Outlier Detection(regarding Econometrics) in different dataset in R

Pratyusha Bala
STAT019
SEM-6, UG-3
Presidency University

3/5/2022

Dataset 1:

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 such as: pm10median, pm25median, o3median, so2median and tmpd.

Explatory Data Analysis:

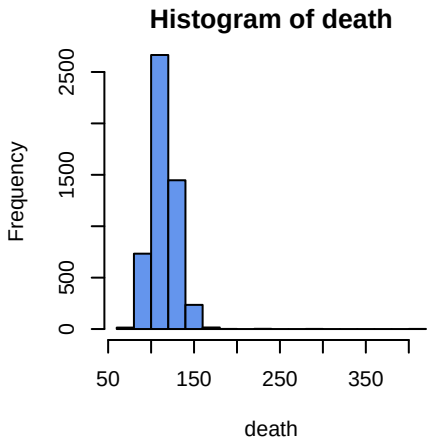
Here is the summary for each variables in the form of a table.

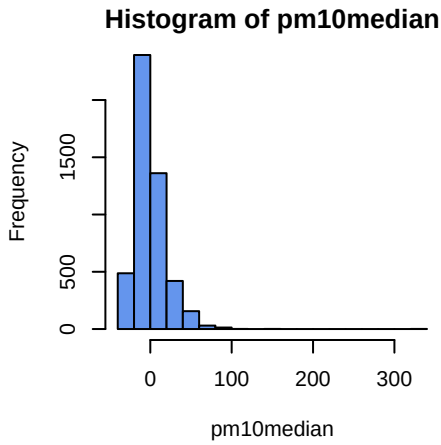
The temperature is in fahrenheit, since its range is (-16, 92). 92 degree celcius is not possible for weather on earth .Also, since there are 4387 NA values in the data for median density of smaller pollutant particles(pm25median) out of 5114 values, we shall ignore it.Let us see the histograms now.

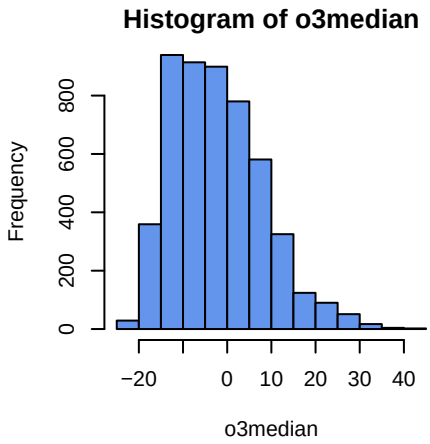
```
## Warning:  package 'gamair' was built under R version 4.1.2
```

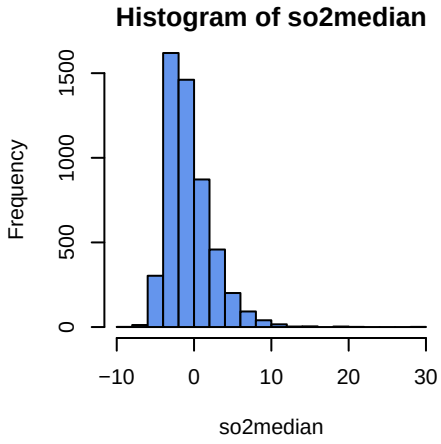
Table 1: Summary for each variables

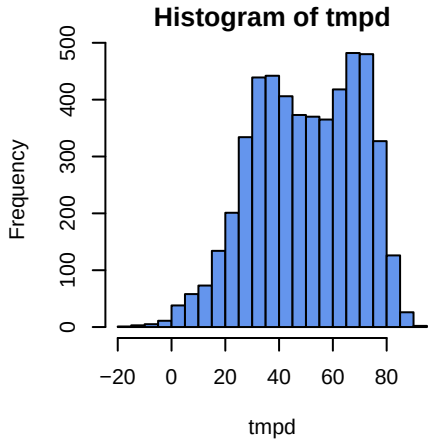
	min	1st Qu	median	mean	3rd Qu	max	variance
death	69	105	114	115.4	124	411	234
pm10	-37.38	-13.1	-3.54	-0.15	8.3	320.72	370.72
pm25	-16.42	-6.59	-1.33	0.243	5.34	38.15	75.2
o3	-24.78	-10.23	-3.33	-2.18	4.47	43.69	104.1
so2	:-8.21	-2.69	-1.22	-0.64	0.83	28.9	8.56
time	-2556	-1278	0	0	1278	2556	111360
tmpd	-16	35	51	50	67	92	378.7



