

# Practical Assignment I

## Econometrics

Aritra Basak  
CRN: STAT003  
UG Sem 6

May 3, 2022

### Analysis of the chicago dataset:

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.

```
## Warning: package 'gamair' was built under R version 4.2.0
```

Here, 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.

### ❖ Summary of the data:

```
##      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
```

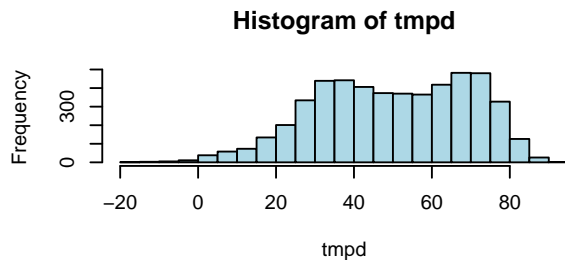
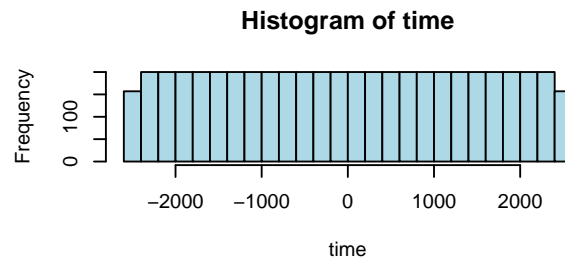
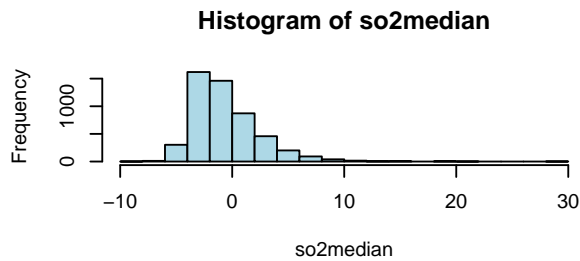
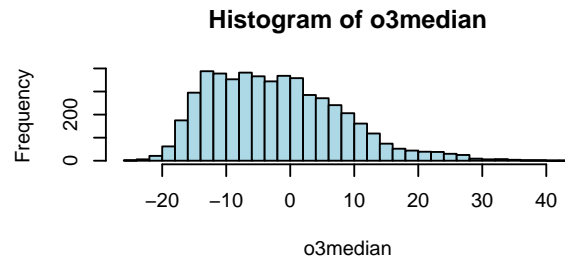
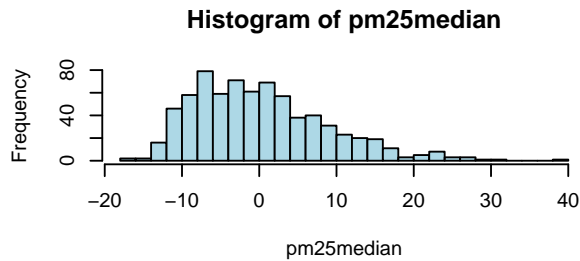
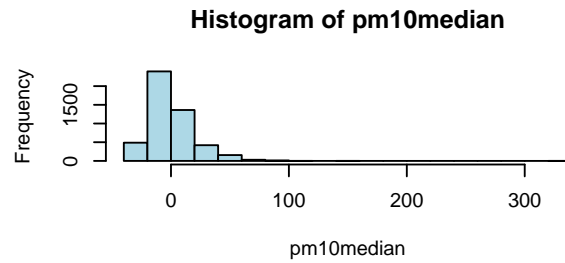
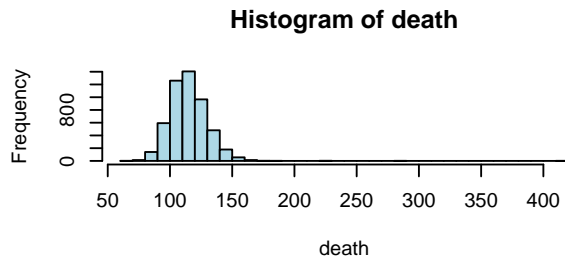
The dataset contains 5114 number of observations.

❖ **Important points regarding the variables to be noted:**

- **Unit of the Temperature:** The maximum temperature is 92 degree Celsius. 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.
- **Ignorance of the pm25median variable:** We shall ignore the pm25median variable in the rest of this problem set as 4387 values of that variable are missing among the total 5114 observations. We cannot work with such a variable having so many missing values.
- **Mean, variance and median of each variable:**

Variable name	Mean	Variance	Median
Death	115.4189	234.0522	114
pm10median	-0.1463896	370.7924	-3.539062
pm25median	0.2430526	75.3241	-1.325843
o3median	-2.179377	104.1139	-3.325857
so2median	-0.6360707	8.562395	-1.218264
time	0	2179843	0
tmpd	50.19329	378.7697	51

