Report for HW2

12032924 李熹成

Problem #1

```
# 1.Significant earthquakes since 2150 B.C.
##1.1
library(tidyr)
library(dplyr)
library(ggplot2)
SED<-read.csv(file = 'signif.txt',header = T,sep = '\t')
class(SED)
Seq_Eqs<-as_tibble(SED)
##1.2
Seq_Eqs %>%
  group_by(COUNTRY) %>%
  select(DEATHS,YEAR,COUNTRY) %>%
  summarise(total_num_dth=sum(DEATHS)) %>%
  arrange(desc(total_num_dth))->rank_death
rank_death[1:10,]
```

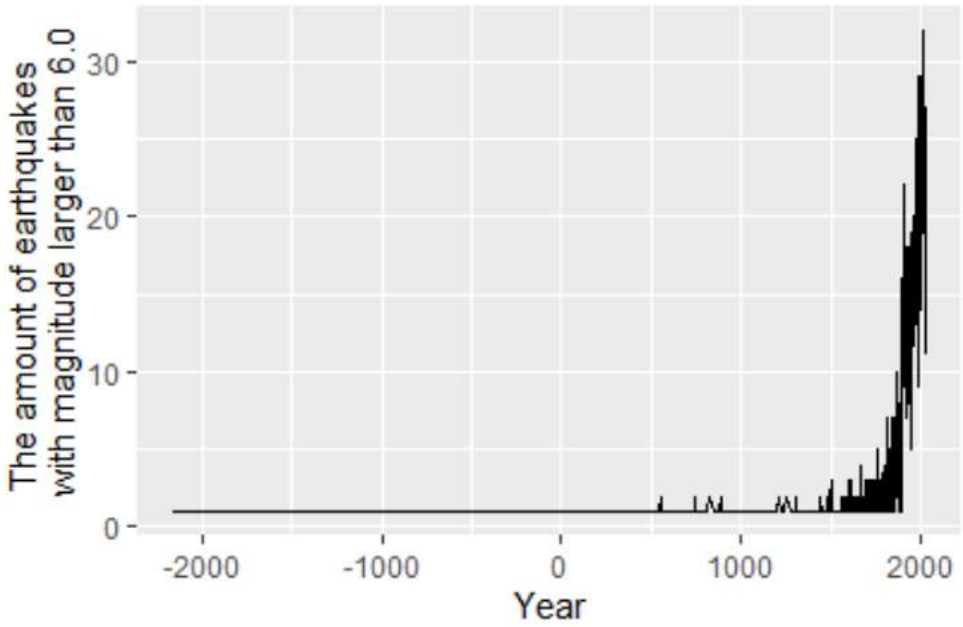
```
##1.3
Seq_Eqs %>%
  filter(EQ_PRIMARY>6) %>%
  group_by(YEAR) %>%
  summarise(ersq_amount=n()) %>%
  ggplot(aes(x=YEAR,y=ersq_amount))+
  geom_line()+
  scale_x_continuous(name = 'Year')+
  scale_y_continuous(name = 'The amount of earthquakes
  with magnitude larger than 6.0')
###From the plot above, the earthquakes recorded is more frequently.
###However, it could also result from the detective techniques advanced.
```

```
##1.4
Acountry<-readline(prompt="Please enter a country you want to observe:")
CountEq_LargestEq<-function(Acountry){
  Seq_Eqs %>%
    filter(COUNTRY==Acountry& EQ_PRIMARY!='NA') %>%
    mutate(ThatDate=paste(YEAR,MONTH,DAY,sep = '-')) %>%
    select(ThatDate,EQ_PRIMARY) %>%
    summarise(ersq_amount2=n(),max_level_date=
                ThatDate[which(EQ_PRIMARY==max(EQ_PRIMARY))])->C_D
  C_D_NH<-unname(C_D)#remove dimname
  return(C_D_NH)
}
```

```
###Get rid of countries with earthquake magnitude equal to 'NA'
Seq_Eqs %>%
  filter(EQ_PRIMARY!='NA')->Seq_Eqs_noNA

i=1
NewMat<-matrix(ncol = 3,nrow = length(unique(Seq_Eqs_noNA$COUNTRY)))
for(CountryName in unique(Seq_Eqs_noNA$COUNTRY)){
  NewMat[i,]<-c(as.character(CountryName),
                as.numeric(CountEq_LargestEq(CountryName)[1,1]),
                as.character(CountEq_LargestEq(CountryName)[1,2]))
  i=i+1
}
#Sort in descending order by earthquake numbers.
NewMat_Order<-NewMat[order(as.numeric(NewMat[,2]),decreasing=T),]
NewMat_Order
```

