LNM_AS1

Zhilin Yang

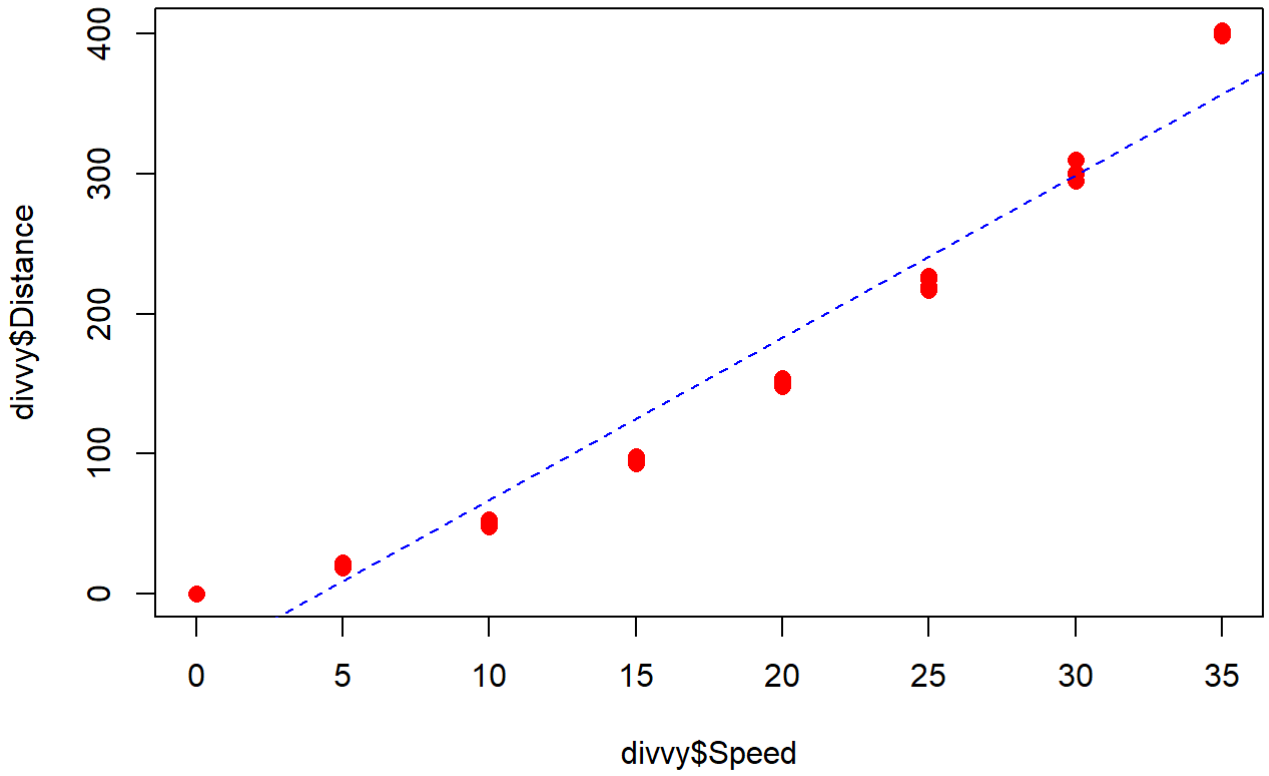
2023-01-22

Q1

```
#(A)
divvy = read.delim('C:/Users/Richa/Downloads/divvy.txt')
divvy
```

```
##    Speed Distance
## 1      0        0
## 2      5       20
## 3     10       50
## 4     15       95
## 5     20      150
## 6     25      220
## 7     30      300
## 8     35      400
## 9      5       21
## 10    10       52
## 11    15       97
## 12    20      152
## 13    25      225
## 14    30      310
## 15    35      401
## 16     0        0
## 17     5       19
## 18    10       48
## 19    15       93
## 20    20      148
## 21    25      217
## 22    30      295
## 23    35      399
## 24     0        0
## 25     5       22
## 26    10       53
## 27    15       98
## 28    20      154
## 29    25      227
## 30    30      301
## 31    35      402
```

```
plot(divvy$Speed,divvy$Distance,pch=16, col='red', cex=1.2)
abline(lm(divvy$Distance ~ divvy$Speed), col='blue' , lty='dashed')
```



As we can see from the plot, the fitted line can describe the trend well.

