HW1 - Chicago and Neighbors

Emulie Chhor

30/12/2022

Question 1 - Load the dataset and run summary on it

```
data("chicago")
help("chicago")
kable(summary(chicago))
```

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

Q1a)

By looking at `help(chicago)`, we see that the temperature is given in Fahrenheit

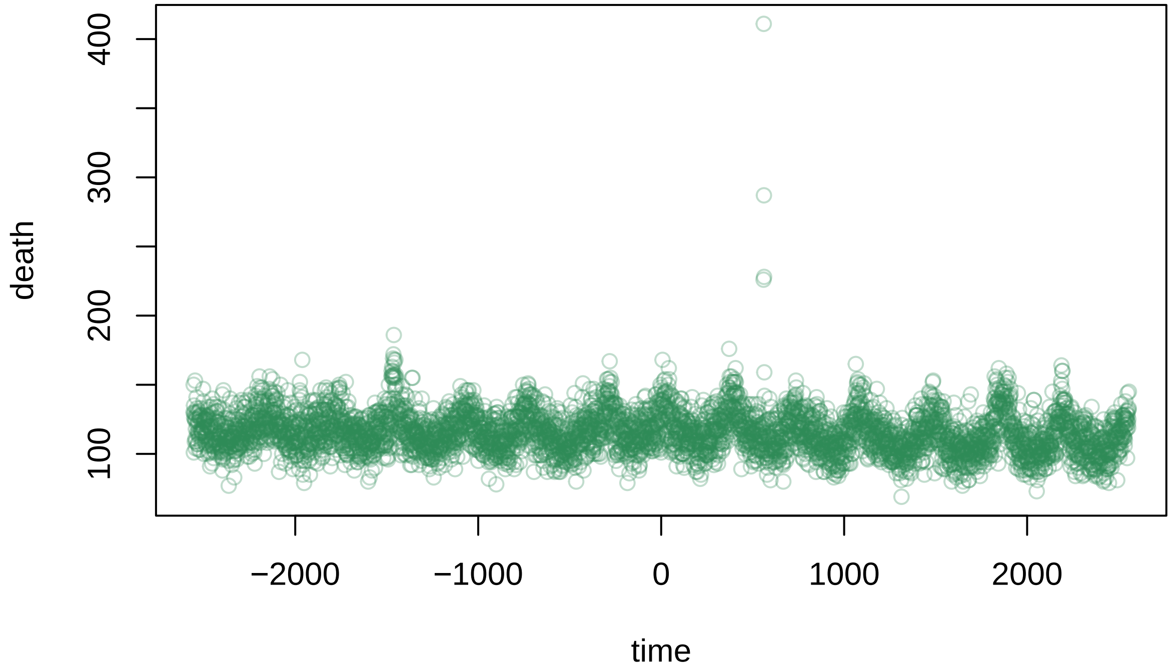
Q1b)

It means that the pollution is less than the median most of the days.