1.2 I used desc() to rank top 10 countries along with the total number of death in 1.2.

1.3 Plot the graph and added labels to the axis. The result is as below:



The earthquake magnitude more than 6 seems to become more frequently. However, it probably because the monitoring technique is getting more advanced over hundreds of years.

1.4 The results are as below:

```
> NewMat_Order
      [,1]                                [,2]  [,3]
 [1,] "CHINA"                             "575" "1668-7-25"
 [2,] "JAPAN"                             "343" "2011-3-11"
 [3,] "INDONESIA"                         "314" "2004-12-26"
 [4,] "IRAN"                              "249" "856-12-22"
 [5,] "USA"                               "215" "1964-3-28"
 [6,] "TURKEY"                            "206" "1912-8-9"
 [7,] "GREECE"                            "152" "365-7-21"
 [8,] "PERU"                              "146" "1716-2-6"
 [9,] "CHILE"                             "145" "1960-5-22"
[10,] "RUSSIA"                            "139" "1952-11-4"
[11,] "PHILIPPINES"                       "132" "1897-9-21"
[12,] "MEXICO"                            "119" "1899-1-24"
[13,] "ITALY"                             "96"  "1915-1-13"
[14,] "TAIWAN"                            "93"  "1920-6-5"
[15,] "PAPUA NEW GUINEA"                  "89"  "1919-5-6"
[16,] "INDIA"                             "81"  "1950-8-15"
[17,] "NEW ZEALAND"                       "62"  "1826-NA-NA"
```

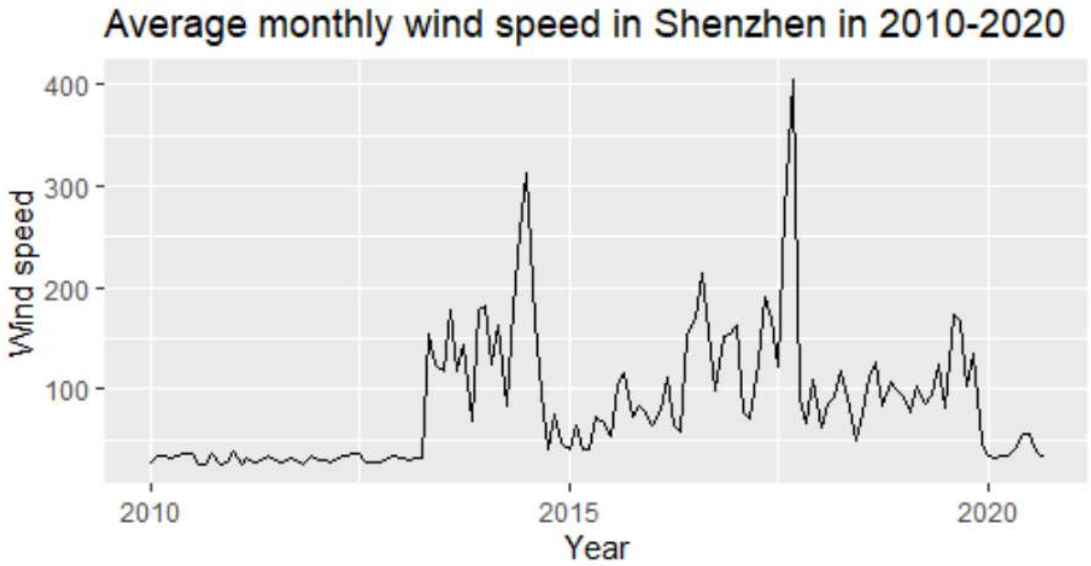
Create a function named *CountEq_LargestEq()* function to pick the times and date-of-maxEq according to country name. Then using a *for* loop to pick all countries names from the table and run *CountEq_LargestEq().* Noted that before the operations above, I remove the country name with earthquake magnitude of NA. Finally, I combined the element in each output dataframe and convert them into a matrix and output the result in decreasing order.