#(B)

```
model = lm(divvy$Distance ~ divvy$Speed)
summary(model)
```

```
##
## Call:
## lm(formula = divvy$Distance ~ divvy$Speed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -34.70 -25.20 -13.61  12.50  48.91
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -48.9070    10.1201  -4.833 4.04e-05 ***
## divvy$Speed  11.5806     0.4762  24.318  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.68 on 29 degrees of freedom
## Multiple R-squared:  0.9533, Adjusted R-squared:  0.9516
## F-statistic: 591.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

#The R square of the model is 0.9533 which is a very large R square value. It shows that 95.33% (a very large portion) of variation in Distance can be explained by the variation in Speed.

#(C)

#As we can see from the plot and the R square value, the predictor and reponse variable has a relatively linear relationship.

#In this model, the following Cautions alingned with the interpretation of R-square

#Caution #2:A large R Square does not necessarily mean that the estimated regression line fits the data well.There might be another function better describes the trend in the data. As we can see the response varialble and predictor are not perfectly linear.

#Caution #4 Correlation does not imply causation. We need further demonstration such as Hypothesis Testing to decide the distribution and the relationship.

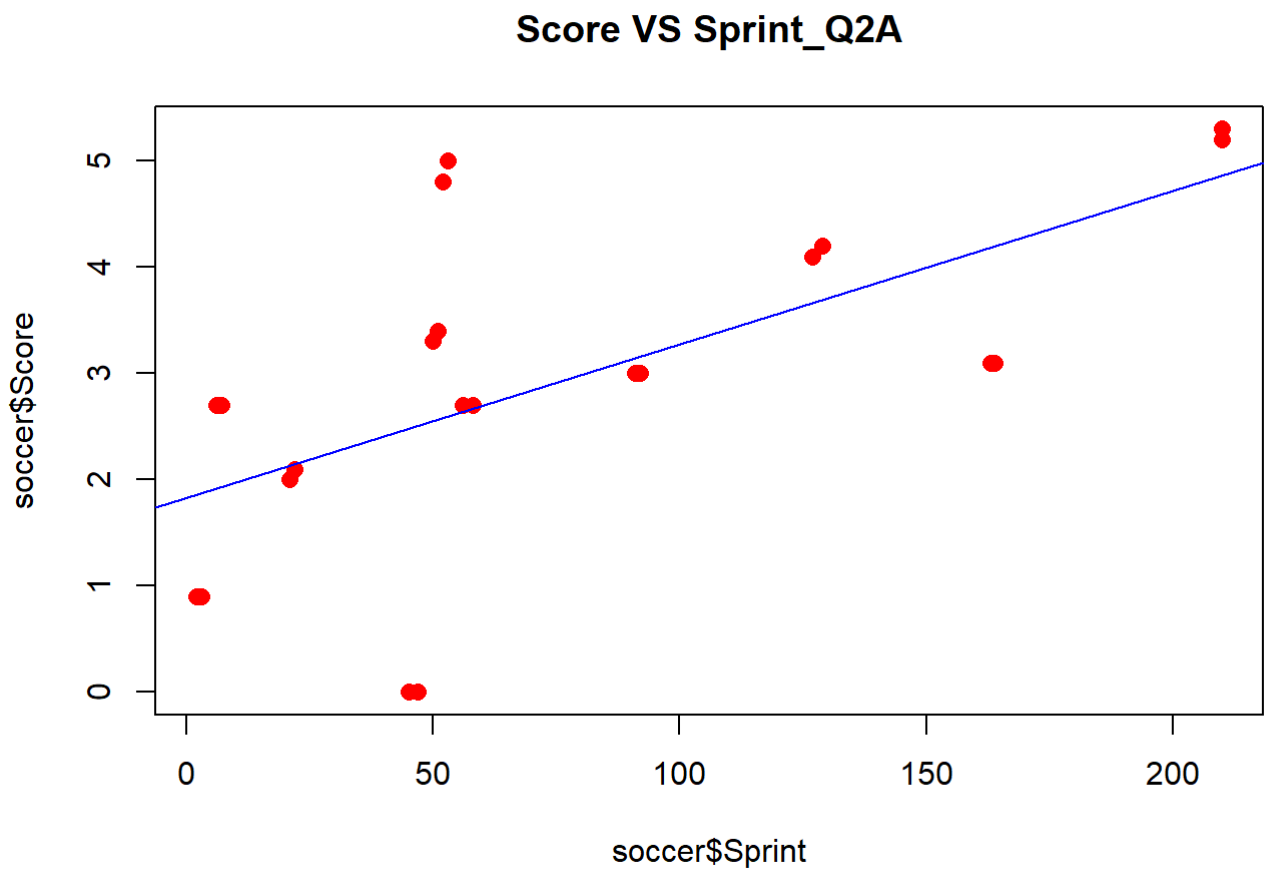
#In this model, the following cautions are less appropriate: #Caution #3: The coefficient of determination R Square and the correlation coefficient R can both be greatly affected by just adding one data(or a few) data point.This is less appropriate because there are no values significant outliers in this data

#Caution #6: A 'statistically significant' r sqaure does not imply slope beta1 is meaningfully different from 0. # In this data, the p-value of slope beta1 is 2.2e-16 which means: the chance that beta1 is insigficant is zero

#Caution #7: A large r squared value does not mean that a useful prediction of y can be made. It might be impossible to get valid prediction intervals. # In this data, the standard error of beta1 is very low, which leads to the fact that the there exists a valid confidence interval

Q2 #(A)