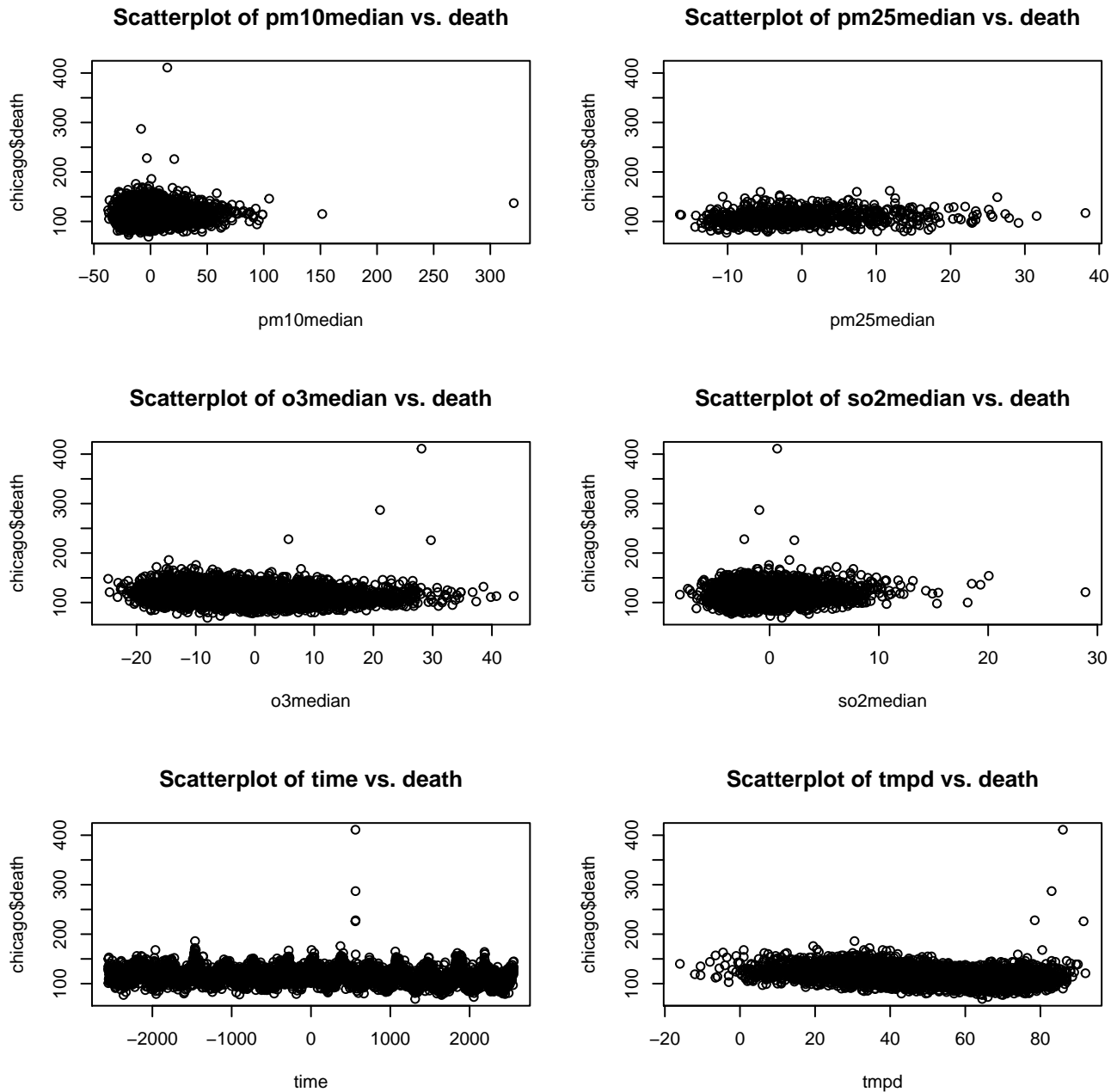
- **Histogram for each variable:**



## Observations:

- **Histogram of death:** The average number of deaths is almost 115 per day, the distribution is more or less symmetric. There is some significant outliers present.
- **Histogram of pm10median:** The large pollutant particles over the city are very light (i.e. not so dense), the distribution is positively skewed having some potential outliers.
- **Histogram of pm25median:** The average median density of smaller pollutant particles is almost 0, the distribution is somewhat symmetrically distributed. There are few outliers.
- **Histogram of o3median:** The average median concentration of ozone (O3) in the air is near -5, the distribution is positively skewed with few potential outliers.
- **Histogram of so2median:** The average median concentration of sulfur dioxide (SO2) in the air is near -2, the distribution is positively skewed with some significant outliers.
- **Histogram of tmpd:** The average mean daily temperature is almost 50, the distribution is negatively skewed.

❖ Scatterplot for each predictor variable against the response variable:



Observations:

- From the scatterplots it is evident that, the bivariate relationships between each predictor and the response variable may be closely 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:

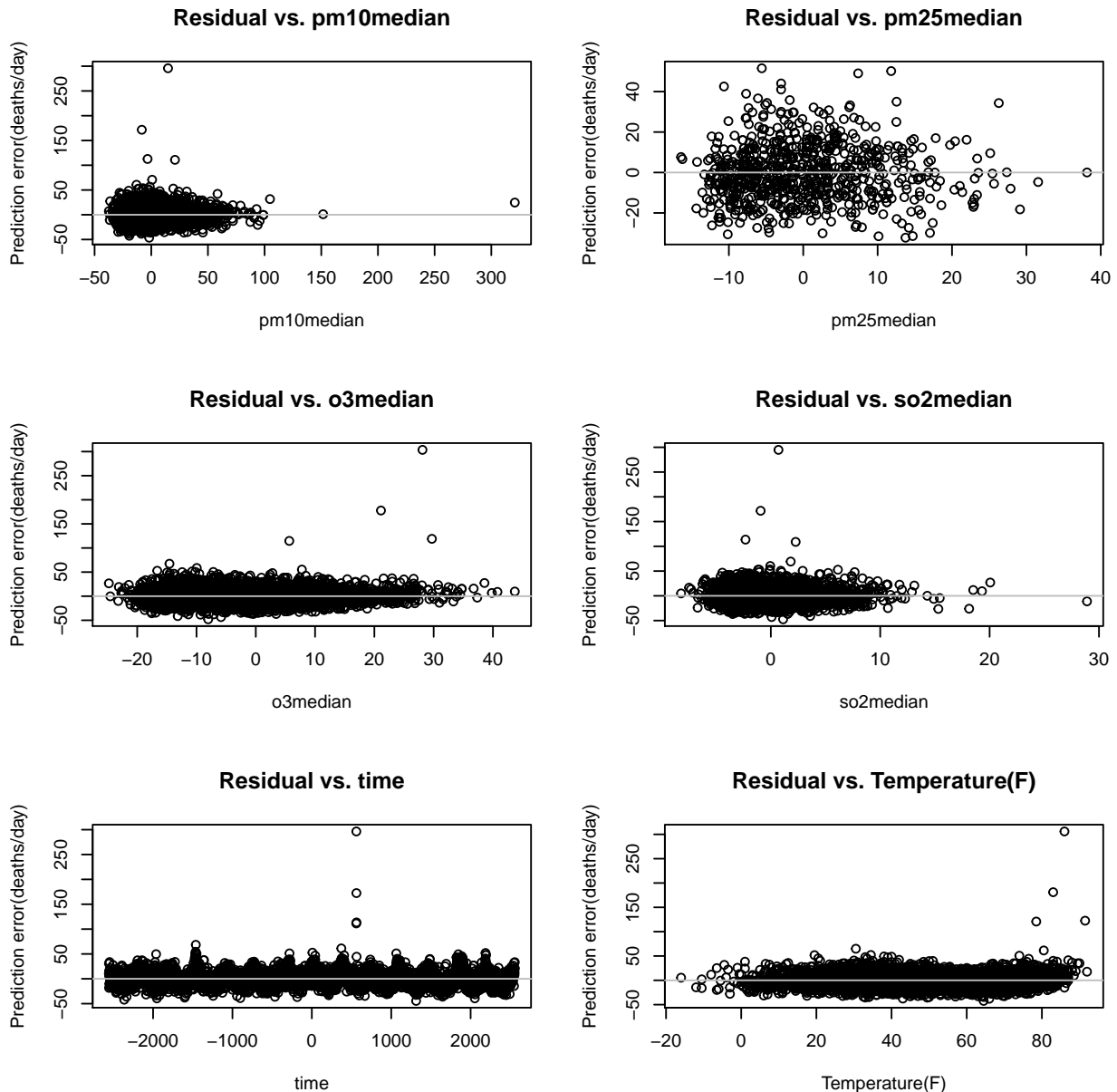
```
## numeric(0)
```

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

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

Clearly, different plots share same outlier days.

- Plotting residuals against each of the predictor variables:



Clearly, only the scatterplot with the residuals on the vertical axis and the pm25median variable on the horizontal axis, looks like a constant-width blur of points around a straight flat line at height zero. For all the other plots, there are deviations from this (in substantial regions of x-axis the average residuals are positive), indicating that a simple linear regression model of the number of deaths on the predictor would be inappropriate. Therefore, the pm25median variable would be the most appropriate for fitting a linear regression model.