Question 2 - Death Over Time

Q2a)

```
with(chicago, plot(time, death, col=alpha('seagreen', 0.3)))
```



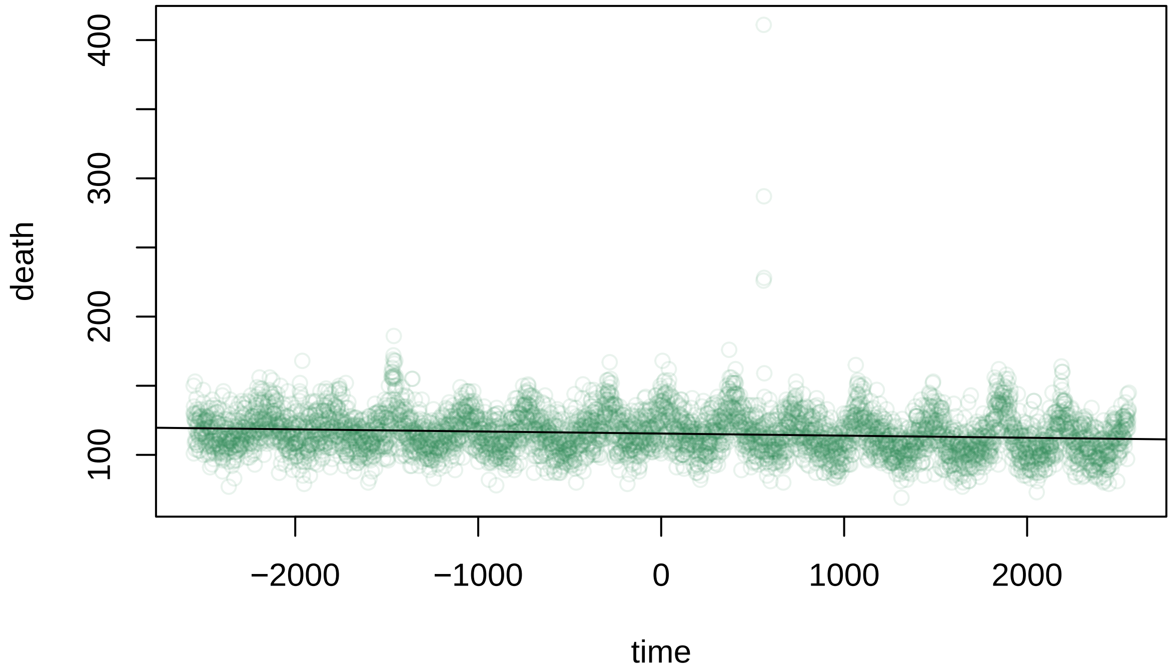
```
# adding calendar date to chicago
day.zero <- as.Date("1993-12-31")
chicago$date <- day.zero + chicago$time

# chicago %>%
#   select("death") %>%
#   as.ts() %>%
#   feasts::autoplot()
```

Q2b)

```
model1 <- lm(death~time, data = chicago)
kable(summary(model1)$coefficients)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	115.4188502	0.2116365	545.36371	0
time	-0.0015207	0.0001434	-10.60775	0

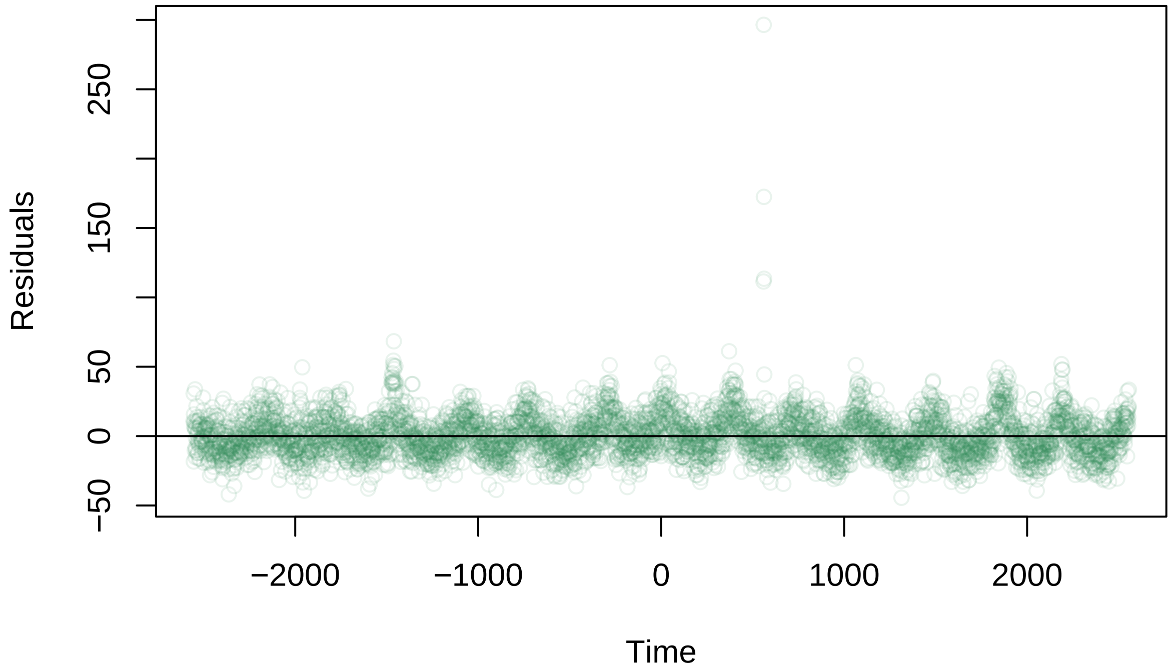
```
with(chicago, plot(time, death, col=alpha('seagreen', 0.1)))
abline(model1)
```



The slope coefficient estimate is -0.0015207. Since its p-value is under $\alpha = 0.05$, we say that it is significantly different from 0.

Q2c)