```
soccer = read.delim('C:/Users/Richa/Downloads/soccer.txt')
plot(soccer$Sprint,soccer$Score,,pch=16, col='red', cex=1.2,main = 'Score VS Sprint_Q2A')
model_soccer = lm(soccer$Score~soccer$Sprint)
abline(model_soccer,, col='blue' )
```



```
summary(model_soccer)$r.squared
```

```
## [1] 0.3755701
```

#The R-Square is 0.3756, there are only 37.6% variation in y can be explained by x. It shows that the model des not describe the trend well

#(B)

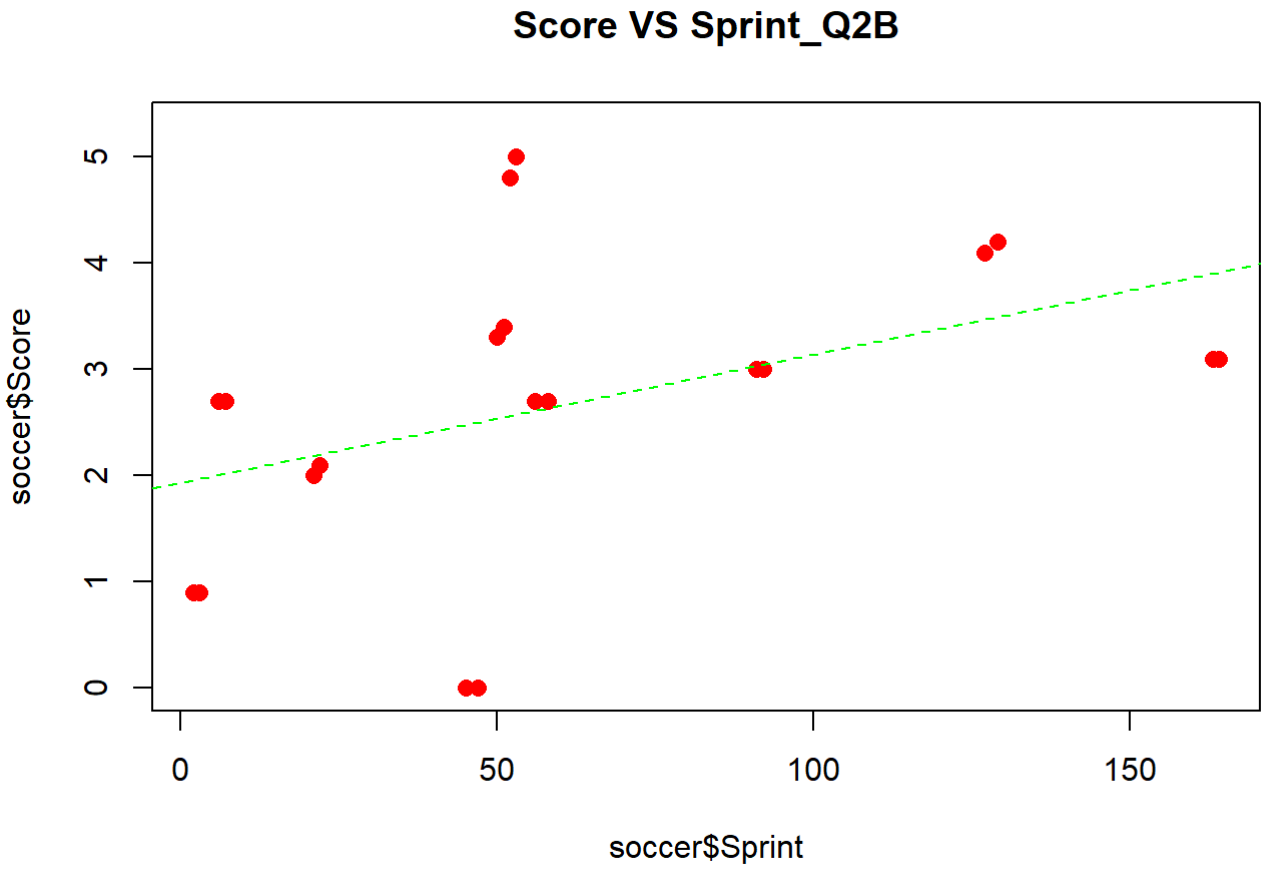
```
soccer[soccer==210]=NA
soccer_B = na.omit(soccer)
soccer_B
```

```
##    Sprint Score
## 1      51   3.4
## 2      52   4.8
## 3      21   2.0
## 4       2   0.9
## 5      56   2.7
## 6     163   3.1
## 7       6   2.7
## 8      47   0.0
## 9     127   4.1
## 10     91   3.0
## 11     50   3.3
## 12     53   5.0
## 13     22   2.1
## 14      3   0.9
## 15     58   2.7
## 16    164   3.1
## 17      7   2.7
## 18     45   0.0
## 19    129   4.2
## 20     92   3.0
```