### ❖ Closer observation of the relationship between death and tmpd:

We will take a closer look at the relationship between death and tmpd. Someone proposes that the relationship follows a normal error linear regression model with  $\epsilon \sim N(0, 14.22)$  and the true regression function  $E[Y|X = x] = 130 - 0.28x$ .

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

$$Y = 130 - 0.28x + \epsilon$$

The assumptions of the model are,

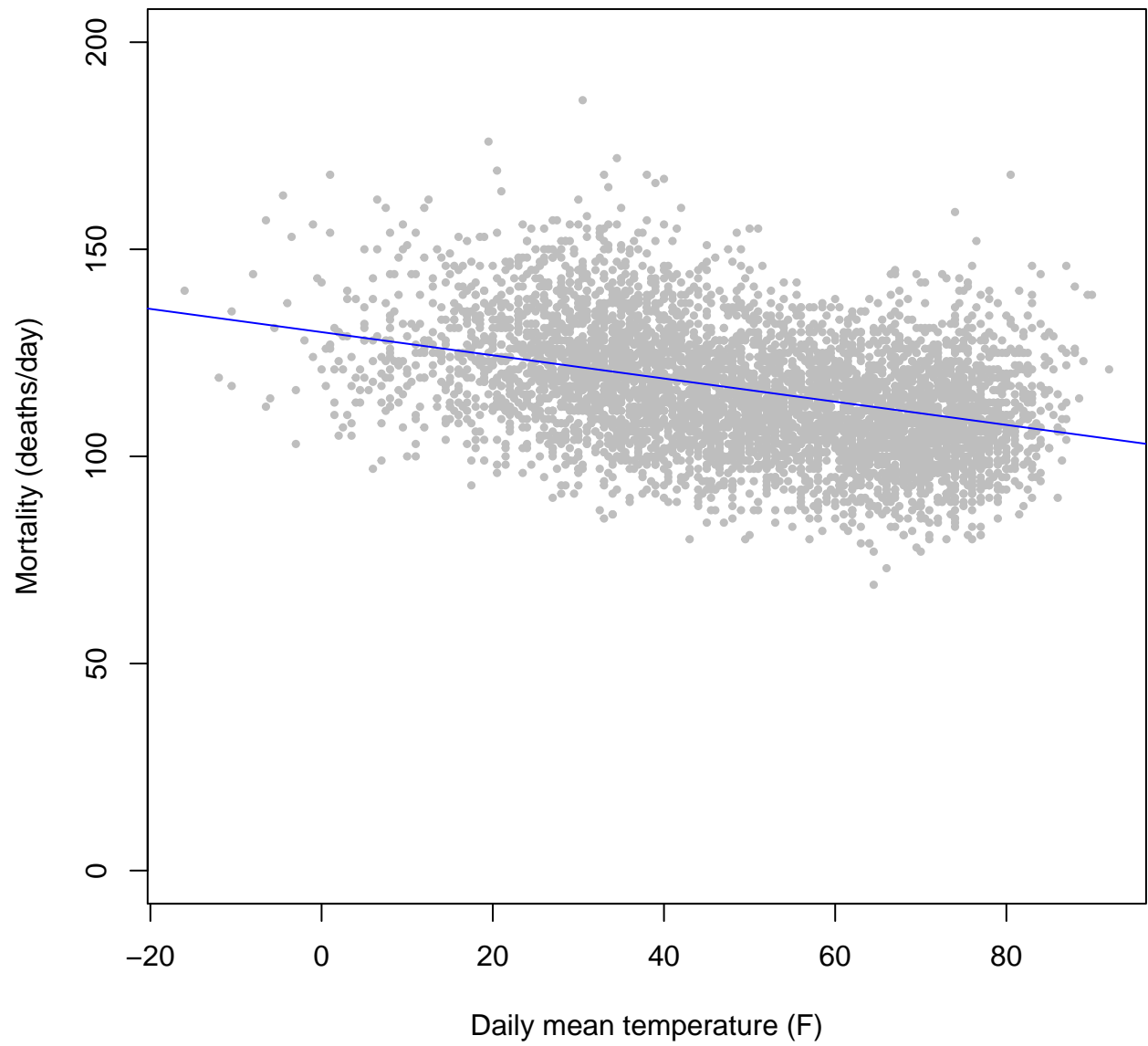
- For all  $x$ ,  $E[\epsilon|X = x] = 0$ ,  $\text{Var}[\epsilon|X = x] = \sigma^2 (= 14.22)$ . specifically,  $\epsilon \sim N(0, 14.22)$
- $\epsilon$  is uncorrelated across observations.

### **Interpretation of the proposed coefficients:**

- On an average, the expected number of non-accidental deaths per day in Chicago is 130 when the mean daily temperature is  $0^\circ F$ .
- 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.

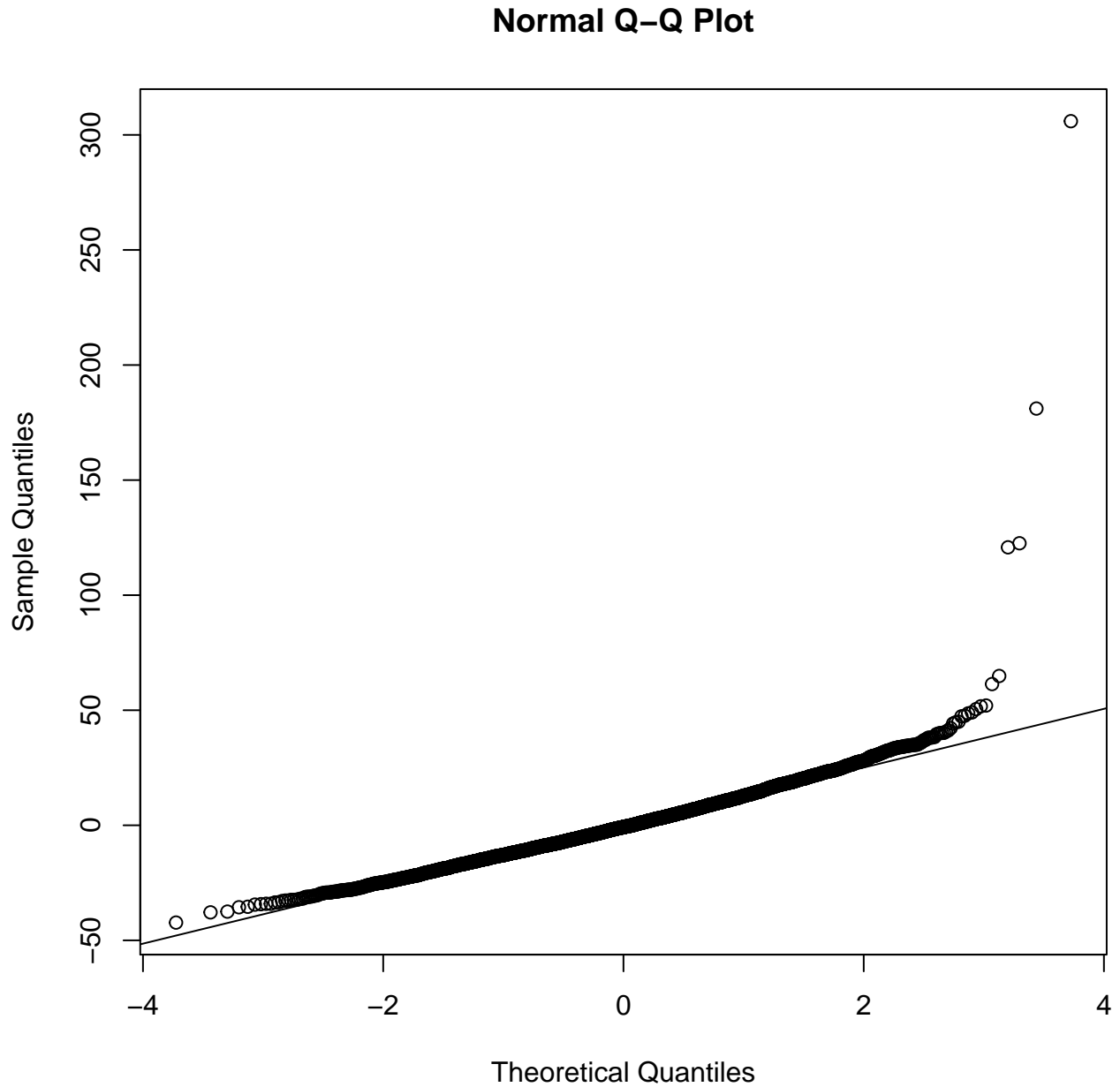
### **Scatterplot of death and tmpd along with the proposed linear function:**