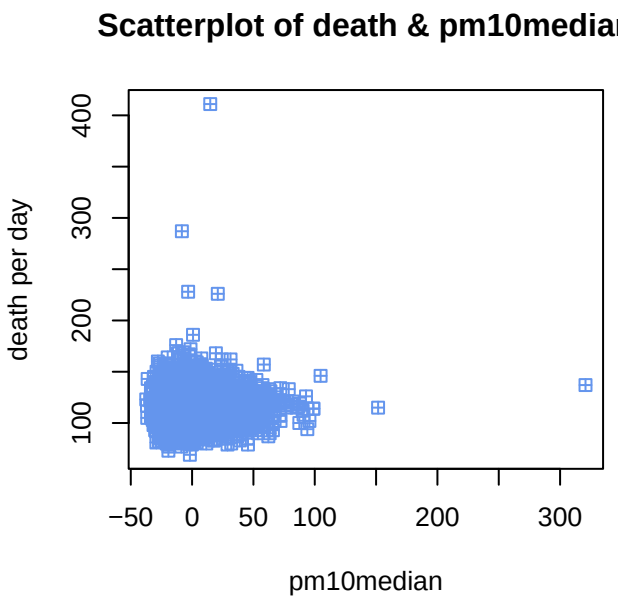


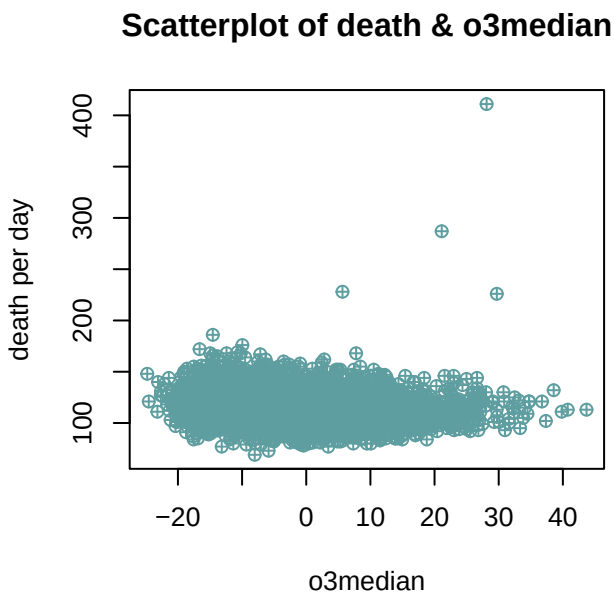


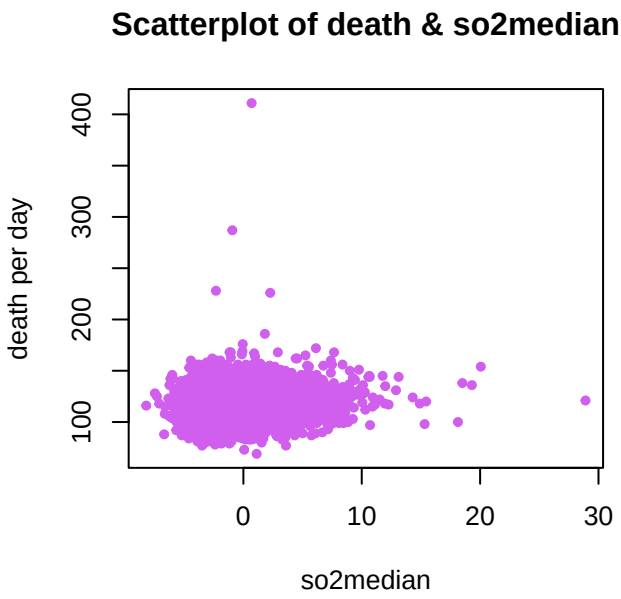


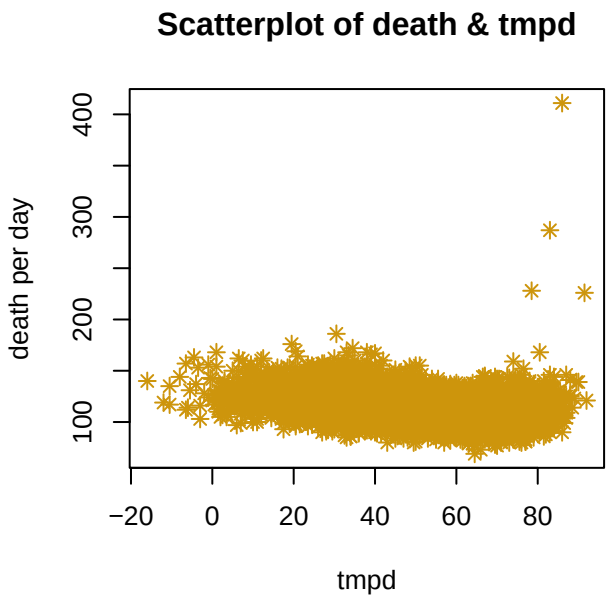


Thus, death per day,the median density of large pollutant particles(pm10median) and the median concentration of SO2 in the air (so2median) are positively skewed. tmpd is negatively skewed with wide spread whereas ozone concentration shows slightly positive skewness. Here only histograms are not able to show any outliers.Let's see what the scatterplots tell us.









There are outliers in all four scatterplots. Death rate seems to be linearly related with o3median and tmpd,but because of the influence from outliers the correlation coefficient turns out slightly negative in both cases.Whereas, we have the other two variables, pm10median and so2median. with increases in their values the death rate remains generally constant, except some outliers.