```
model_B = lm(soccer_B$Score~soccer_B$Sprint)
summary(model_B)$r.squared
```

```
## [1] 0.1908406
```

```
plot(soccer$Sprint,soccer$Score,,pch=16, col='red', cex=1.2,main = 'Score VS Sprint_Q2B')
abline(model_B,, col='green',lty='dashed')
```

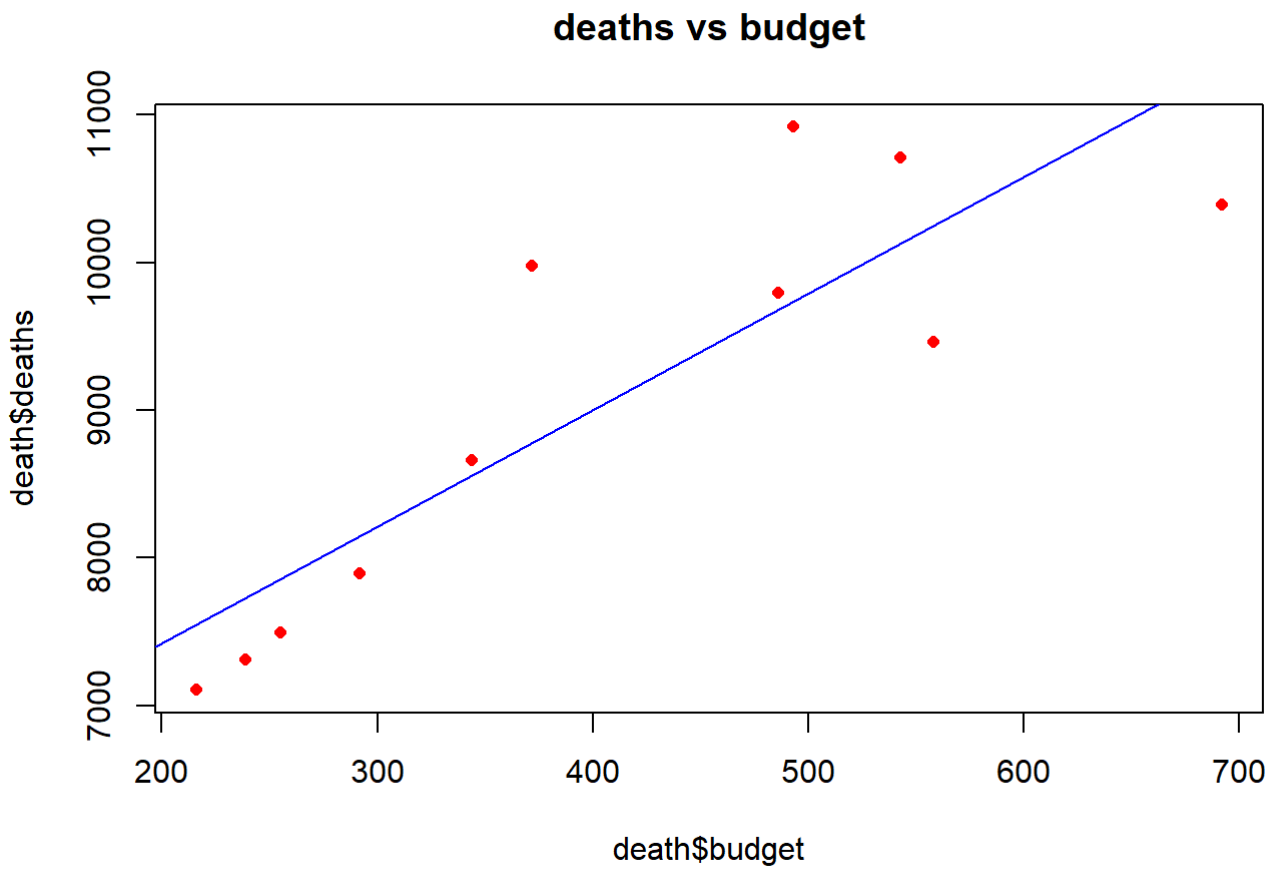


#The R Sqaure Value decreased from 0.3756 to 0.1908 after removing the two data points.