**Scatterplot of death vs tmpd**



Checking of the normal error regression model assumption:  
Q-Q Plots:





Clearly, it is evident that here the normal error regression model assumption is not appropriate.

**Note:**

1. 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  $2^{\circ}C$  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.

2. The relationship between temperature and deaths is not casual.

Since non-accidental deaths can also differ by some other reasons such as, due to pollution through different pollutants etc, i.e. there exists third variable which is the underlined factor of such relationship between temperature and deaths.

## Analysis of the “econ.csv” dataset:

The data file **econ.csv** contains information about the economies of the 366 “metropolitan statistical areas” ( $\approx$  cities) of the US in 2006. In particular, it lists, for each city, the population, the total value of all goods and services produced for sale in the city that year per person (“per capita gross metropolitan product”, pcgmp), and the share of economic output coming from four selected industries.

It has 366 rows and 7 columns. It contains the name of the cities (metropolitan statistical areas) in a column corresponding to each observation along with the above mentioned 6 columns.

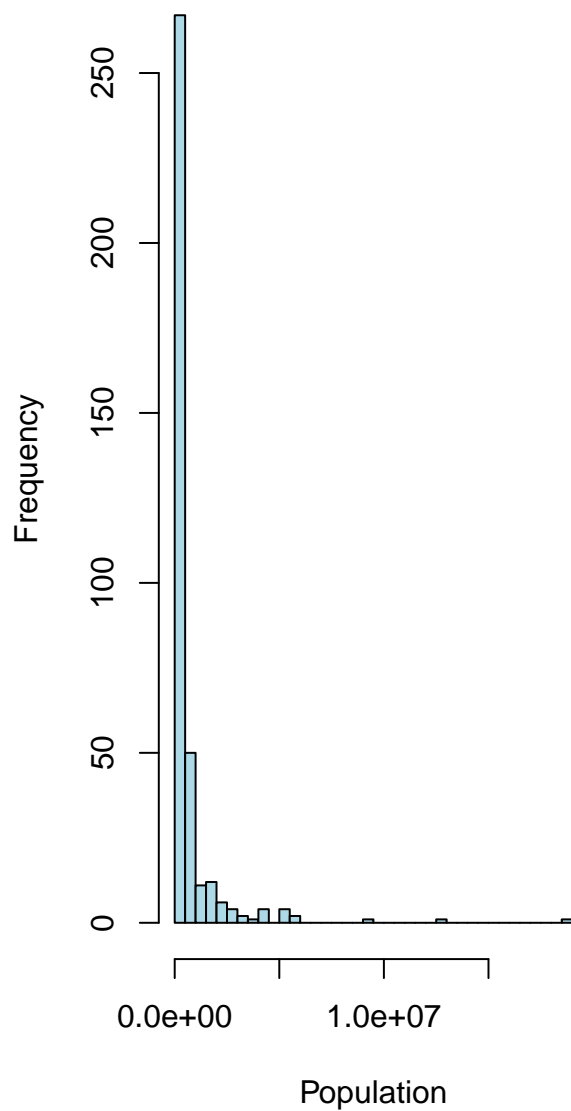
### ❖ Summary of the data:

##	pcgmp	pop	finance	prof.tech
##	Min. :14920	Min. : 54980	Min. :0.03845	Min. :0.01474
##	1st Qu.:26533	1st Qu.: 135625	1st Qu.:0.10403	1st Qu.:0.02932
##	Median :31615	Median : 231500	Median :0.14140	Median :0.04213
##	Mean :32923	Mean : 680898	Mean :0.15082	Mean :0.04905
##	3rd Qu.:38213	3rd Qu.: 530875	3rd Qu.:0.18122	3rd Qu.:0.05932
##	Max. :77860	Max. :18850000	Max. :0.38480	Max. :0.19080
##			NA's :12	NA's :112
##	ict	management		
##	Min. :0.00349	Min. :0.00042		
##	1st Qu.:0.01215	1st Qu.:0.00294		
##	Median :0.02218	Median :0.00651		
##	Mean :0.03910	Mean :0.00908		
##	3rd Qu.:0.04072	3rd Qu.:0.01191		
##	Max. :0.58600	Max. :0.05431		
##	NA's :76	NA's :157		

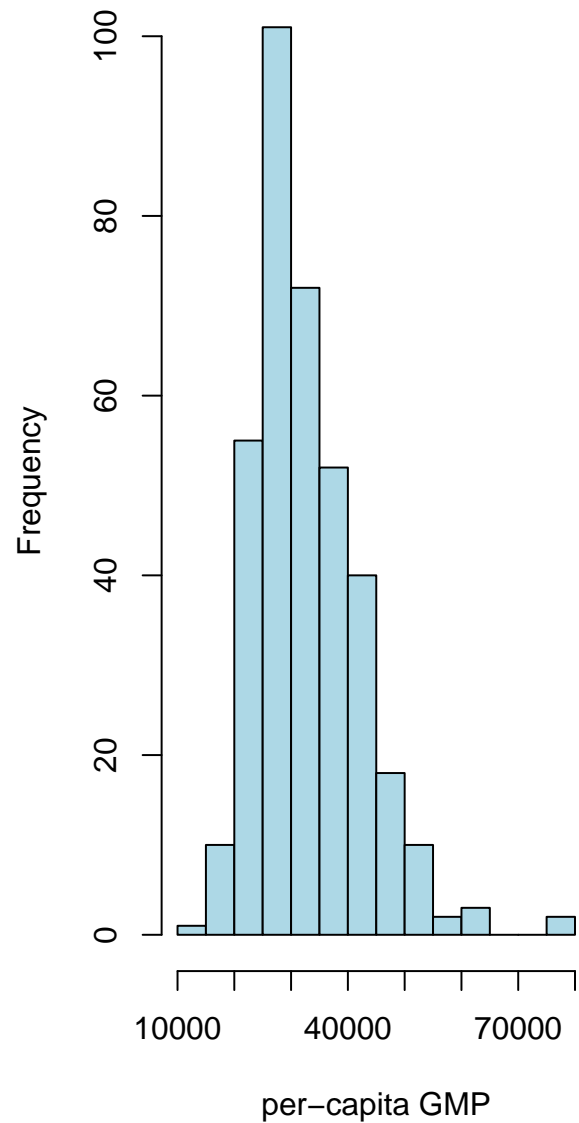
### ❖ Exploratory data Analysis:

- Histograms (univariate EDA plot) for population and for per-capita GMP:

### Histogram of Population

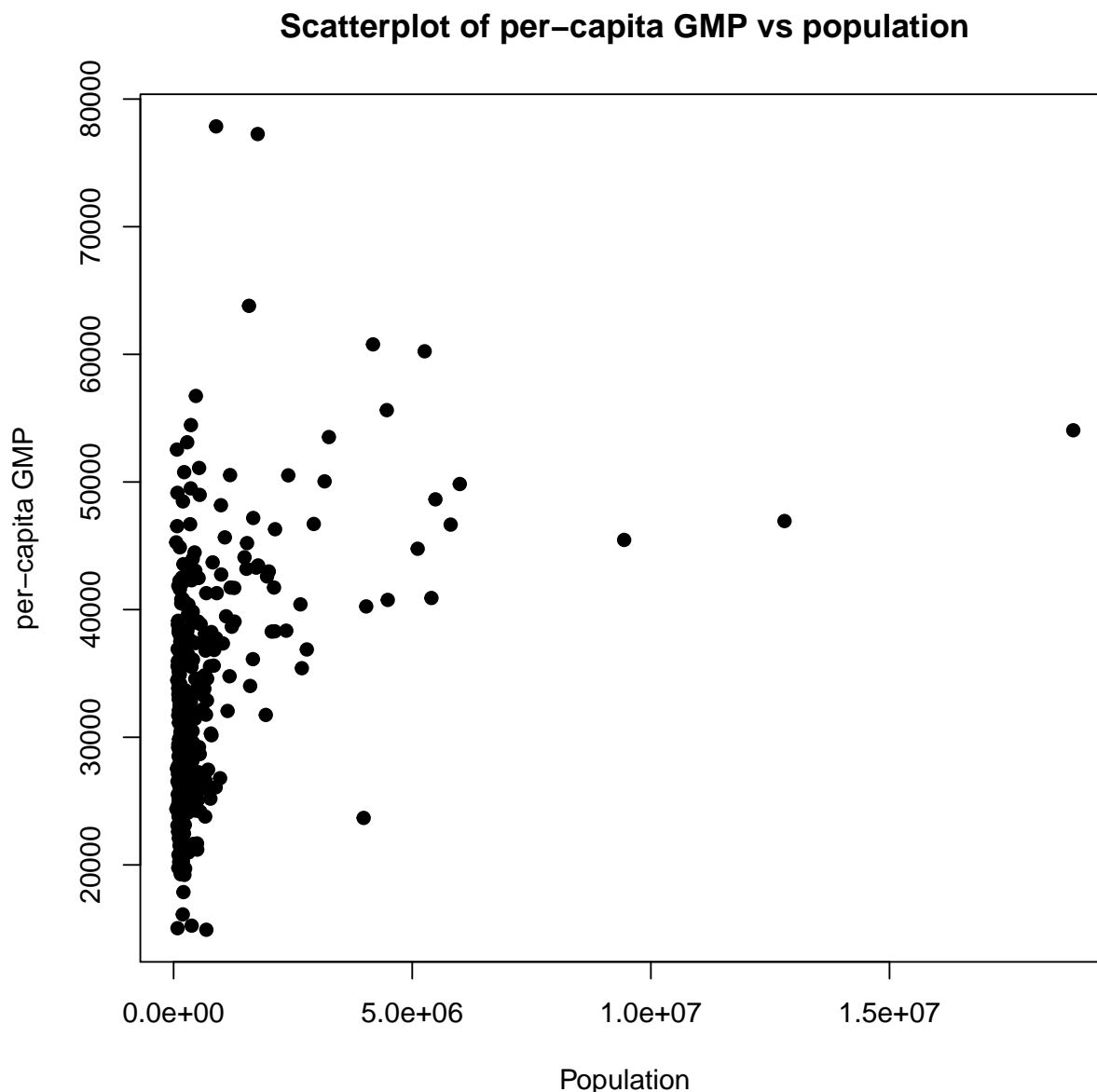


### Histogram of per-capita GMP



The distribution of population is highly positively skewed i.e. a large number of cities have a little amount of population, and very few cities have huge population. There are some outliers of excessively high magnitude. The distribution of per-capita GMP is slightly positively skewed with some potential outliers.

- Scatterplot (bivariate EDA plot) for per-capita GMP as a function of population:



• **Fitting a simple linear regression model of per-capita GMP on population:**

Suppose our model is,

$$Y = \alpha + \beta X + \epsilon$$

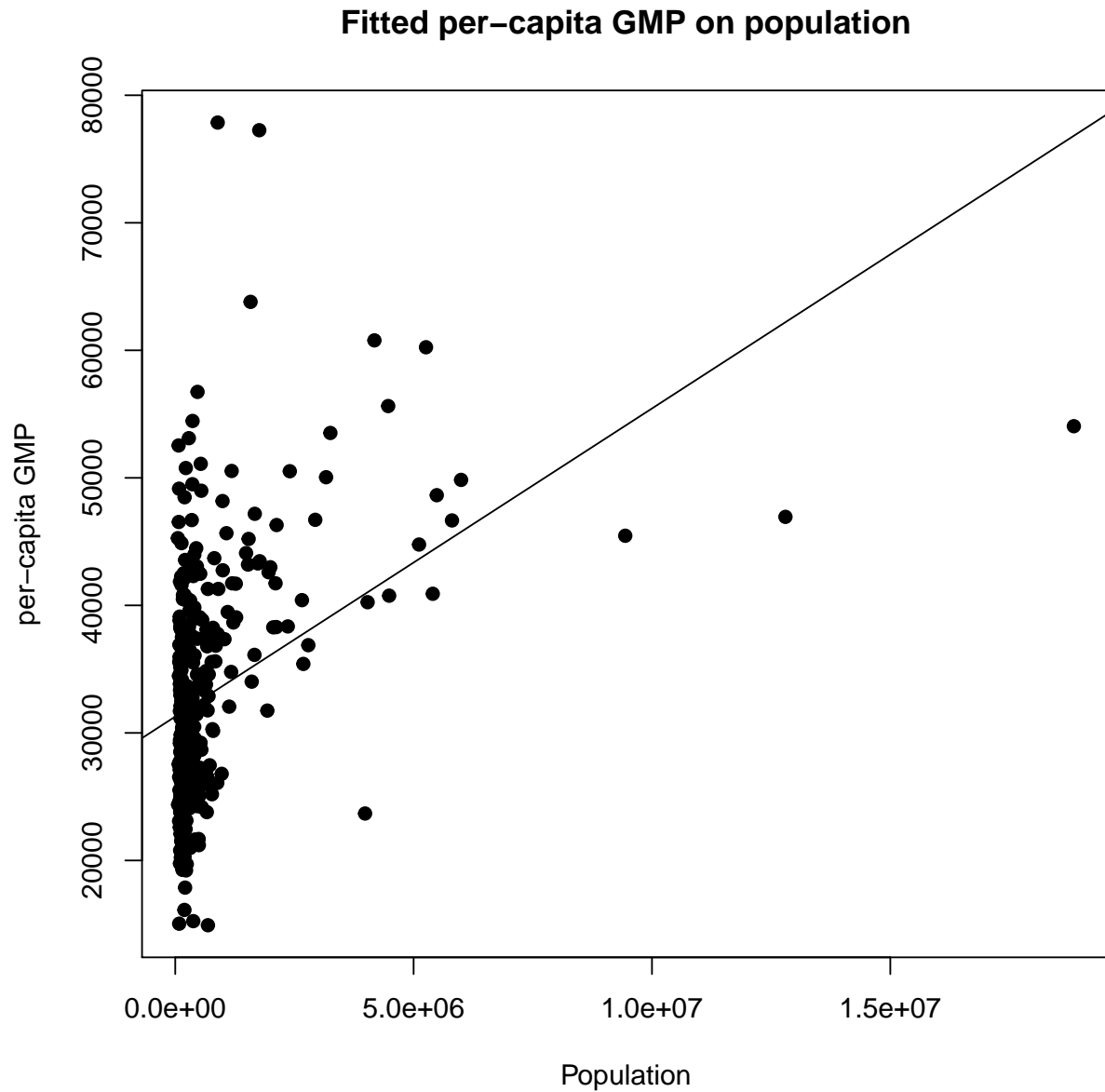
where,  $Y$  denotes per-capita GMP,  $X$  denotes population and  $\alpha$  and  $\beta$  are the parameters of the model and  $\epsilon$  is the random error term.

The assumptions of the model are,

- For all  $x$ ,  $E[\epsilon|X = x] = 0$ ,  $\text{Var}[\epsilon|X = x] = \sigma^2$ .
- $\epsilon$  is uncorrelated across observations.

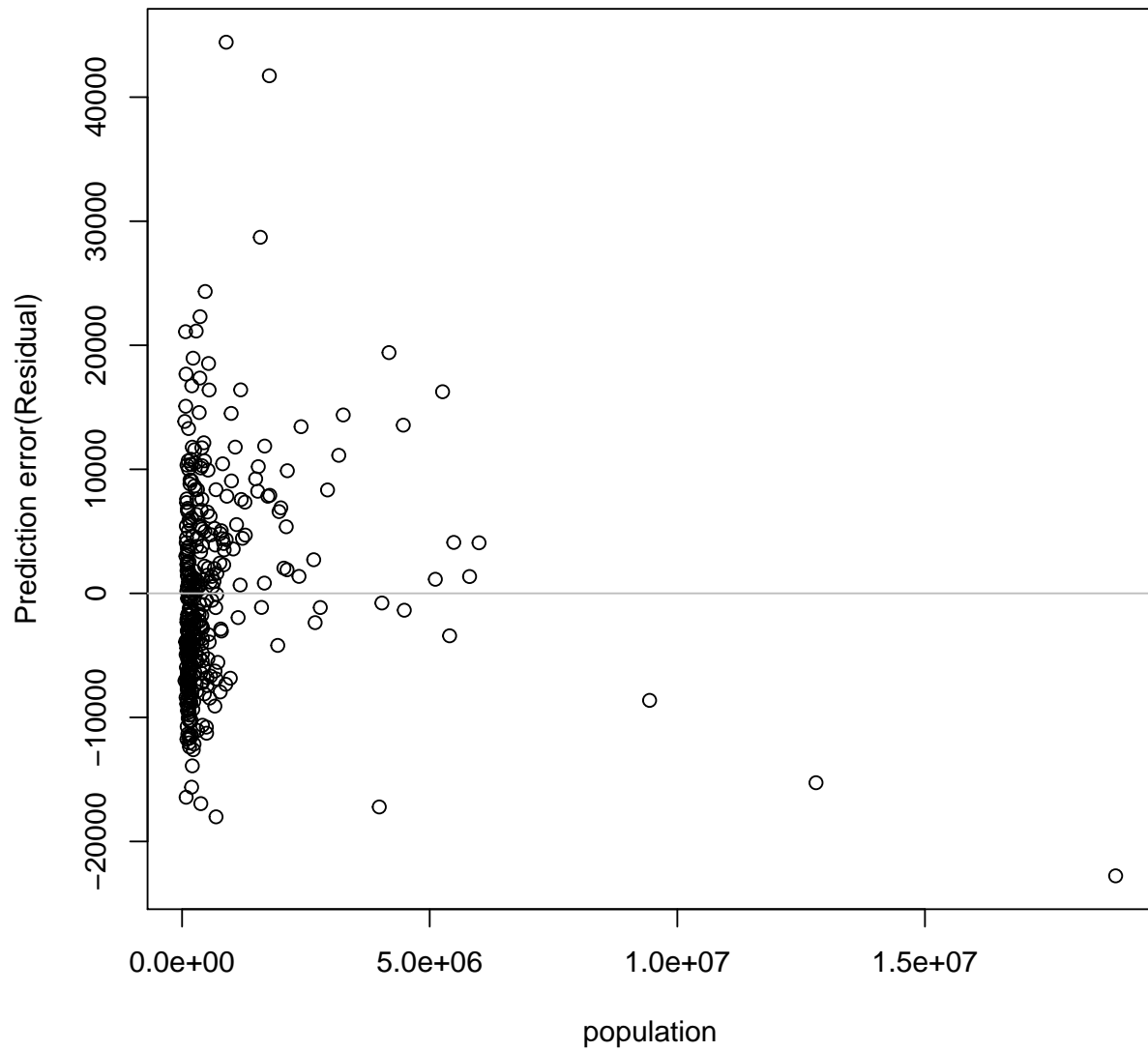
The least-square estimate of the slope is  $\hat{\beta} = 0.002416201$  and least-square estimate of the intercept is  $\hat{\alpha} = 31277.57$

These values are same as the coefficients obtained by the `lm` function.