Analysis

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

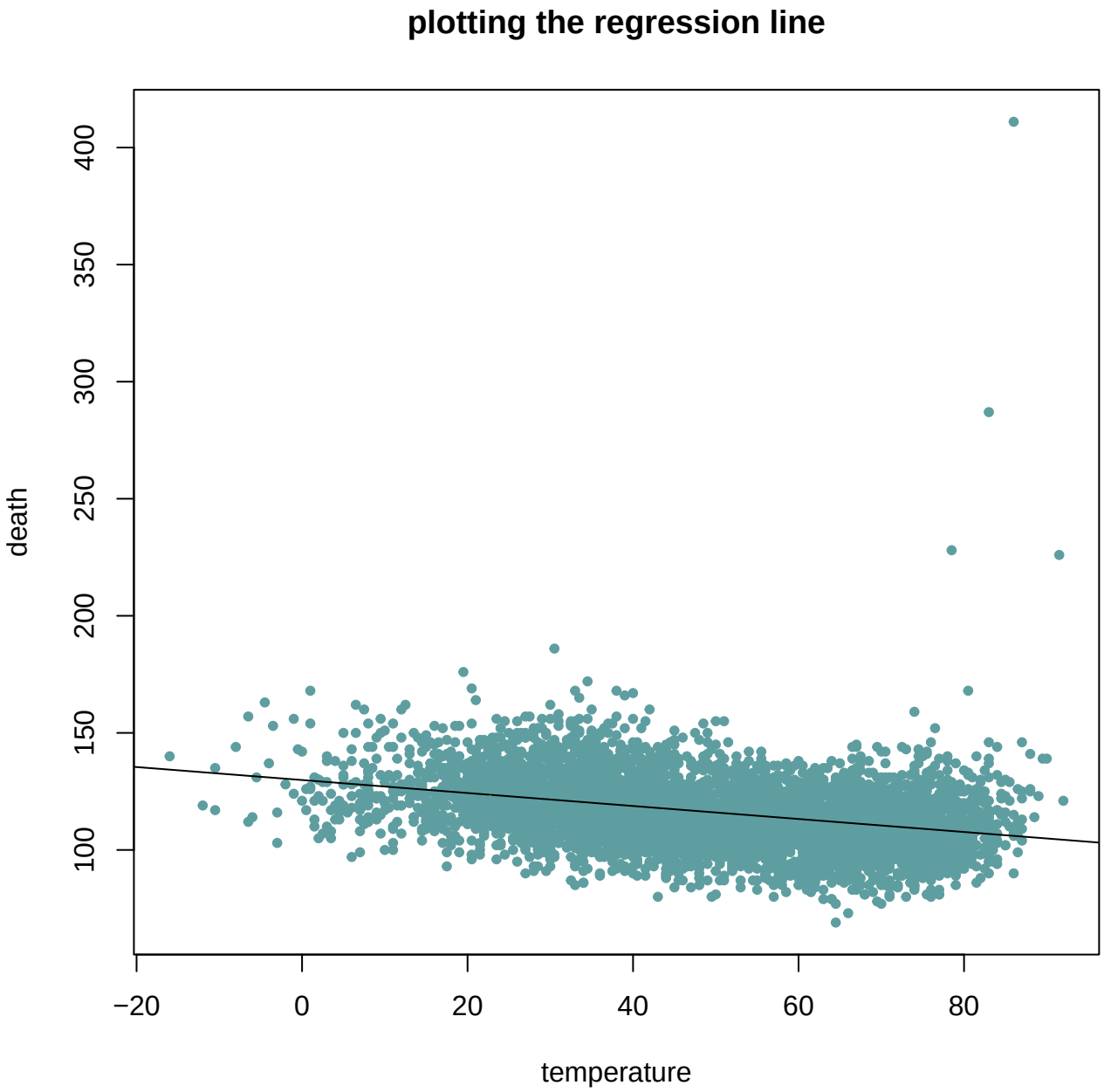
Then the regression model is:

$$y_i = 130 - 0.28x_i + e_i$$

where

$$e_i \sim N(0, 14.22) \forall i$$

At 0 degree fahrenheit, the death rate is 130 unit and for unit change in temperature the death rate decreases by 0.28 units.



The proposed function fits the data very well, and hence it's safe to assume the normal error regression model. Also the predicted change in the number of deaths from a 2°C degree warming over the course of a whole year is 126 units.

Conclusion:

It's clear that death and temperature are related , but they do not have a causal relationship. More temperature does not imply more death rate, but if we see that the temperature is very high, we can atleast expect to see more deaths.

Dataset 2:

The data file econ.csv contains information about the economies of the 366 "metropolitan statistical areas" (≈ cities) of the US in 2006. It lists, for each city, the population, per capita gross metropolitan product or pcgmp, and the share of economic output coming from four selected industries