```
plot(chicago$time, residuals(model1), xlab = "Time", ylab = "Residuals",
     col=alpha('seagreen', 0.1))
abline(h=mean(residuals(model1)))
```



It seems that: - mean residual is 0 - homoscedacity: variance is constant - residuals are NOT independant: there seem to be a periodicity

Q2d)

Since the slope coefficient is statistically significant, we say that everyday that goes by since 1993-12-31, we expect the death rate to decrease by around -0.0015207.

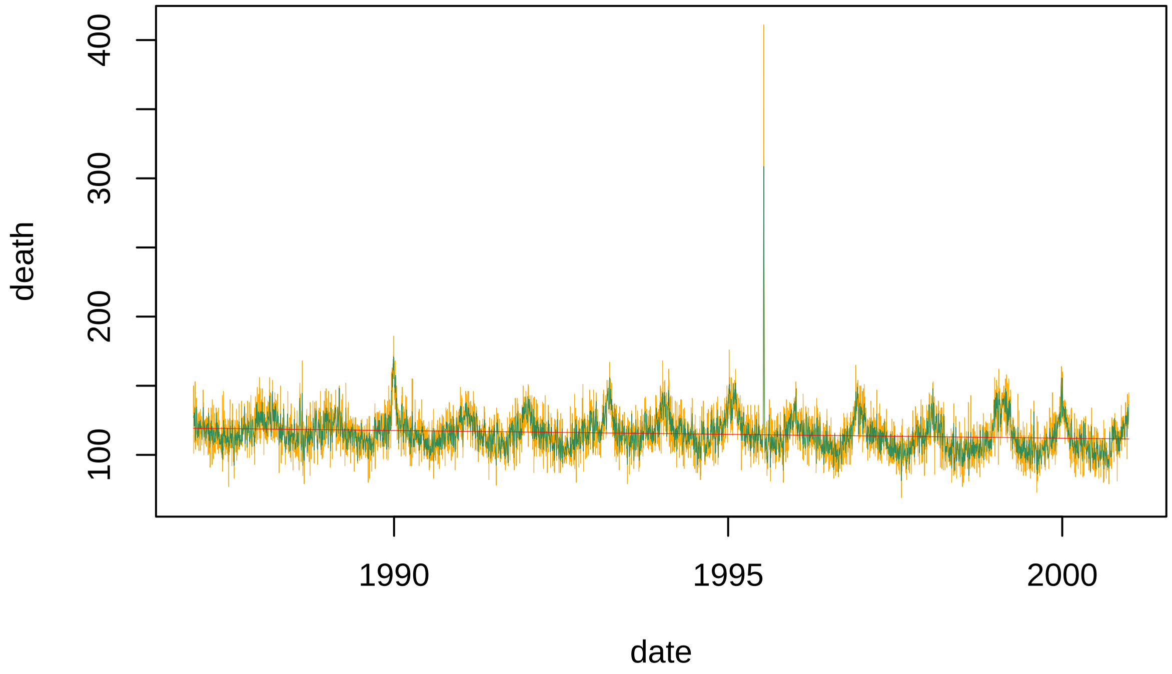
Q2e)

There is reason to doubt the validity of the significance test since the data doesn't verify all the linear regression assumptions. Therefore, it is unreasonable to think that the data is linear as it is.

Question 3 - Neighbors in Time

Q3a)

```
# predict deaths using KNN
knn_model1 <- with(chicago, knn.reg(train = time,
                       test = as.matrix(time, ncol=1),
                       y = death, k = 3))

