Dandong Tu

S631 HW2

1.

a)

```
> head(UN11)
                   region    group fertility   ppgdp lifeExpF pctUrban
Afghanistan          Asia    other     5.968   499.0    49.49       23
Albania            Europe    other     1.525  3677.2    80.40       53
Algeria            Africa africa       2.142  4473.0    75.00       67
Angola             Africa africa       5.135  4321.9    53.17       59
Anguilla        Caribbean    other     2.000 13750.1    81.10      100
Argentina      Latin Amer    other     2.172  9162.1    79.89       93
> Fertility=UN11$fertility
> fertility.r=round(Fertility)
> head(lifeexp)
[1] 49.49 80.40 75.00 53.17 81.10 79.89
> mean(lifeexp)
[1] 72.29319
> var(lifeexp)
[1] 102.491
```

b)

```
> AA=data.frame(UN11$lifeExpF,fertility.r)
> head(AA)
  UN11.lifeExpF fertility.r
1         49.49           6
2         80.40           2
3         75.00           2
4         53.17           5
5         81.10           2
```

6 79.89 2

```
> E1=AA[which.names("1",AA$fertility.r),]
> E2=AA[which.names("2",AA$fertility.r),]
> E3=AA[which.names("3",AA$fertility.r),]
> E4=AA[which.names("4",AA$fertility.r),]
> E5=AA[which.names("5",AA$fertility.r),]
> E6=AA[which.names("6",AA$fertility.r),]
> E7=AA[which.names("7",AA$fertility.r),]
```

```
> mean(E1$UN11.lifeExpF)
[1] 80.96565
> mean(E2$UN11.lifeExpF)
[1] 77.77853
> mean(E3$UN11.lifeExpF)
[1] 68.85352
> mean(E4$UN11.lifeExpF)
[1] 64.70913
> mean(E5$UN11.lifeExpF)
[1] 57.55556
> mean(E6$UN11.lifeExpF)
[1] 54.38778
> mean(E7$UN11.lifeExpF)
[1] 55.77
```

c)

```
> var(E1$UN11.lifeExpF)
[1] 13.15358
> var(E2$UN11.lifeExpF)
```

[1] 22.69346

> var(E3$UN11.lifeExpF)

[1] 86.26717

> var(E4$UN11.lifeExpF)

[1] 55.31225

> var(E5$UN11.lifeExpF)

[1] 38.8089

> var(E6$UN11.lifeExpF)

[1] 19.76342

> var(E7$UN11.lifeExpF)

[1] NA

2.

a)

Since fertility is the dependence on ppgdp, so that the fertility is the response and the ppgdp is predictor.

b)

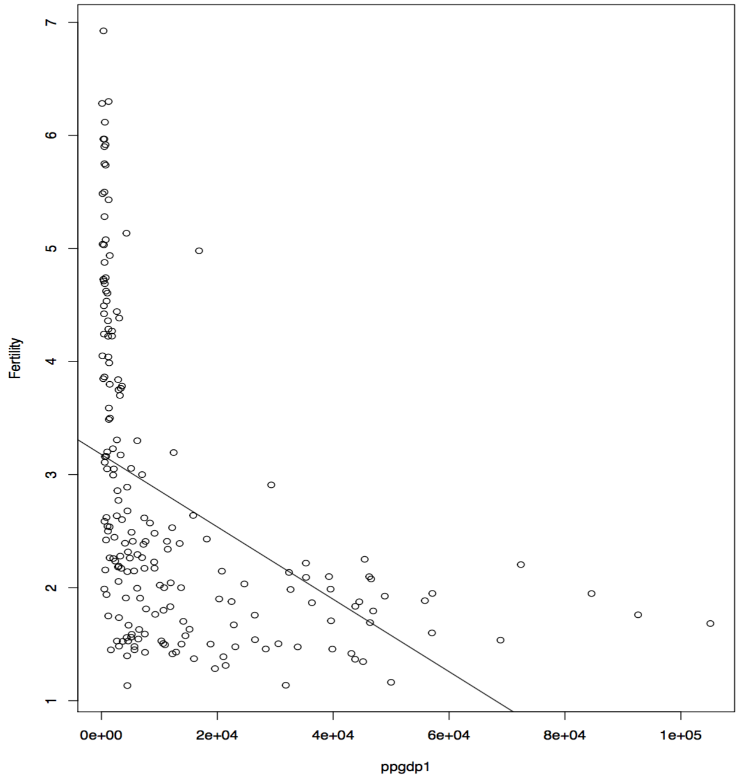
ppgdp1=UN11$ppgdp

> head(ppgdp1)

[1]   499.0   3677.2   4473.0   4321.9 13750.1   9162.1

> plot(x=ppgdp1,y=Fertility)

> abline(lm(Fertility~ppgdp1))



Based on the graph, there is a big portion of points fall near the vertical axis(Fertility) in which the horizontal value near(a little bit higher than "0e+00"). And when predictor value(ppgdp1) increasing, there is not such a big trend shows the straight-line relation between predictor and response. Therefore, it seems a straight-line mean function is implausible for this graphy.