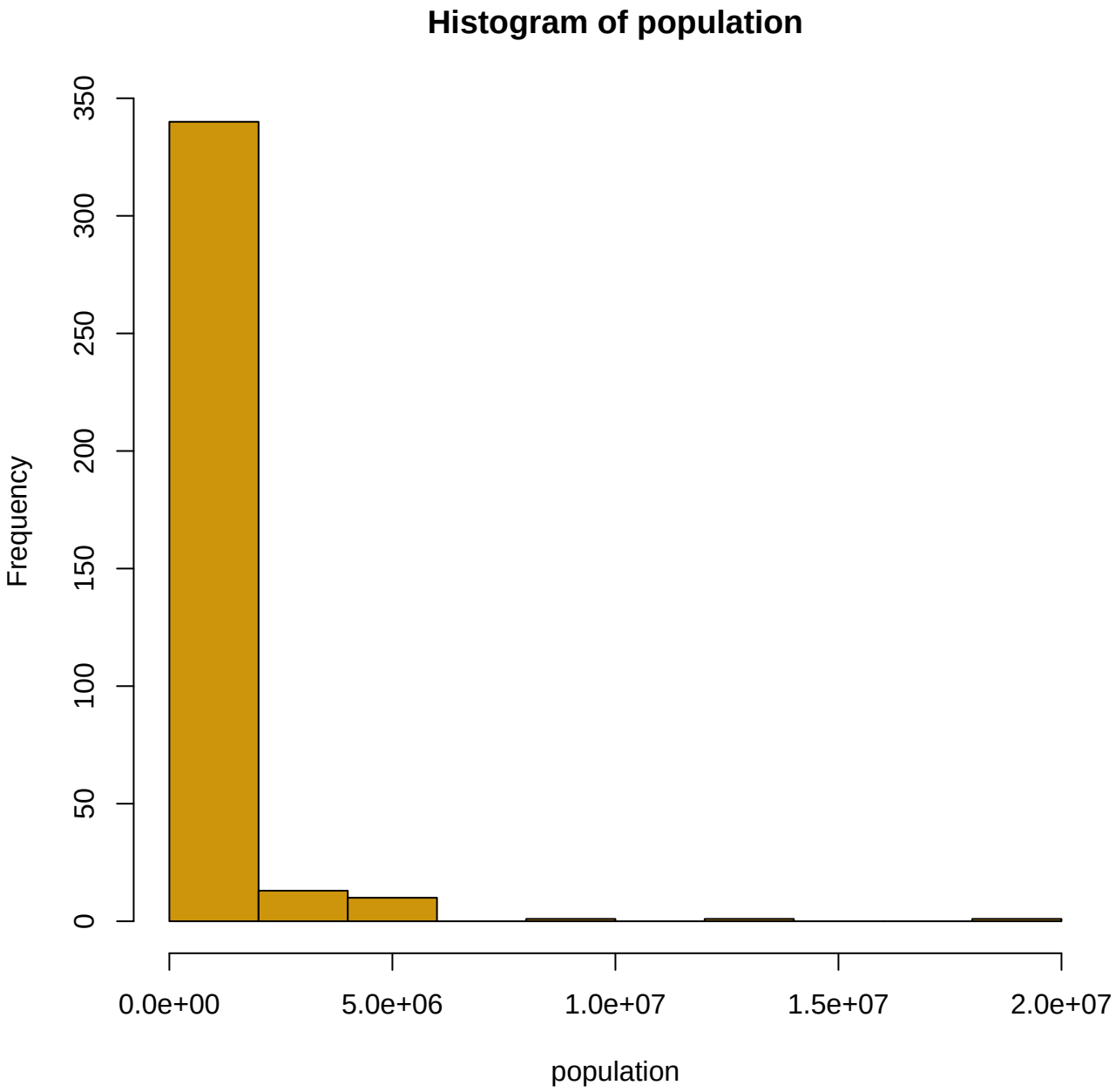
Explatory Data Analysis:

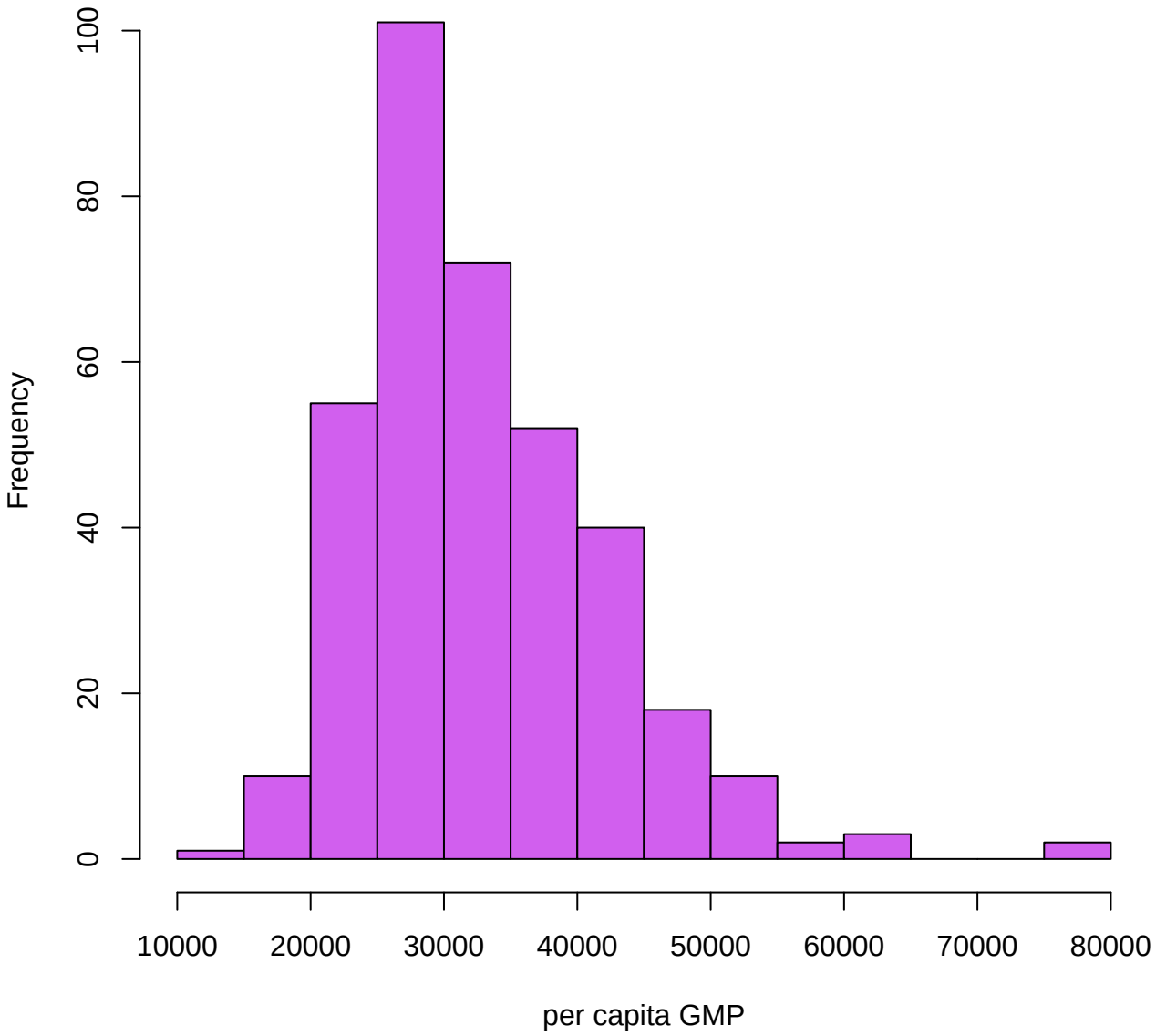
Here is the summary for each variables in the form of a table.

Table 2: Summary Statistics

	1st Qu	median	mean	3rd Qu
pcgmp	26533	31615	32923	38213
pop	135625	231500	680898	530875
finance	0.10403	0.1414	0.1508	0.1812
prof. tech	0.029	0.042	0.049	0.059
ict	0.012	0.022	0.039	0.04
management	0.003	0.0065	0.009	0.011

Now let us plot some histograms.





As it is clear from the histograms, population is positively skewed with outliers having very large values. Also, the distribution is very much leptokurtic, there are more cities having moderate population and only a few having large population.

Per capita GMP is relatively lesser skewed, but positive in nature. the distribution almost resembles normal distribution, there are a few cities with high pcgmp.