# plot KNN predictions
plot(death~date, data = chicago, type = 'l', lwd = 0.3, col='orange')
lines(chicago$date, knn_model1$pred, lwd = 0.3, col='seagreen')
lines(chicago$date, predict(model1), lwd = 0.3, col='red')
```



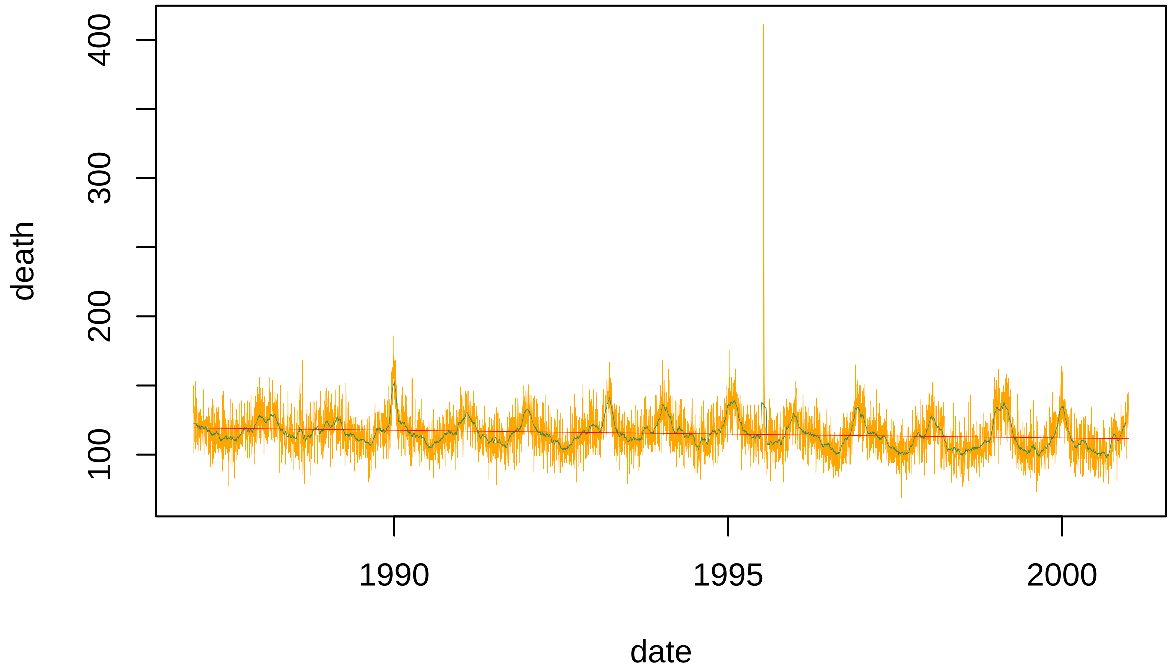
The estimated function has a similar shape than the observed response variable whereas a linear regression doesn't fit as well.

Q3b)

The predicted values are calculated by computing the mean of the deaths of the 3 closest dates.

Q3c)

```
# predict deaths using KNN
knn_model2 <- with(chicago, knn.reg(train = time,
                      test = as.matrix(time, ncol=1),
                      y = death, k = 30))

