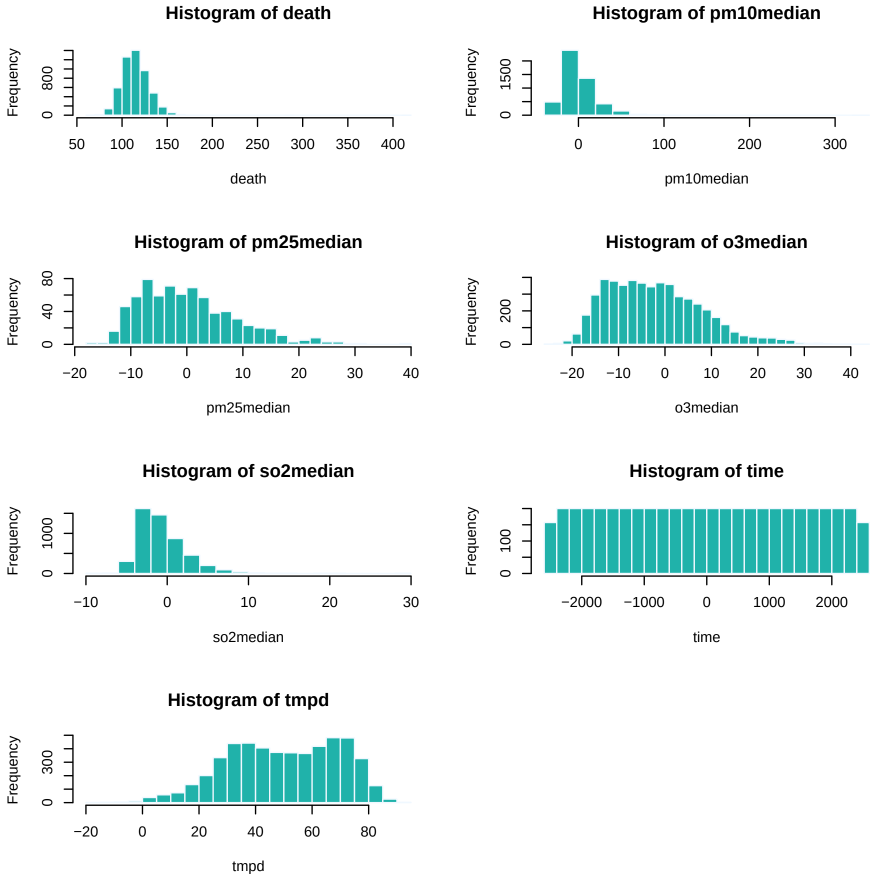
```
## [1] 92
```

Clearly, if the temperature is measured in degrees Celsius then 92 degree Celsius is extremely high to be the temperature of a city. So, the temperature must be given in degrees Fahrenheit.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## -16.426  -6.588  -1.326   0.243   5.344  38.150    4387
## [1] 5114
```

Clearly, 4387 values of the variable **pm25median** are missing among the total 5114 observations. We cannot work with such a variable with so many missing values. So, we shall ignore this variable.

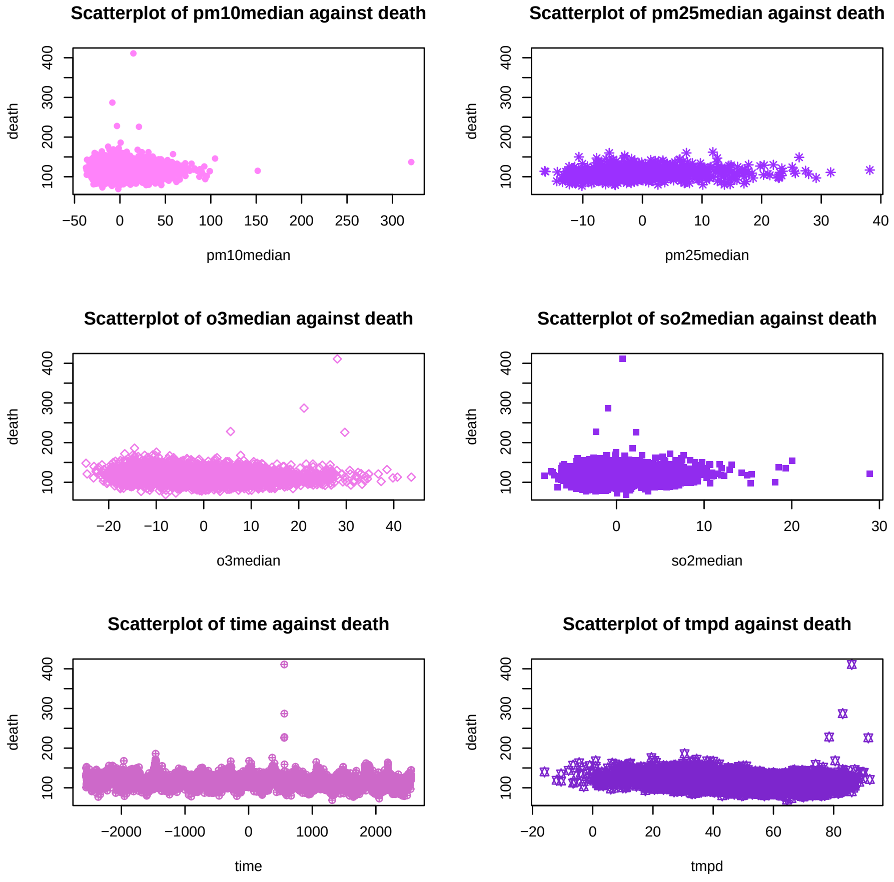
```
##       mean variance median
## 1 115.4189 234.0522    114
##         mean variance    median
## 1 -0.1463896 370.7924 -3.539062
##        mean variance    median
## 1 0.2430526  75.3241 -1.325843
##        mean variance    median
## 1 -2.179377 104.1139 -3.325857
##         mean variance    median
## 1 -0.6360707 8.562395 -1.218264
##   mean variance median
## 1    0  2179843      0
##       mean variance median
## 1 50.19329 378.7697     51
```



From the histograms we can learn a lot of things regarding the variables,

- From the histogram of the death variable we can observe that, the central tendency in near 115 deaths per day, the distribution is almost symmetric and is close to normally distributed. There are many outliers present.
- From the histogram of the pm10median variable we can observe that, the central tendency in near 0, the distribution is positively skewed with a high number of potential outliers.
- From the histogram of the pm25median variable we can observe that, the central tendency in near 0, the distribution is positively skewed. There are no such potential outliers.
- From the histogram of the o3median variable we can observe that, the central tendency in near -5, the distribution is quite positively skewed with few potential outliers.

- From the histogram of the so2median variable we can observe that, the central tendency in near -2, the distribution is positively skewed with a high number of potential outliers.
- From the histogram of the time variable we can observe that, the central tendency in near 0, the distribution is symmetric and is close to uniformly distributed.
- From the histogram of the tmpd variable we can observe that, the central tendency in near 50, the distribution is negatively skewed.



From the scatterplots it is evident that, all the scatterplots are almost linear with slope 0, except the scatterplot of death against time.

So, the response variable death may not have any type of relation (i.e. may not be dependent on) with any of the predictor variables pm10median, pm25median, o3median, so2median and tmpd. There is a sinusoidal relation between time and death.

From the scatterplots clearly we can see that there are outliers in each plot.

The corresponding days where we see outliers in the plot of pm10median against death:

```
##   [1] -2515.5 -2514.5 -2458.5 -2449.5 -2388.5 -2354.5 -2353.5 -2184.5 -2063.5
##  [10] -2040.5 -2039.5 -2033.5 -2026.5 -2020.5 -2004.5 -2003.5 -2002.5 -2001.5
##  [19] -1955.5 -1951.5 -1787.5 -1690.5 -1689.5 -1688.5 -1668.5 -1649.5 -1639.5
##  [28] -1603.5 -1602.5 -1561.5 -1540.5 -1450.5 -1349.5 -1332.5 -1308.5 -1307.5
##  [37] -1298.5 -1252.5 -1223.5 -1222.5 -1206.5  -963.5  -961.5  -960.5  -955.5
##  [46]  -948.5  -935.5  -919.5  -918.5  -898.5  -897.5  -896.5  -882.5  -858.5
##  [55]  -823.5  -697.5  -666.5  -609.5  -603.5  -599.5  -595.5  -589.5  -588.5
##  [64]  -563.5  -562.5  -548.5  -547.5  -501.5  -471.5  -457.5  -456.5  -455.5
##  [73]  -434.5  -430.5  -429.5  -330.5  -291.5  -251.5  -218.5  -217.5  -142.5
##  [82]  -139.5  -126.5   -85.5   -84.5   -81.5   -67.5    44.5    77.5    83.5
##  [91]    93.5   112.5   114.5   115.5   140.5   155.5   165.5   166.5   167.5
## [100]   170.5   230.5   236.5   238.5   244.5   254.5   257.5   263.5   279.5
## [109]   293.5   300.5   515.5   530.5   531.5   532.5   557.5   558.5   576.5
## [118]   603.5   606.5   612.5   635.5   637.5   648.5   650.5   654.5   831.5
## [127]   869.5   902.5   909.5   910.5   947.5   976.5   977.5   978.5  1200.5
## [136]  1214.5  1220.5  1266.5  1270.5  1301.5  1353.5  1354.5  1356.5  1370.5
## [145]  1371.5  1374.5  1500.5  1545.5  1546.5  1595.5  1598.5  1599.5  1608.5
## [154]  1635.5  1654.5  1655.5  1728.5  1760.5  1914.5  1949.5  1950.5  1998.5
## [163]  2020.5  2021.5  2022.5  2069.5  2070.5  2071.5  2094.5  2110.5  2126.5
## [172]  2127.5  2128.5  2133.5  2142.5  2147.5  2148.5  2244.5  2258.5  2287.5
## [181]  2295.5  2308.5  2315.5  2316.5  2319.5  2349.5  2351.5  2352.5  2398.5
## [190]  2428.5  2452.5  2453.5  2465.5  2476.5  2477.5  2482.5  2484.5  2490.5
```

The corresponding days where we see outliers in the plot of pm25median against death:

```
##  [1] 1595.5 1882.5 1960.5 1999.5 2071.5 2231.5 2300.5 2433.5 2487.5 2490.5
## [11] 2522.5 2540.5
```

The corresponding days where we see outliers in the plot of o3median against death:

```
##  [1] -2420.5 -2389.5 -2388.5 -2355.5 -2353.5 -2026.5 -2021.5 -2020.5 -2018.5
## [10] -2015.5 -2005.5 -2004.5 -1981.5 -1971.5 -1654.5 -1614.5 -1297.5 -1276.5
## [19] -1252.5  -925.5  -920.5  -897.5  -896.5  -895.5  -859.5  -548.5  -237.5
## [28]   149.5   168.5   533.5   538.5   539.5   558.5   559.5   560.5   908.5
## [37]   910.5   917.5   918.5  1275.5  1288.5  1638.5  1996.5  2021.5  2072.5
## [46]  2073.5  2350.5  2351.5  2401.5
```

The corresponding days where we see outliers in the plot of so2median against death:

```
##   [1] -2553.5 -2551.5 -2543.5 -2530.5 -2506.5 -2505.5 -2499.5 -2477.5 -2474.5
##  [10] -2458.5 -2457.5 -2449.5 -2436.5 -2432.5 -2431.5 -2388.5 -2310.5 -2270.5
##  [19] -2269.5 -2254.5 -2217.5 -2205.5 -2184.5 -2181.5 -2164.5 -2158.5 -2147.5
##  [28] -2142.5 -2039.5 -2004.5 -1938.5 -1864.5 -1863.5 -1862.5 -1828.5 -1826.5
##  [37] -1825.5 -1764.5 -1759.5 -1758.5 -1743.5 -1712.5 -1540.5 -1475.5 -1474.5
##  [46] -1469.5 -1468.5 -1463.5 -1395.5 -1352.5 -1253.5 -1206.5 -1158.5 -1143.5
##  [55] -1100.5 -1099.5 -1095.5 -1082.5 -1069.5 -1064.5 -1063.5 -1061.5  -882.5
##  [64]  -870.5  -868.5  -857.5  -856.5  -782.5  -771.5  -770.5  -768.5  -715.5
##  [73]  -702.5  -643.5  -633.5  -603.5  -602.5  -595.5  -588.5  -548.5  -456.5
##  [82]  -375.5  -354.5  -353.5  -352.5  -351.5  -349.5  -346.5  -345.5  -339.5
##  [91]  -331.5  -330.5  -329.5  -306.5  -305.5  -239.5  -237.5   -30.5   -18.5
## [100]   -17.5     7.5     9.5    18.5    19.5    20.5    21.5    31.5    42.5
## [109]    90.5   285.5   322.5   346.5   347.5   354.5   355.5   391.5   635.5
## [118]   801.5   908.5   909.5   910.5  1074.5  1109.5  1110.5  1125.5  1126.5
## [127]  1205.5  1213.5  1444.5  1445.5  1482.5  1491.5  1501.5  1796.5  1806.5
## [136]  1841.5  1899.5  2031.5  2073.5  2126.5  2137.5  2191.5  2250.5  2257.5
## [145]  2274.5  2281.5  2308.5  2311.5  2350.5  2351.5  2477.5  2495.5  2496.5
## [154]  2548.5  2551.5
```

The corresponding days where we see outliers in the plot of time against death:

```
## numeric(0)
```

The corresponding days where we see outliers in the plot of tmpd against death:

```
## [1] 17.5
```

```
## [1] 3
## [1] 13
## [1] 22
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 9
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
```

- Hence, the plot of pm10median against death and the plot of pm25median against death share 3 outlier days.
- The plot of pm10median against death and the plot of o3median against death share 13 outlier days.
- The plot of pm10median against death and the plot of so2median against death share 22 outlier days.
- The plot of o3median against death and the plot of so2median against death share 9 outlier days.

From the scatterplots it is evident that, all the scatterplots are almost linear with slope 0, except the scatterplot of death against time. So, except time for each predictor variable , fitting a linear regression model of the number of deaths on the predictor will be meaningless.

Only for the time variable, we can justify the modeling assumptions to hold as from the scatterplot it is evident that there is some sinusoidal relation between time and death.

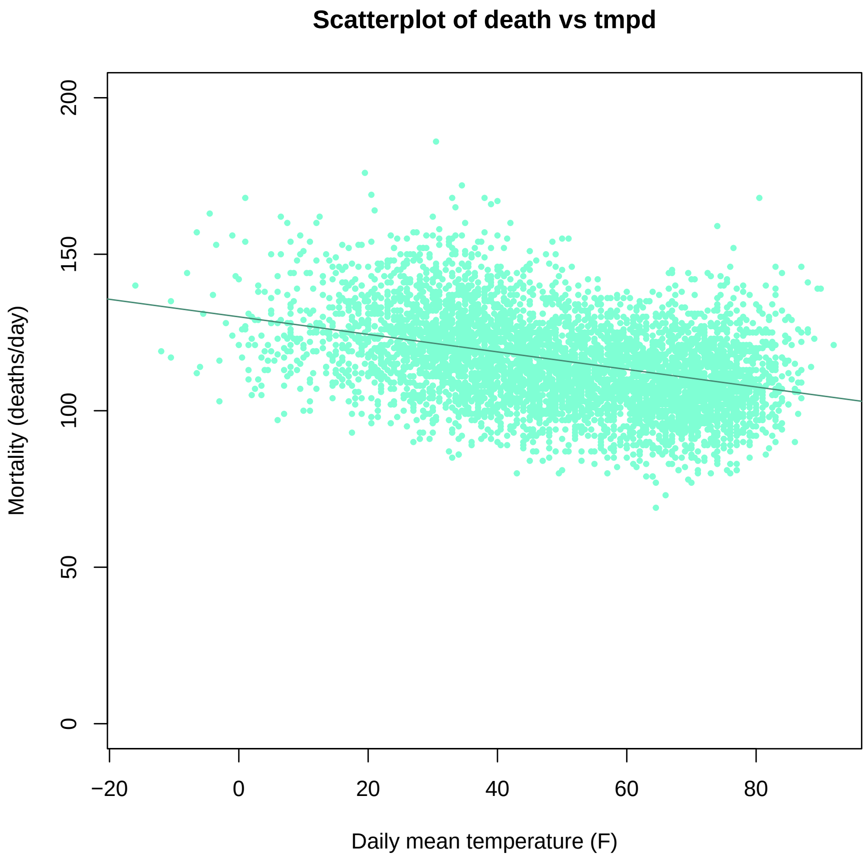
The theoretical regression model between death and tmpd is as follows:

```
##
## Call:
## lm(formula = death ~ tmpd, data = chicago)
##
## Coefficients:
## (Intercept)         tmpd
##    129.9571      -0.2896
```

The assumptions of the model are,
The distribution of X is arbitrary.
If X=x, then Y=$\alpha + \beta x + \epsilon$, for some parameters α and β, and some random noise variable ϵ.
For all x, E[$\epsilon | X = x$]=0, Var[$\epsilon | X = x$]=σ^2.
ϵ is uncorrelated across observations.

The interpretation of the constant term in the proposed regression function is that, on an average, the expected number of non-accidental deaths per day in Chicago is 130 when the mean daily temperature is $0^{\circ}F$.

The interpretation of the slope in the proposed regression function is that, if we select two sets of cases from the un-manipulated distribution where the mean daily temperature differs by $1^{\circ}F$, we expect the number of non-accidental deaths per day in Chicago to differ by 0.28.



We can assume that the model's error variables follow a normal distribution. Since, the noise might be due to adding up the effects of lots of little random causes, all nearly independent of each other and of X, where each of the effects are of roughly similar magnitude. Then due to the central limit theorem the sum of random errors will indeed be pretty Gaussian.

Also, the Gaussian noise model helps us to work out a complete theory of inference and prediction for the model by helping us to find closed forms for estimates of the parameters,variances etc.

The relation between the Fahrenheit scale and the Celsius scale is: $F = \left(\frac{9}{5} \times C\right) + 32$.

So, $2^{\circ}C$ is equivalent to $35.6^{\circ}F$.

Now, for unit increase in temperature, we can expect the number of deaths to decrease by 0.28 per day, according to the proposed linear regression model.

So, for an increase of $35.6^{\circ}F$ in temperature, we can expect the number of deaths to decrease by $(0.28 \times 35.6 =)\, 9.968$ per day.

Hence,the predicted change in number of deaths in a year will be, $9.968 \times 365 = 3638.32 \approx 3638$.

So, for $2°C$ increase in average temperature over the course of a whole year, we can expect the number of deaths to decrease by 3638 over the year.

No, we cannot claim the relationship between temperature and deaths casual.

Since non-accidental deaths can differ by some other reason such as, cancer or some severe illness, suicide etc, i.e. there existsa third variable which is the underlined factor of such relationship between temperature and deaths.

References:

Roger D. Peng, Leah J. Welty and Aiden McDermott. R package NMMAPSdata.