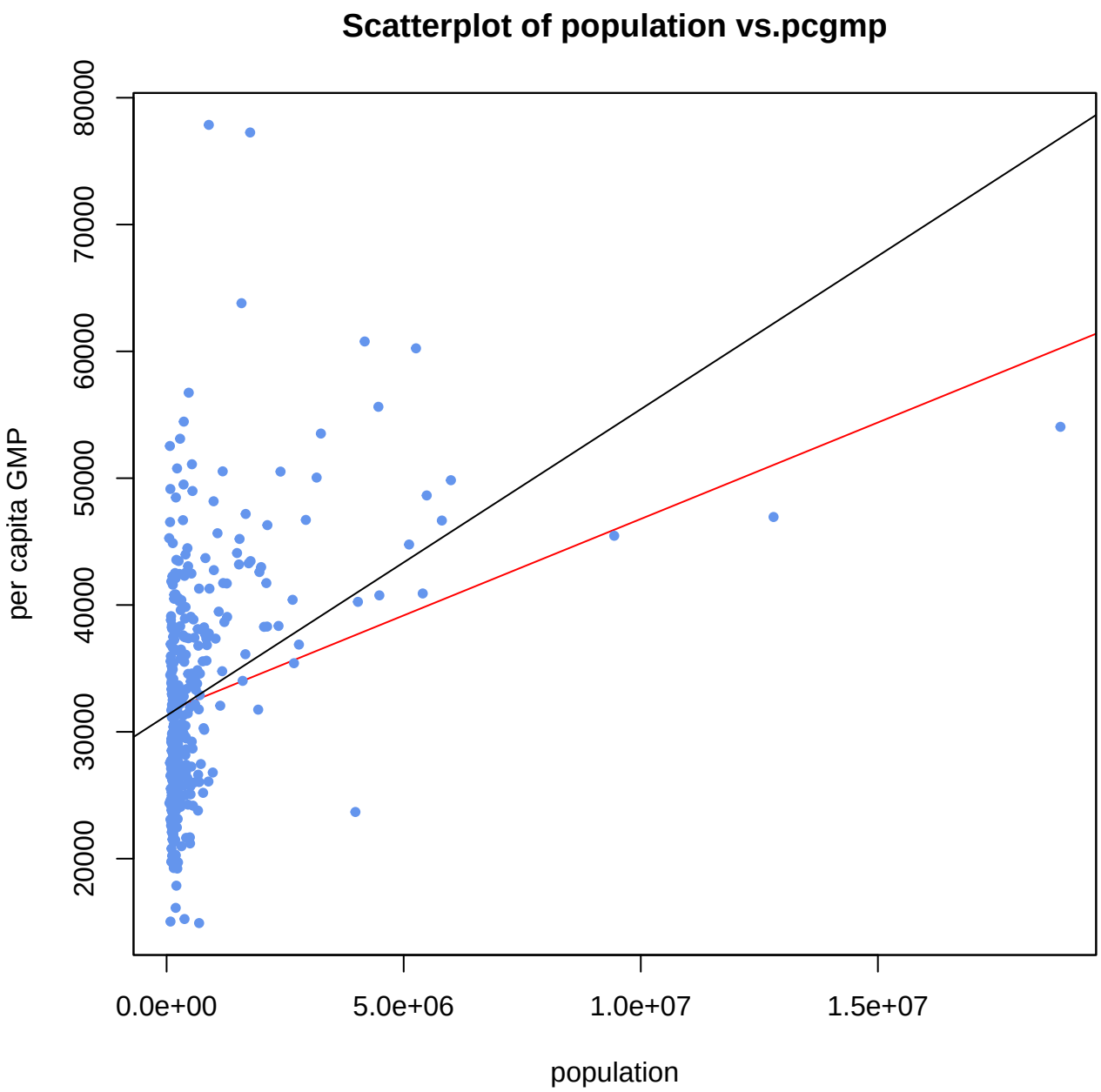
Analysis:

Let us use simple linear regression of pcgmp on population and draw the regression line on the same plot.

The Least Square estimates of the coefficients are,

$\hat{\beta} = \frac{cov(pcgmp,pop)}{var(pop)} = 0.002422821$

and, $\hat{\alpha} = pc\bar{g}mp - \hat{\beta} * p\bar{o}p = 31273.07$.



The red line is the regression line and the black one is the line returned by the function lm.

Conclusion:

The regression line works only for population lesser than 2000. The reason for this is the influence of the outliers,which alters the slope of the line. When the population is close to 0, the model seems to fit best, but as the population grows , it becomes more and more useless. Alternatively, the line calculated from least square estimates fits well to the outliers here. The MSE of the regression is 71085591 which is very large. The estimated slope is 0.002422, that means for every unit change in population, there is 0.002422 units of change in pcgmp.