# plot KNN predictions
plot(death~date, data = chicago, type = 'l', lwd = 0.3, col='orange')
lines(chicago$date, knn_model2$pred, lwd = 0.3, col='seagreen')
lines(chicago$date, predict(model1), lwd = 0.3, col='red')
```

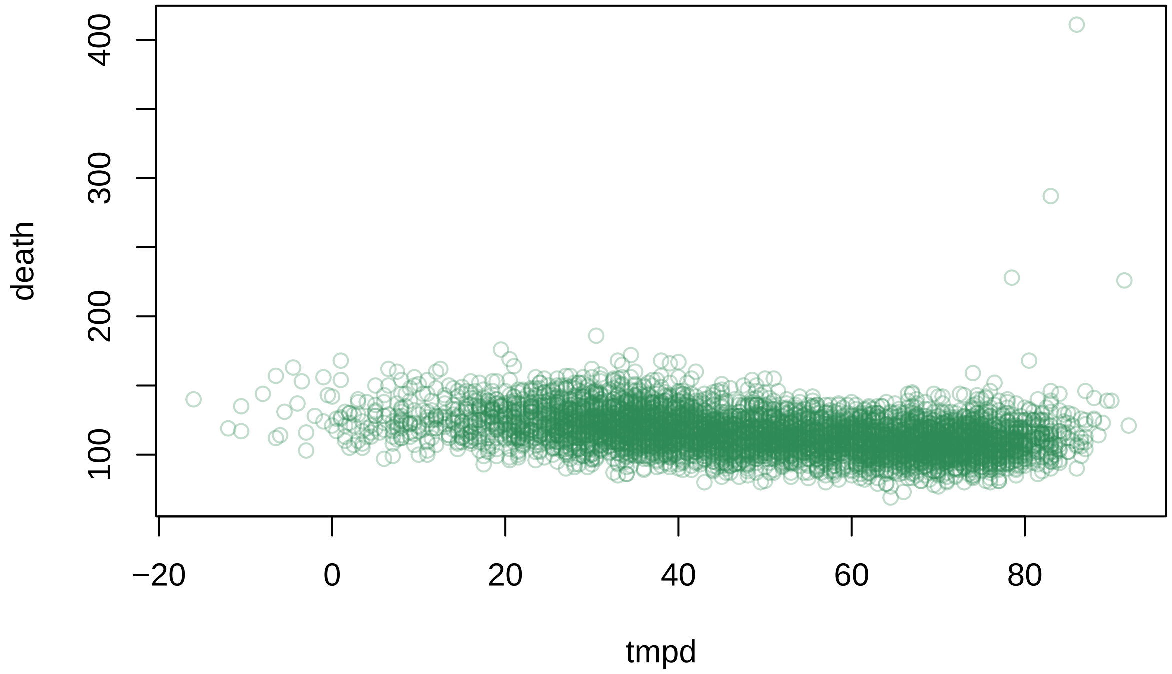


The new estimate using k=30 is smoother than with k=3 because the outliers are less important in the prediction. Therefore, the variance of the model with k=30 is less than with k=3

Question 4 -

Q4a)

```
with(chicago, plot(tmpd, death, col=alpha('seagreen', 0.3)))
```



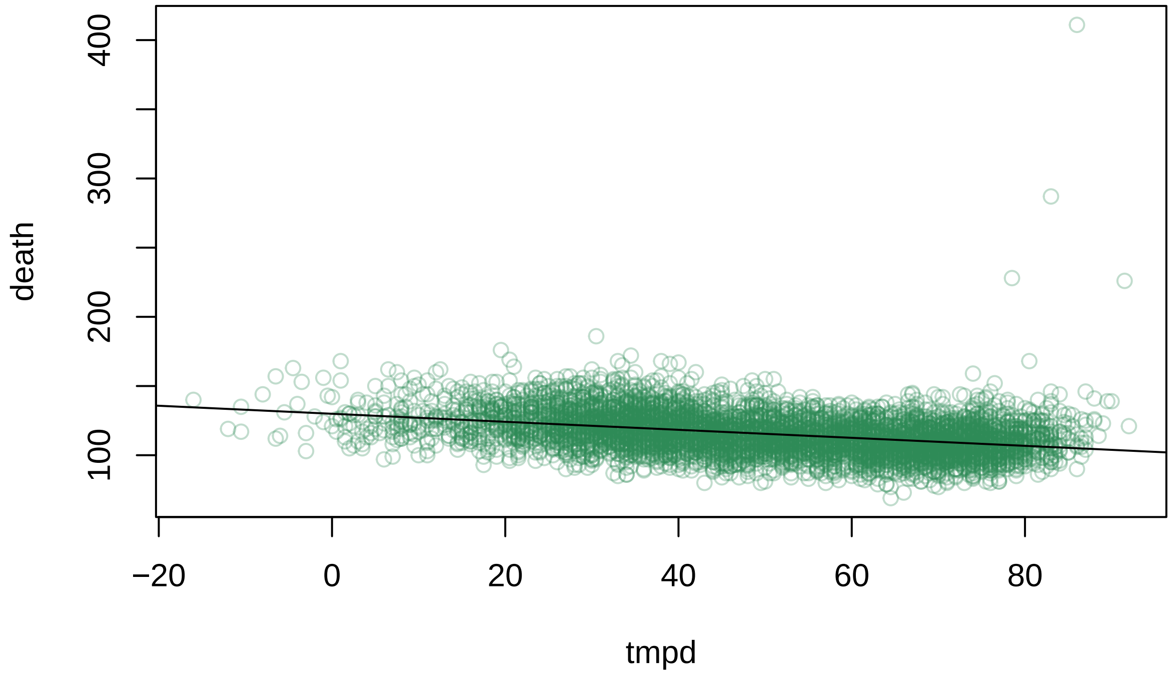
It seems that the data is linear

Q4b)