#(C) #As we can see from the above results, the R sqaure can be largely affected by the outliers and other extreme values in the data. Therefore, it is important to have a scatterplot to understand whether the correlation coefficient is affected by outliers or not.

Q3

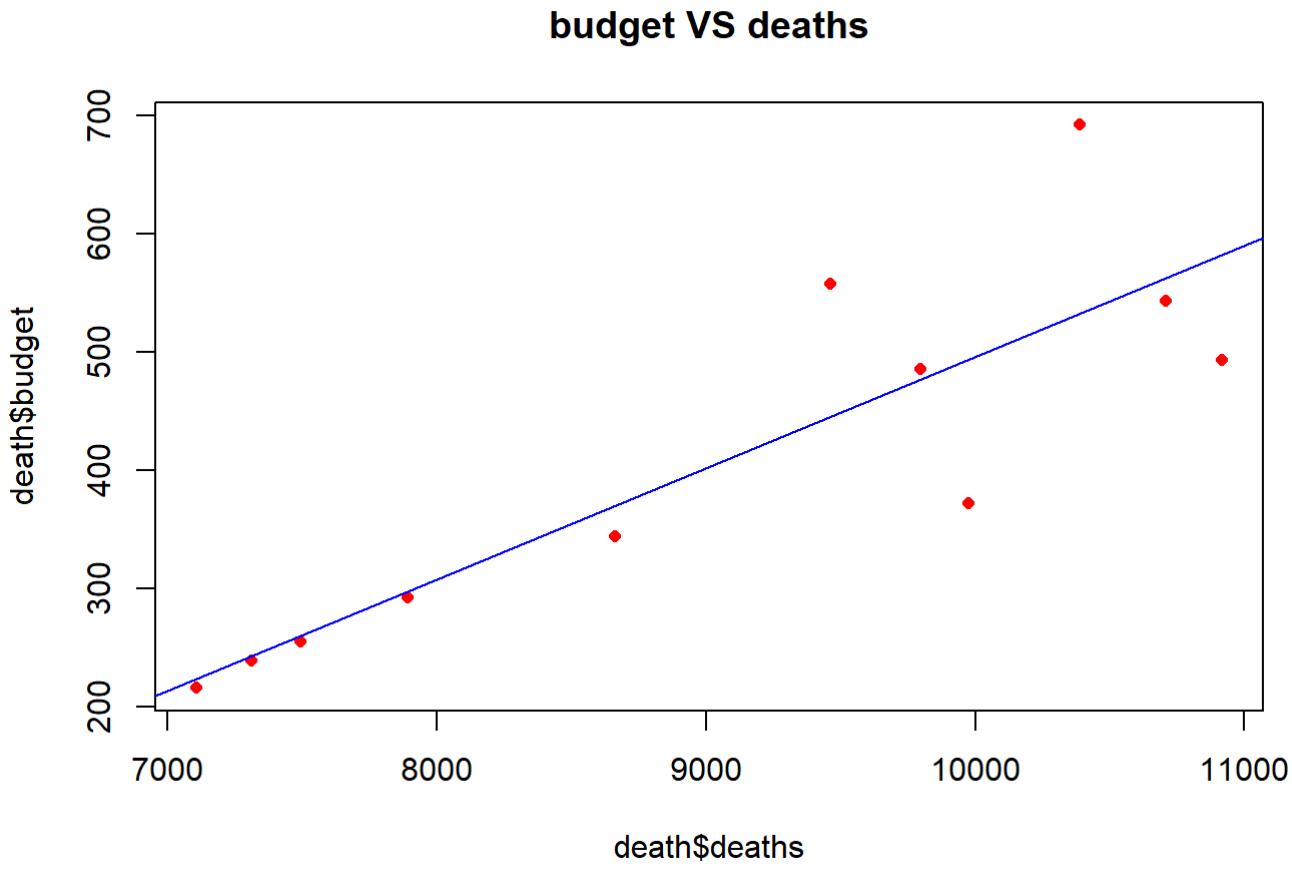
```
#(A)
death = read.delim('C:/Users/Richa/Downloads/death.txt')