Problem #2

```
#2. Wind speed in Shenzhen during the past 10 years

ShenzhenData<-read.csv(file = '2281305.csv',header = T)
class(ShenzhenData)
SZD_tbl<-as_tibble(ShenzhenData)

SZD_tbl %>%
  mutate(WD_angle=as.numeric(substr(WND,1,3)),
         WD_DQC=substr(WND,5,5),
         WD_TC=substr(WND,7,7),
         WD_Speed=as.numeric(substr(WND,9,12)),
         WD_SQC=substr(WND,14,14),
         Months=substr(DATE,1,7)) %>%
  mutate(WD_angle_New=ifelse(WD_angle==999,'NA',WD_angle),
         WD_DQC_New=ifelse(WD_DQC=='3'|WD_DQC=='7','NA',WD_DQC),
         WD_TC_New=ifelse(WD_TC=='9',NA,WD_TC),
         WD_Speed_New=ifelse(WD_TC==9999,'NA',WD_Speed),
         WD_SQC_New=ifelse(WD_SQC=='3'|WD_SQC=='7','NA',WD_SQC)) %>%
  # Filter(WD_angle_New!='NA',
  #        WD_DQC_New!='NA',
  #        WD_TC_New!='NA',
  #        WD_Speed_New!='NA',
  #        WD_SQC_New!='NA') %>%
  select(WD_angle_New, WD_DQC_New,
         WD_TC_New,WD_Speed_New,WD_SQC_New,Months) %>%
  group_by(Months) %>%
  summarise(MonAvgWS=mean(WD_Speed_New)) %>%
  mutate(Months_day=paste(Months,'1',sep='-')) %>%
  ggplot(aes(x=as.Date(Months_day),y=MonAvgWS,group = 1))+
  geom_line()+
  xlab('Year')+
  ylab('Wind speed')+
  labs(title = 'Average monthly wind speed in Shenzhen in 2010-2020')
```

I refer to the descriptive document and remove the data of complete error.

The results are as below:

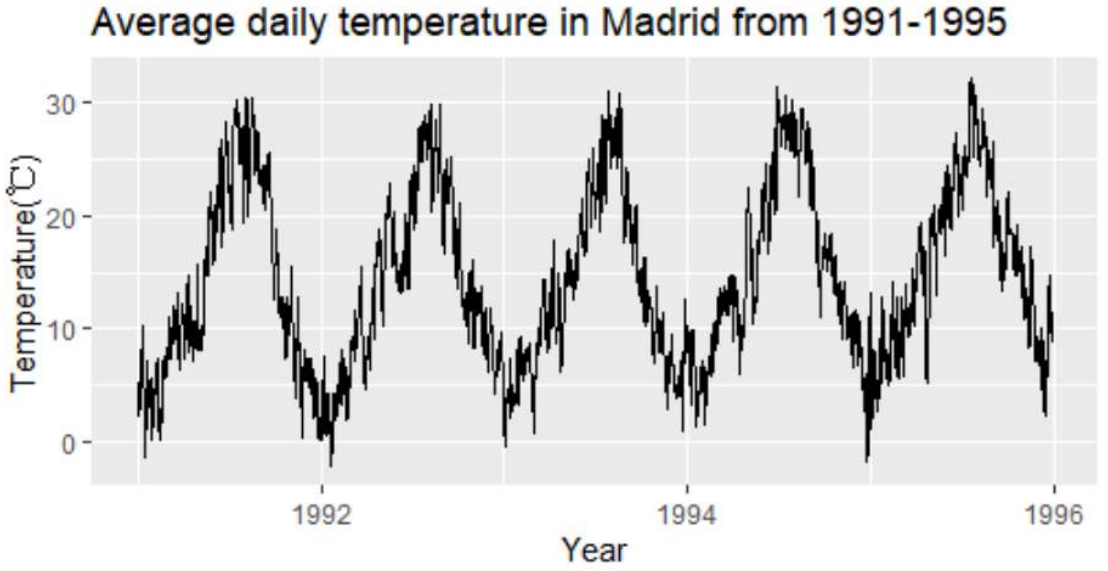


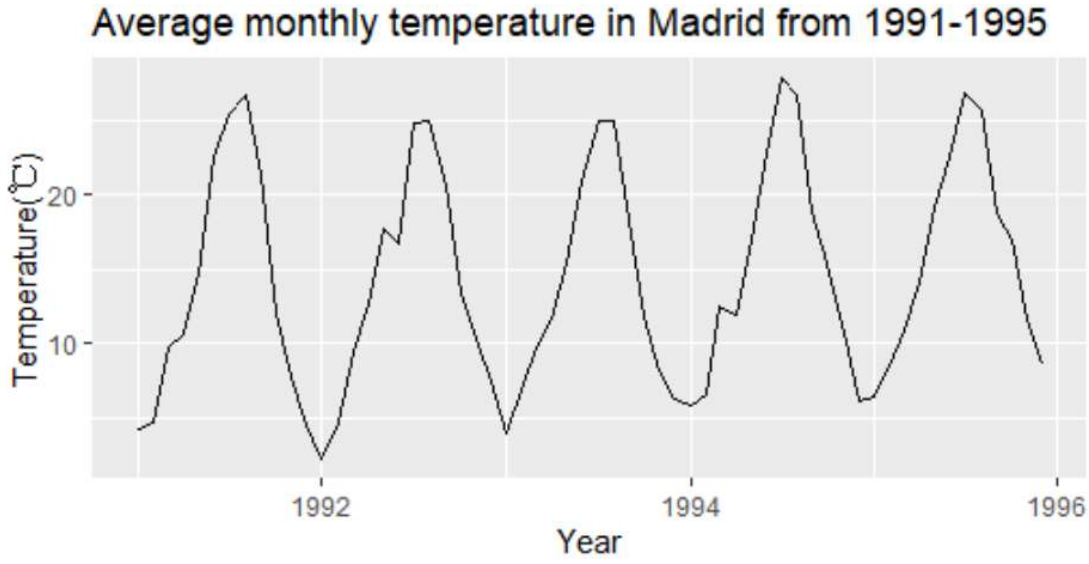
Problem#3

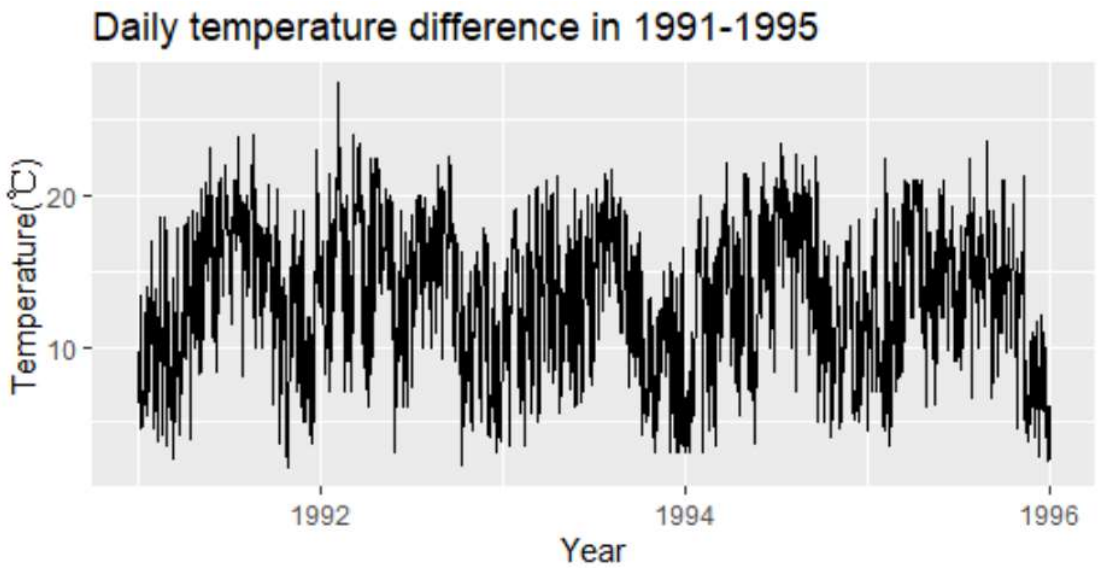
```
#3. Revisit a data set
Madrid<-read.csv(file = 'c82210-1.csv',header = T)
class(Madrid)
class(Madrid$Tm)
Madrid_tbl<-as_tibble(Madrid)

Madrid_tbl %>%
  mutate(Date=as.Date(paste(Y,M,D,sep = '-'))) %>%
  ggplot(aes(x=Date,y=AT))+
  geom_line()+
  xlab('Year')+
  ylab('Temperature(℃)')+
  labs(title = 'Average daily temperature
       in Madrid from 1991-1995')

Madrid_tbl %>%
  mutate(MonthM=as.Date(paste(Y,M,'1',sep = '-'))) %>%
  group_by(MonthM) %>%
  summarise(avgmonthAT=mean(AT)) %>%
  ggplot(aes(x=MonthM,y=avgmonthAT))+
  geom_line()+
  xlab('Year')+
  ylab('Temperature(℃)')+
  labs(title = 'Average monthly temperature
       in Madrid from 1991-1995')

Madrid_tbl %>%
  mutate(Date2=as.Date(paste(Y,M,D,sep = '-'))) %>%
  mutate(diff9195=TM-Tm) %>%
  ggplot(aes(x=Date2,y=diff9195))+
  geom_line()+
  xlab('Year')+
  ylab('Temperature(℃)')+
  labs(title = 'Daily temperature difference in 1991-1995')

class(diff1991)
shapiro.test(diff1991)
### p-value is 1.515e-10 satisfying normal distribution.

Madrid_tbl %>%
  mutate(Date2=as.Date(paste(Y,M,D,sep = '-'))) %>%