```
lm_model_temp <- lm(death~tmpd, data = chicago)
summary(lm_model_temp)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 129.9570512 0.55022802 236.18763  0.00000e+00
## tmpd         -0.2896443 0.01022089 -28.33845 3.23449e-164
```

```
with(chicago, plot(tmpd, death, col=alpha('seagreen', 0.3)))
abline(lm_model_temp)
```



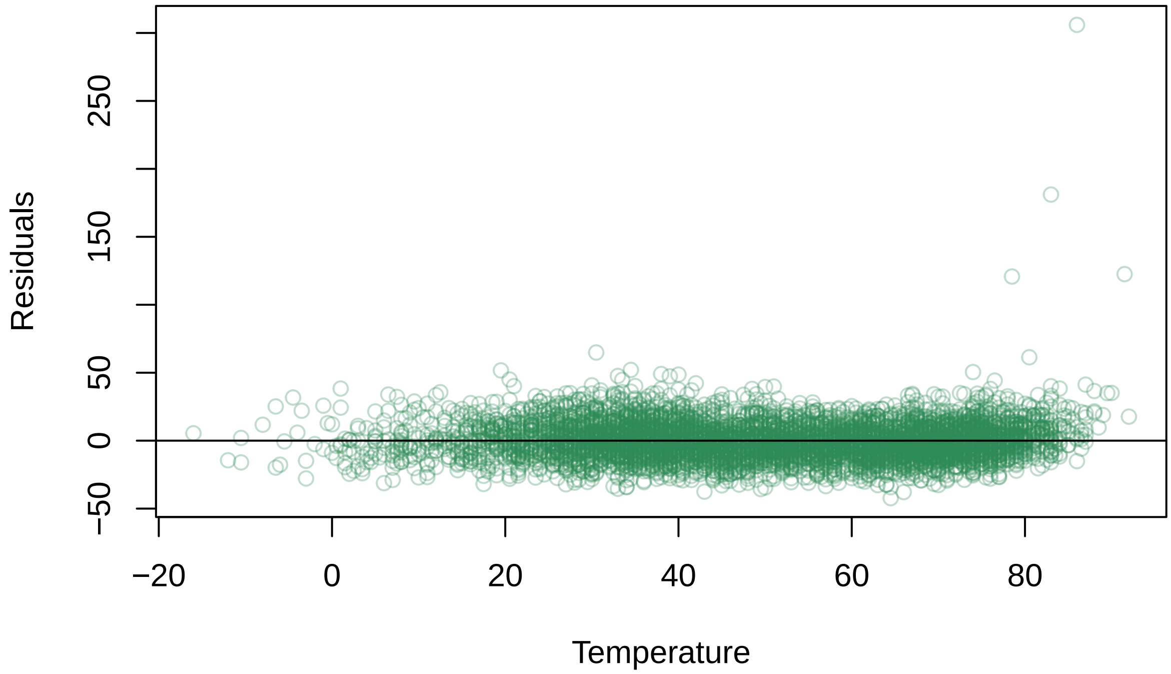
The slope coefficient is -0.2896443 and is significant because its p-value is smaller than $\alpha = 0.05$

Q4c)

For every increase of 1 degrees Fahrenheit, we expect the number of death decrease by -0.2896443 on average.

Q4d)