- **Comment on the fit:** The line doesn't fit the data well.
- Verification of the assumptions of the simple linear regression model:
  - Plotted residuals against the population:

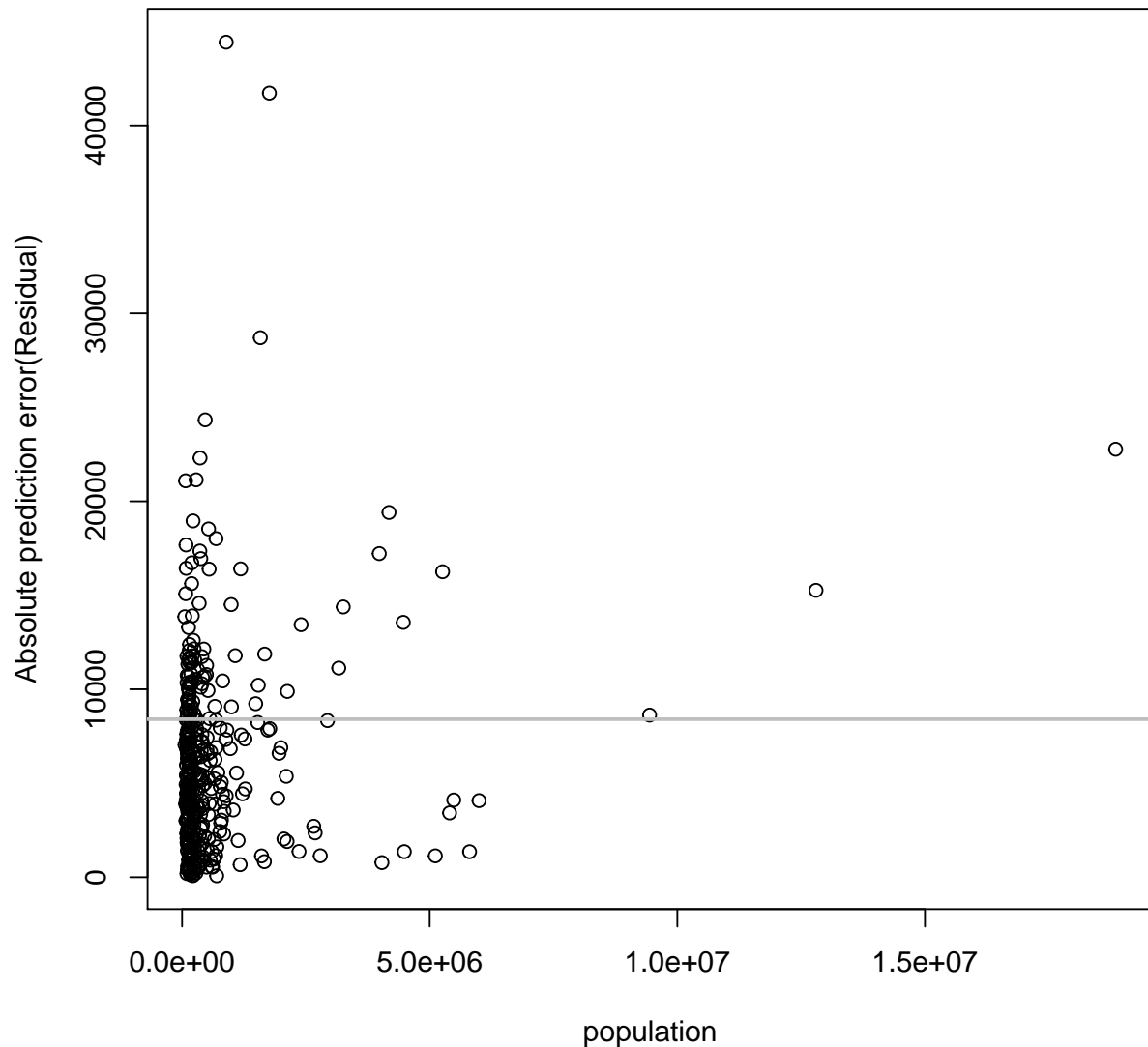
## Residual vs. population



Clearly, the scatterplot looks like a changing-width (shrinking) blur of points around a straight, flat line at height zero. This means that the simple-linear part of the simple linear regression model is wrong.

- Plotted absolute residuals against the population:

### Absolute residual vs. population



Clearly, the points are not scattered around the flat line and after  $1.0e+07$  population the residuals are persistently above zero. This could be due to non-constant noise variance (technically called “heteroscedasticity”), or due to getting the functional form of the regression wrong.

Therefore, we can say that here the assumptions of the simple linear regression model don’t hold.

#### ❖ Observations on Pittsburgh city:

- The population of Pittsburgh is 2361000.
- The per-capita GMP of Pittsburgh is 38350.
- The per-capita GMP of Pittsburgh predicted by the model is 36982.22.
- The residual for Pittsburgh is 1367.775.
- The residual square for Pittsburgh is only 2.6% (approx.) of the MSE (The mean squared error (MSE) of the regression is 70697145). Now, as the residual for Pittsburgh is greater than 1, so we can say that the residual

for Pittsburgh is quite small compared to the mean squared error.

### [ Interpretation of the estimated slope:

If we select two sets of cases from the un-manipulated distribution where the population differs by 1, we expect per-capita GMP to differ by 0.002416201 unit.]

- The predicted per-capita GMP for a city with  $10^5$  more people than Pittsburgh is 37223.84.
- If  $10^5$  people were added to the population, by a policy intervention, then the predicted Pittsburgh per-capita GMP would become more closer to the observed per-capita GMP i.e. the residual would decrease.

## Appendix:

### (R codes)

1. Analysis of the Chicago dataset:

(a) *#...loading the chicago dataset in R*  
`library(gamair)`  
`data(chicago)`

(b) *#...summary on each variable*  
`summary(chicago)`  
`dim(chicago)`

- (c) Examining the variables:

i. *#...maximum temperature in Chicago*  
`max(chicago$tmpd)`

ii. *#...summary on pm25median variable*  
`summary(chicago$pm25median)`  
`nrow(chicago)`

iii. *#...function giving mean, median and variance of each variable*  
`mvm<-function(x)`  
`{`  
`a=mean(x,na.rm=T) #mean`  
`b=var(x,na.rm=T) #variance`  
`c=median(x,na.rm=T) #median`  
`return(data.frame(mean=a,variance=b,median=c))`  
`}`  
`mvm(chicago$death)`  
`mvm(chicago$pm10median)`  
`mvm(chicago$pm25median)`  
`mvm(chicago$o3median)`  
`mvm(chicago$so2median)`  
`mvm(chicago$time)`  
`mvm(chicago$tmpd)`



```

iv. #...histogram for each variable with appropriate number of breaks
par(mfrow=c(4,2))
for(j in 1:(ncol(chicago)))
{
    hist(chicago[,j],xlab=colnames(chicago)[j],main=paste("Histogram of",colnames(chicago)[j]))
}

```

```

(d) #...plotting of each predictor variable against the response variable
par(mfrow=c(3,2))
for(j in 2:(ncol(chicago)))
{
    plot(x=chicago[,j],y=chicago$death,xlab=colnames(chicago)[j],main=paste("Scatterplot of", colnames(chicago)[j]))
}

```

i. written.

```

ii. #...outlier days for the plot of pm10median against death
out=boxplot.stats(chicago$pm10median)$out
#out=the values of any data points which lie beyond the extremes of the whiskers.
out_ind=which(chicago$pm10median%in%out)
day1=chicago$time[out_ind]
day1