  mutate(diff9195=TM-Tm) %>%
  ggplot(aes(x=as.numeric(H),y=diff9195))+
  geom_point()+
  xlab('Humidity')+
  ylab('difference of temperature')+
  labs(title = 'Scatter plot of Humidity in respect of Temperature Difference')


Madrid_tbl %>%
  filter(H!='-') %>%
  mutate(Date2=as.Date(paste(Y,M,D,sep = '-'))) %>%
  mutate(diff9195=TM-Tm) %>%
  pull(diff9195)->diff9195

Madrid_tbl %>%
  mutate(Date2=as.Date(paste(Y,M,D,sep = '-'))) %>%
  filter(H!='-') %>%
  pull(as.numeric(H))->H

##p-value = 5.204e-16,is normally distributed
class(diff9195)
shapiro.test(diff9195)
##p-value < 2.2e-16,is normally distributed
class(H)
shapiro.test(as.numeric(H))

r3<-cor(as.numeric(H),diff9195)
```

```
Madrid_tbl %>%
  filter(H!='-') %>%
  mutate(diff9195=TM-Tm) %>%
  group_by(H) %>%
  summarise(diffavg=mean(diff9195)) %>%
  mutate(nH=as.numeric(H)) %>%
  ggplot(aes(x=nH, y=diffavg)) +
  geom_point()+
  xlab('Humidity')+
  ylab('difference of temperature')+
  labs(title = 'Scatter plot of Humidity in respect of Temperature Difference
       (doing mean to Humidity in each Humidity)')

hdregression<-lm(HD$diffavg~HD$nH,data=HD)
summary(hdregression)
abline(hdregression)
```

The results are as below:

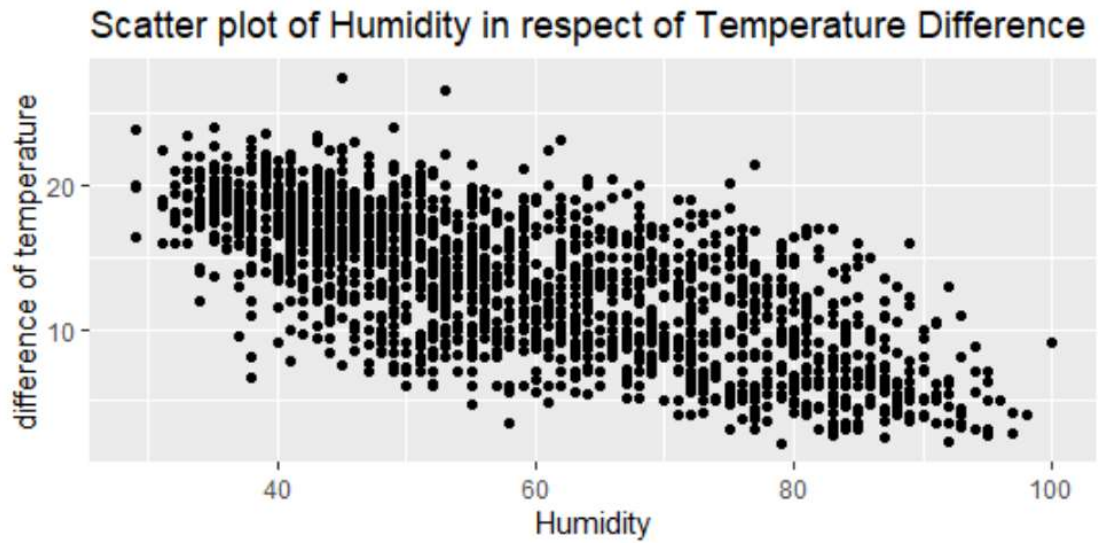






After reiterating the steps done in homework 1. I tried to find the relationship between Difference of Temperature and Humidity.

I firstly draw a scatter plot of all element.



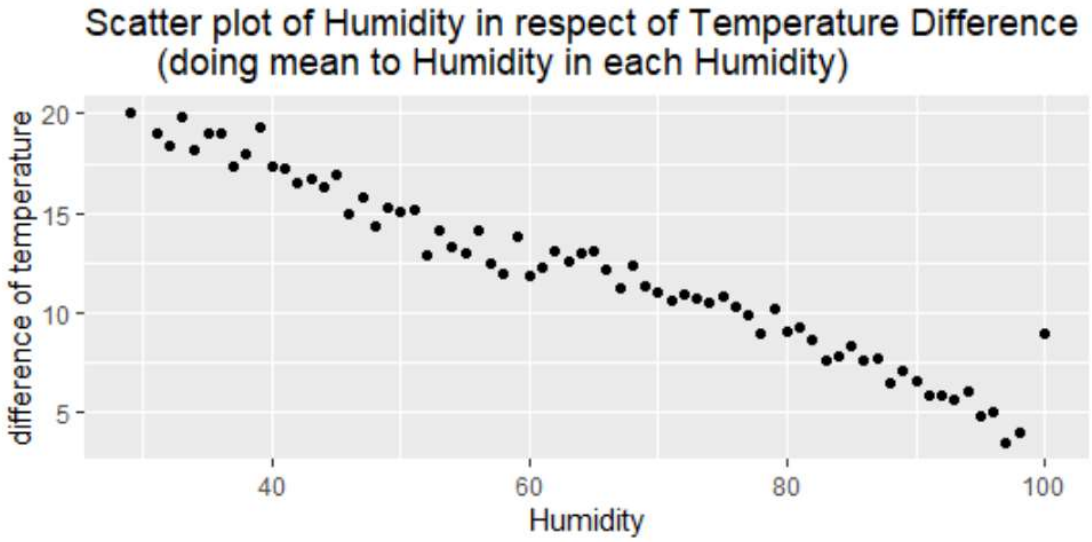
It seems there is a linearly relationship between the two variables.

Then I test whether the two variable is normally distributed.

The results show they did:

```
> shapiro.test(diff9195)

        Shapiro-Wilk normality test

data:  diff9195
W = 0.97827, p-value = 5.204e-16

> ##p-value < 2.2e-16,is normally distributed
> class(H)
[1] "character"
> shapiro.test(as.numeric(H))

        Shapiro-Wilk normality test

data:  as.numeric(H)
W = 0.96192, p-value < 2.2e-16
```

The two p-values are both in a tiny amount. The two sets both obey the normal distribution.




```
> summary(hdregression)

Call:
lm(formula = HD$diffavg ~ HD$nH, data = HD)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8678 -0.5157  0.0702  0.4196  4.2858

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.617606   0.360754   71.01   <2e-16 ***
HD$nH       -0.209034   0.005336  -39.17   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9043 on 68 degrees of freedom
Multiple R-squared:  0.9576,    Adjusted R-squared:  0.9569
F-statistic:  1535 on 1 and 68 DF,  p-value: < 2.2e-16
```

The humidity and the temperature difference can well fit linear regression.