model_3a = lm(death$deaths~death$budget,data = death)
plot(death$budget,death$deaths,,pch=16, col='red', cex=0.8,main = 'deaths vs budget')
abline(model_3a, col='blue')
```




```
summary(model_3a)$r.squared
```

```
## [1] 0.743423
```

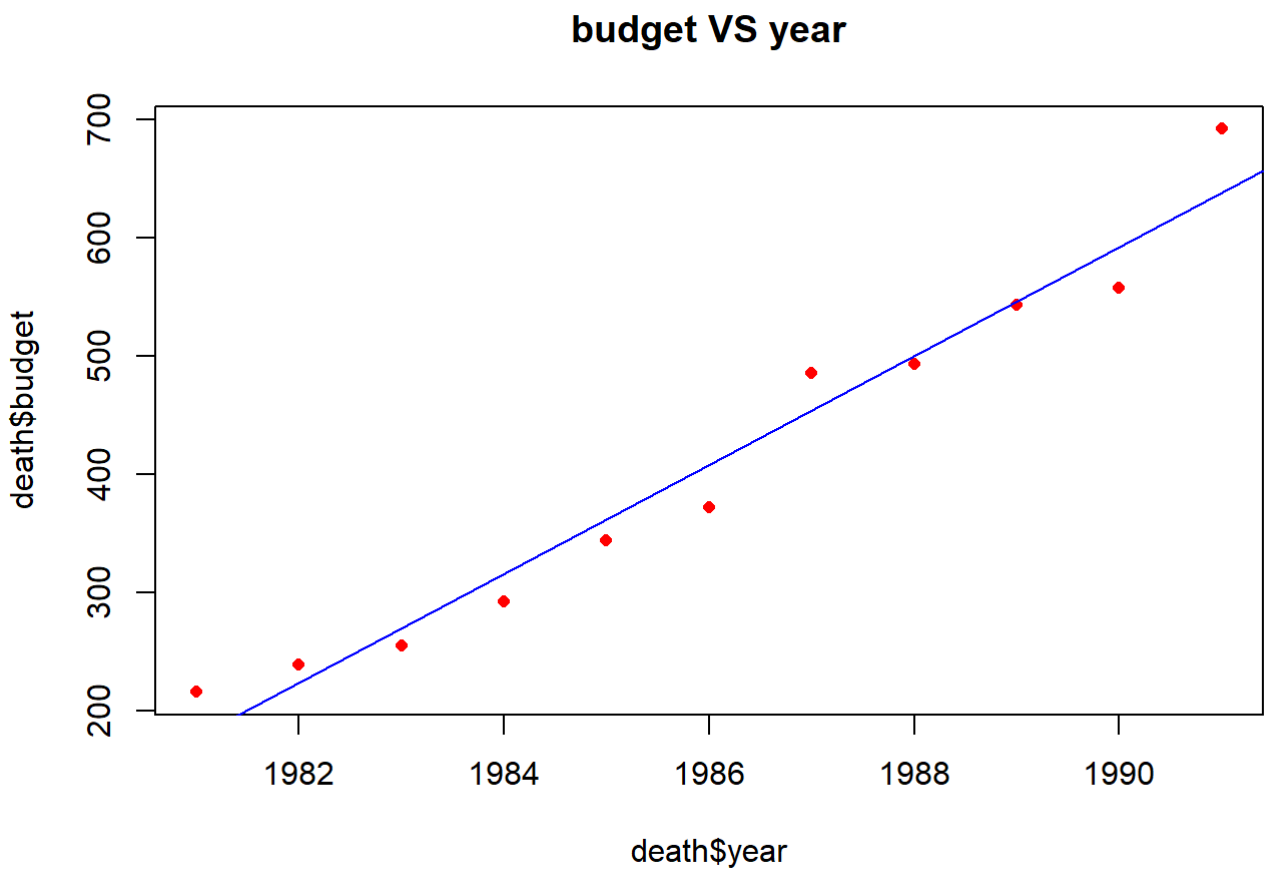
```
#I do not think the budget caused the deaths. Because there are no strong evidences from the scatterplot. The data points are not closely distribution around the fitted line which means that the variance is large and changing as well.

#(B)
model_3B = lm(death$budget~death$deaths,data = death)
plot(death$deaths,death$budget,,pch=16, col='red', cex=0.8,main = 'budget VS deaths')
abline(model_3B, col='blue')
```