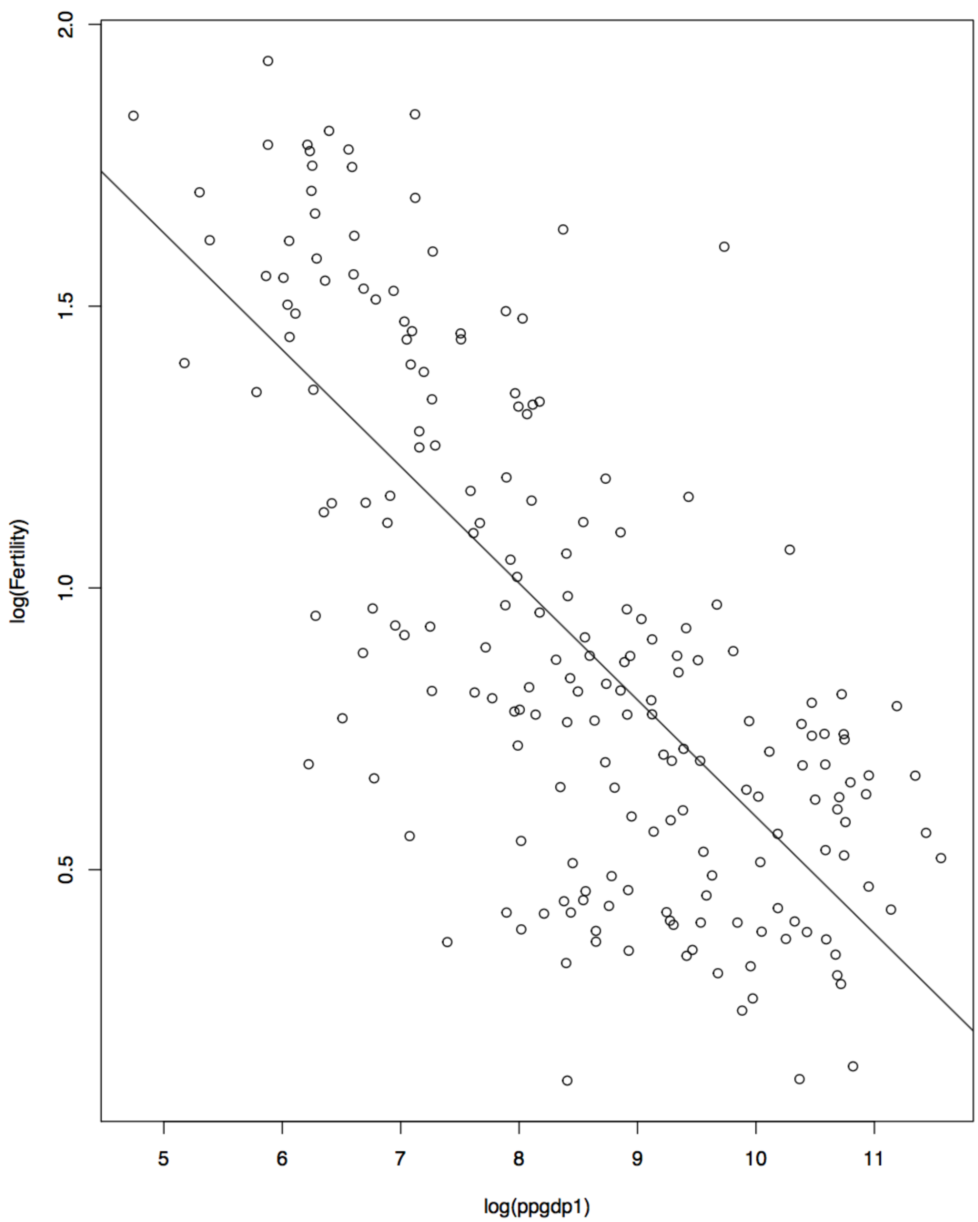
```
> summary(lm(Fertility~ppgdp1))

Call:
lm(formula = Fertility ~ ppgdp1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9006 -0.8801 -0.3547  0.6749  3.7585
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.178e+00  1.048e-01  30.331  < 2e-16 ***
ppgdp1      -3.201e-05  4.655e-06  -6.877  7.9e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.206 on 197 degrees of freedom
Multiple R-squared:  0.1936,    Adjusted R-squared:  0.1895
F-statistic: 47.29 on 1 and 197 DF,  p-value: 7.903e-11
```

The summary shows that the value of R-squared equal 0.1936 which indicates that only 19.36% of a straight-line mean function can explain the variability of the response data. Also, the 1.206 of residual value indicates a high difference between the points to the line. So that, a straight-line mean function does NOT seem to be plausible for a summary of this graph.

c)

```
> plot(x=log(ppgdp1),y=log(Fertility))
> abline(lm(log(Fertility)~log(ppgdp1)))
```