```
plot(chicago$tmpd, residuals(lm_model_temp), xlab = "Temperature",
     ylab = "Residuals", col=alpha('seagreen', 0.3))
abline(h=mean(residuals(lm_model_temp)))
```

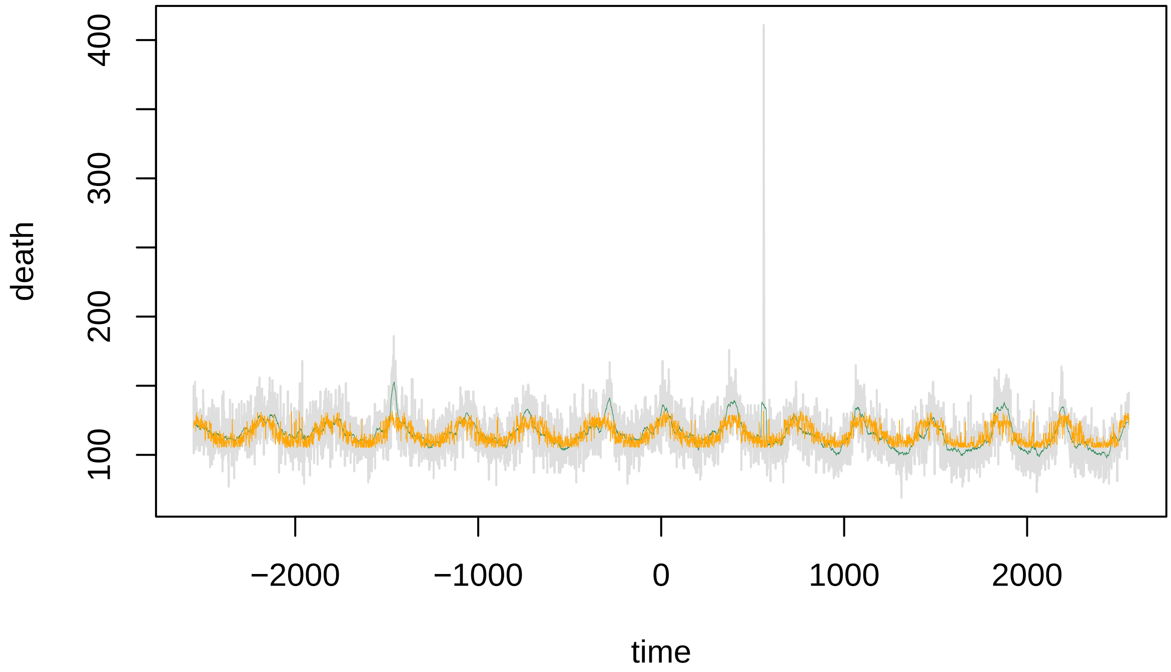


It seems that all the linear regression assumptions are met: - means residual is 0 - independance of residuals: there is no correlation between errors - homoscedacity: in general, the variance is constant except for a few outliers toward the end.

Question 5 -

Q5a)

```
# prediction using KNN on temperature
knn_model_temp <- with(chicago, knn.reg(train = tmpd,
      test = as.matrix(tmpd, ncol = 1), y = death, k=30))

# comparing knn model on time and temperature for k = 30
with(chicago, plot(time, death, type = 'l', col = alpha('grey', 0.5)))
lines(chicago$time, knn_model2$pred, col='seagreen', lwd = 0.3)
lines(chicago$time, knn_model_temp$pred, col='orange', lwd = 0.3)
```



Q5b)

The estimated values we get using a linear regression looks like a line where as with the KNN, the estimated values can encapsulate non-linearity. The estimated value looks like a time series with some seasonality.

Question 6 -

Q6a)

```
# make the temperature 4 degrees Celsius hotter
temp.celsius <- (chicago$tmpd - 32) * 5/9 + 4
chicago$warmer <- temp.celsius * 9/5 + 32
```

Q6b)

```
lm.temp.diffs <- predict(lm_model_temp,  newdata = list(tmpd=chicago$warmer)) - predict(lm_model_temp, 
kable(mean(lm.temp.diffs))
```

x
-2.085439

Q6c)

```
knn.temp.diffs <- with(chicago, knn.reg(train = tmpd,
                                        test = as.matrix(warmer, ncol = 1),
                                        y = death, k = 30)$pred -
```

```
  knn.reg(train = tmpd, test = as.matrix(tmpd, ncol = 1),
          y = death, k = 30)$pred)

kable(mean(knn.temp.diffs))
```

x
-0.4822383