```
summary(model_3B)$r.squared
```

```
## [1] 0.743423
```

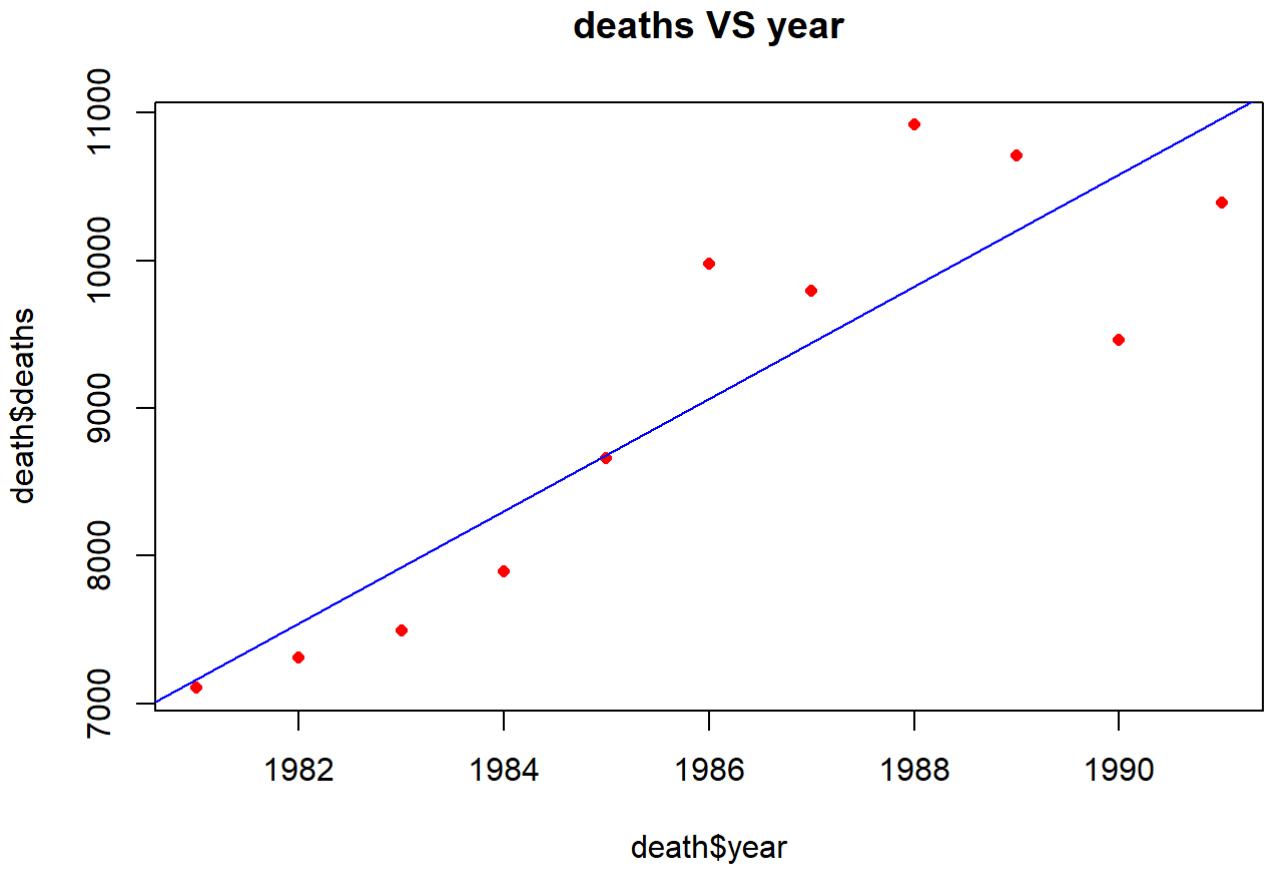
```
#I do not think the deaths caused the budget. Although there are some linear relationship, th
ere are not enough evidence the prove causality.

#(C)
model_3c = lm(death$budget~death$year,data = death)
plot(death$year,death$budget,,pch=16, col='red', cex=0.8,main = 'budget VS year')
abline(model_3c, col='blue')
```




```
summary(model_3c)$r.squared
```

```
## [1] 0.9618623
```

```
#(D)
model_3c = lm(death$deaths~death$year,data = death)
plot(death$year,death$deaths,,pch=16, col='red', cex=0.8,main = 'deaths VS year')
abline(model_3c, col='blue')
```




```
summary(model_3c)$r.squared
```

```
## [1] 0.7831877
```

#(E)

#As we can see from (A) and (B), there is a positive correlation between budget and deaths.

#As we can see from (C) and (D), both of the budget and deaths are increasing along with time. Therefore, this lead to a fact: The positive correlation between deaths and budget may be larged affected by the lurking variable time.

#Along with time, the government expenditure budget, the total population and the accessibility of drugs due to technology progress might all contributed to the drug-induced deaths. The two variables (deaths and budget) have positive correlations might because they are both increasing along with time.