In particular, Pittsburgh, has population=2361000, pcgmp=38350. The residual according to our model would be 1356.654. For a city with 10^5 more people than Pittsburgh, the model predicts that it will have 37235.63 pcgmp.

R codes:

Dataset 1

```
library(gamair)
data(chicago)
head(chicago)#...(a)

summary(chicago)
var(na.omit(chicago[,1]))*5113/5114
var(na.omit(chicago[,2]))*4862/4863
var(na.omit(chicago[,3]))*726/727
var(na.omit(chicago[,4]))*5113/5114
var(na.omit(chicago[,5]))*5086/5087
var(na.omit(chicago[,6]))*5113/5114
var(na.omit(chicago[,7]))*5113/5114#...(b)

c=chicago[,c(-3,-6)]
par(mfrow=c(3,3),mar=c(4,4,2,0.5))
for (j in 1:(ncol(c)))
{
hist(c[,j], xlab=colnames(c)[j],main=paste("Histogram of", colnames(c)[j]),col="cornflowerblue", breaks=20)
} #...(c)

par(mfrow=c(2,2),mar=c(5.1,4.1,4.1,2.1))
for(i in 2:5)
{
p=c("cornflowerblue","cadetblue","mediumorchid2","darkgoldenrod3")
h=c(12,10,20,8)
plot(c[,i],c[,1],col=p[i-1],pch=h[i-1],ylab="death per day",xlab=colnames(c)[i],main=paste("Scattorplot of deat
}
cor(c)#...(d)

par(mfrow=c(1,1),mar=c(5.1,4.1,4.1,2.1))
sim.gnslrm <- function(x, intercept, slope, sigma.sq)
{
n <- length(x)
y <- intercept + slope * x + rnorm(n, mean = 0, sd = sqrt(sigma.sq))
return(data.frame(x = x, y = y))
}
sim=sim.gnslrm(c[,5],130,-0.28,14.22)
model=lm(y~x,data=sim)
plot(c$tmpd,c$death,pch=20,col="cadetblue")
abline(model)
sim.gnslrm(x=35.6,130,-0.28,14.22)#...(e)
```

Dataset 2:

```
a=read.table("C:\\Users\\pratyusha\\Downloads\\rdatasets\\econ.csv", header=T, sep=",")
head(a)
dim(a)#...(a)

summary(a[,-1])#...(b)

hist(a$pop,xlab="population",main=paste("Histogram of population"))
hist(a$pcgmp,xlab="per capita GMP",main=paste("Histogram of pcgmp"))#...(c)

plot(a$pop,a$pcgmp,xlab="population",ylab="per capita GMP")
model=lm(a$pcgmp~a$pop,data=a)
model
```

```r
abline(model)#...(d)

beta=cov(a$pcgmp,a$pop)*366/(var(a$pop)*365);beta
alpha=mean(a$pcgmp)-beta*mean(a$pop);alpha#...(e)

p=seq(min(a$pop),max(a$pop),length=10)
q=alpha+beta*p
plot(p,q,xlab="",ylab="",xaxt="n",yaxt="n",xlim=c(0,1.5e+07),ylim=c(10000,90000),col="red",type="l")
par(new=T)
plot(a$pop,a$pcgmp,xlab="population",ylab="per capita GMP",main="Scatterplot of population vs.pcgmp",pch=20,col
model=lm(a$pcgmp~a$pop,data=a)
abline(model)#...(e)

a[a$MSA=="Pittsburgh, PA",]
predicted.pcgmp=alpha+beta*a[a$MSA=="Pittsburgh, PA","pop"]
residual.pcgmp=a[a$MSA=="Pittsburgh, PA","pcgmp"]-predicted.pcgmp
residual.pcgmp#...(f)

sum((model$residuals)^2)/df.residual(model)#...(g)

alpha+beta*(a[a$MSA=="Pittsburgh, PA","pop"]+10^5)#...(h)
```