```

```

#...outlier days for the plot of pm25median against death
out=boxplot.stats(chicago$pm25median)$out
out_ind=which(chicago$pm25median%in%out)
day2=chicago$time[out_ind]
day2

```

```

#...outlier days for the plot of o3median against death
out=boxplot.stats(chicago$o3median)$out
out_ind=which(chicago$o3median%in%out)
day3=chicago$time[out_ind]
day3

```

```

#...outlier days for the plot of so2median against death
out=boxplot.stats(chicago$so2median)$out
out_ind=which(chicago$so2median%in%out)
day4=chicago$time[out_ind]
day4

```

```

#...outlier days for the plot of time against death
out=boxplot.stats(chicago$time)$out
out_ind=which(chicago$time%in%out)
day5=chicago$time[out_ind]
day5

```

```

#...outlier days for the plot of tmpd against death
out=boxplot.stats(chicago$tmpd)$out
out_ind=which(chicago$tmpd%in%out)
day6=chicago$time[out_ind]
day6

```

```

#...code for finding if different plots share outlier days
sum(day1%in%day2)
sum(day1%in%day3)

```

```

sum(day1%in%day4)
sum(day1%in%day5)
sum(day1%in%day6)
sum(day2%in%day3)
sum(day2%in%day4)
sum(day2%in%day5)
sum(day2%in%day6)
sum(day3%in%day4)
sum(day3%in%day5)
sum(day3%in%day6)
sum(day4%in%day5)
sum(day4%in%day6)
sum(day5%in%day6)

```

iii. `par(mfrow=c(3,2))`

```

#...Plotting residuals against pm10median variable
death.pm10.lm<-lm(death~pm10median,data=chicago)
plot(chicago$pm10median[!is.na(chicago$pm10median)],residuals(death.pm10.lm),xlab="pm10median",
abline(h=0,col="grey")

#...Plotting residuals against pm25median variable
death.pm25.lm<-lm(death~pm25median,data=chicago)
plot(chicago$pm25median[!is.na(chicago$pm25median)],residuals(death.pm25.lm),xlab="pm25median",
abline(h=0,col="grey")

#...Plotting residuals against o3median variable
death.o3.lm<-lm(death~o3median,data=chicago)
plot(chicago$o3median[!is.na(chicago$o3median)],residuals(death.o3.lm),xlab="o3median",ylab="P
abline(h=0,col="grey")

#...Plotting residuals against so2median variable
death.so2.lm<-lm(death~so2median,data=chicago)
plot(chicago$so2median[!is.na(chicago$so2median)],residuals(death.so2.lm),xlab="so2median",yla
abline(h=0,col="grey")

#...Plotting residuals against time variable
death.time.lm<-lm(death~time,data=chicago)
plot(chicago$time[!is.na(chicago$time)],residuals(death.time.lm),xlab="time",ylab="Prediction
abline(h=0,col="grey")

#...Plotting residuals against tmpd variable
death.temp.lm<-lm(death~tmpd,data=chicago)
plot(chicago$tmpd[!is.na(chicago$tmpd)],residuals(death.temp.lm),xlab="Temperature(F)",ylab="P
abline(h=0,col="grey")

```

(e) written.

- i. written.
- ii. written.

iii. *#...scatterplot of death vs tmpd*  
`plot(death~tmpd,data=chicago,ylim=c(0,200),xlab="Daily mean temperature (F)",ylab="Mortality (`  
*#...proposed function*  
`abline(a=130,b=-0.28,col="blue")`

```
#...Q-Q Plot to verify the distribution of the residuals
qqnorm(residuals(death.temp.lm))
qqline(residuals(death.temp.lm))
```

## 2. Analysis of the econ.csv data file:

- (a) *#...loading the data file*  
`econ.data=read.table("D:\\AKG Linear Models (5th Sem)\\econ.csv",header=T,sep=",")`  
`attach(econ.data)`  
*#...dimension of the data*  
`dim(econ.data)`
- (b) *#...summary of the six numerical-valued columns*  
`summary(econ.data[,1])`
- (c) *#...univariate EDA plots for population and for per-capita GMP*  
`par(mfrow=c(1,2))`  
`hist(pop,xlab="Population",main="Histogram of Population",col="lightblue",breaks=50)`  
`hist(pcgmp,xlab="per-capita GMP",main="Histogram of per-capita GMP",col="lightblue",breaks=20)`
- (d) *#...bivariate EDA plot for per-capita GMP as a function of population*  
`plot(pop,pcgmp,xlab="Population",ylab="per-capita GMP", main="Scatterplot of per-capita GMP vs pop")`
- (e) *#...slope of the least-square regression line*  
`slope=cov(pop,pcgmp)/var(pop)`  
*#...intercept of the least-square regression line*  
`intercept=mean(pcgmp)-(slope*mean(pop))`
- (f) *#...slope returned by the function lm*  
`coefficients(lm(pcgmp~pop,data=econ.data))[2]`  
*#...intercept returned by the function lm*  
`coefficients(lm(pcgmp~pop,data=econ.data))[1]`
- (g) *#...bivariate EDA plot for per-capita GMP as a function of population*  
`plot(pop,pcgmp,xlab="Population",ylab="per-capita GMP", main="Scatterplot of per-capita GMP vs pop")`  
  
*#...adding the fitted line*  
`abline(a=intercept,b=slope)`  
  
*#...Plotting residuals against the population*  
`model<-lm(pcgmp~pop,data=econ.data)`  
`plot(pop,residuals(model),xlab="population",ylab="Prediction error(Residual)",main="Residual vs. pop")`  
`abline(h=0,col="grey")`  
  
*#...Plotting absolute residuals against the population*  
`plot(pop,abs(residuals(model)),xlab="population",ylab="Absolute prediction error(Residual)",main="Absolute Residual vs. pop")`  
`abline(h=sqrt(mean(residuals(model)^2)),lwd=2,col="grey")`

- (h) *#...finding Pittsburgh in the dataset*  
`Pitts.ind=charmatch("Pittsburgh",MSA)`  
  
*#...population of Pittsburgh*  
`pop[Pitts.ind]`  
  
*#...per-capita GMP of Pittsburgh*  
`pcgmp[Pitts.ind]`  
  
*#...per-capita GMP predicted by the model*  
`pred.pcgmp=(coefficients(model)[1])+(coefficients(model)[2])*(pop[Pitts.ind])`  
  
*#...residual for Pittsburgh*  
`Pitts.rsd=pcgmp[Pitts.ind]-pred.pcgmp`
- (i) *#...mean squared error of the regression*  
`mse=mean(residuals(model)^2)`  
`mse`
- (j) *#...ratio of Residual square for Pittsburgh to the Mean Residual Square or MSE*  
`ratio=(Pitts.rsd^2)/mse`  
`ratio`
- (k) written.
- (l) *#...predicted per-capita GMP for a city with more people than Pittsburgh*  
`pred.pcgmp=(coefficients(model)[1])+(coefficients(model)[2])*((10^5)+pop[Pitts.ind])`  
`pred.pcgmp`