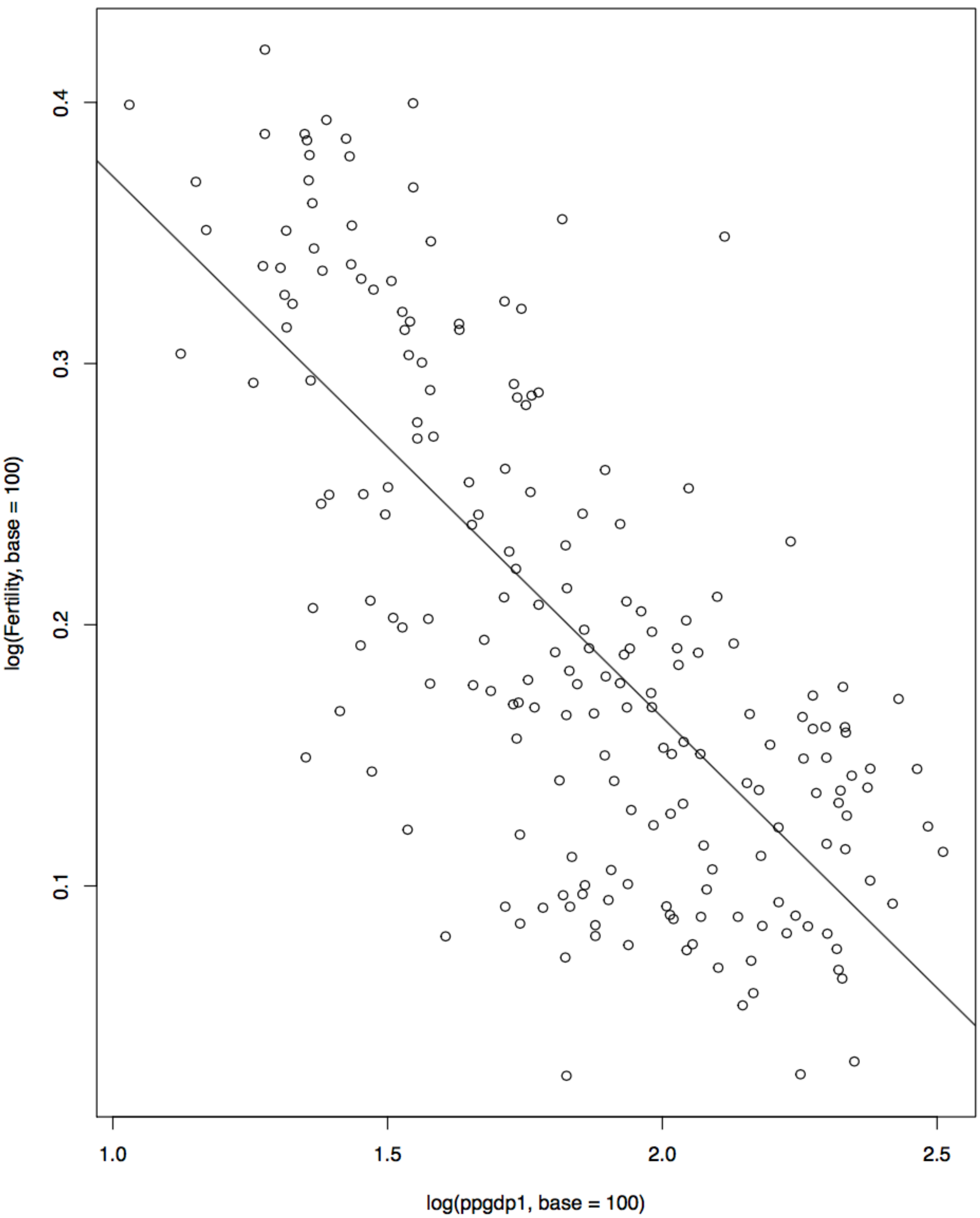
Based on the graph, the trend is that log(Fertility) decreases with log(ppgdp1), however, it is not exact. For example, there are many points near log(ppgdp1) value of "8" far away low than points in the top of "9" "10" and "11. So that the trend is more clear but knowing the log(ppgdp1) still not allow us to predict the Fertility exactly.

```
> summary(lm(log(Fertility)~log(ppgdp1)))
```

```
Call:
lm(formula = log(Fertility) ~ log(ppgdp1))

Residuals:
     Min       1Q   Median       3Q      Max
-0.79828 -0.21639  0.02669  0.23424  0.95596

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.66551    0.12057   22.11   <2e-16 ***
log(ppgdp1) -0.20715    0.01401  -14.79   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3071 on 197 degrees of freedom
Multiple R-squared:  0.526, Adjusted R-squared:  0.5236
F-statistic: 218.6 on 1 and 197 DF,  p-value: < 2.2e-16
```

From the summary, the residual value is 0.3071 which is far away less than the previous one(which is 1.206). Also, we can see that the value of R-squared equal 0.526 which indicates that 52.6% of a straight-line mean function can explain the variability of the response data. Therefore it seems a straight-line mean function should to be more plausible for a summary of this graph but still not allow us to predict exactly.

```
plot(x=log(ppgdp1,base=100),y=log(Fertility,base=100))
abline(lm(log(Fertility,base=100)~log(ppgdp1,base=100)))
```



Different base doesn’ t change the shape.

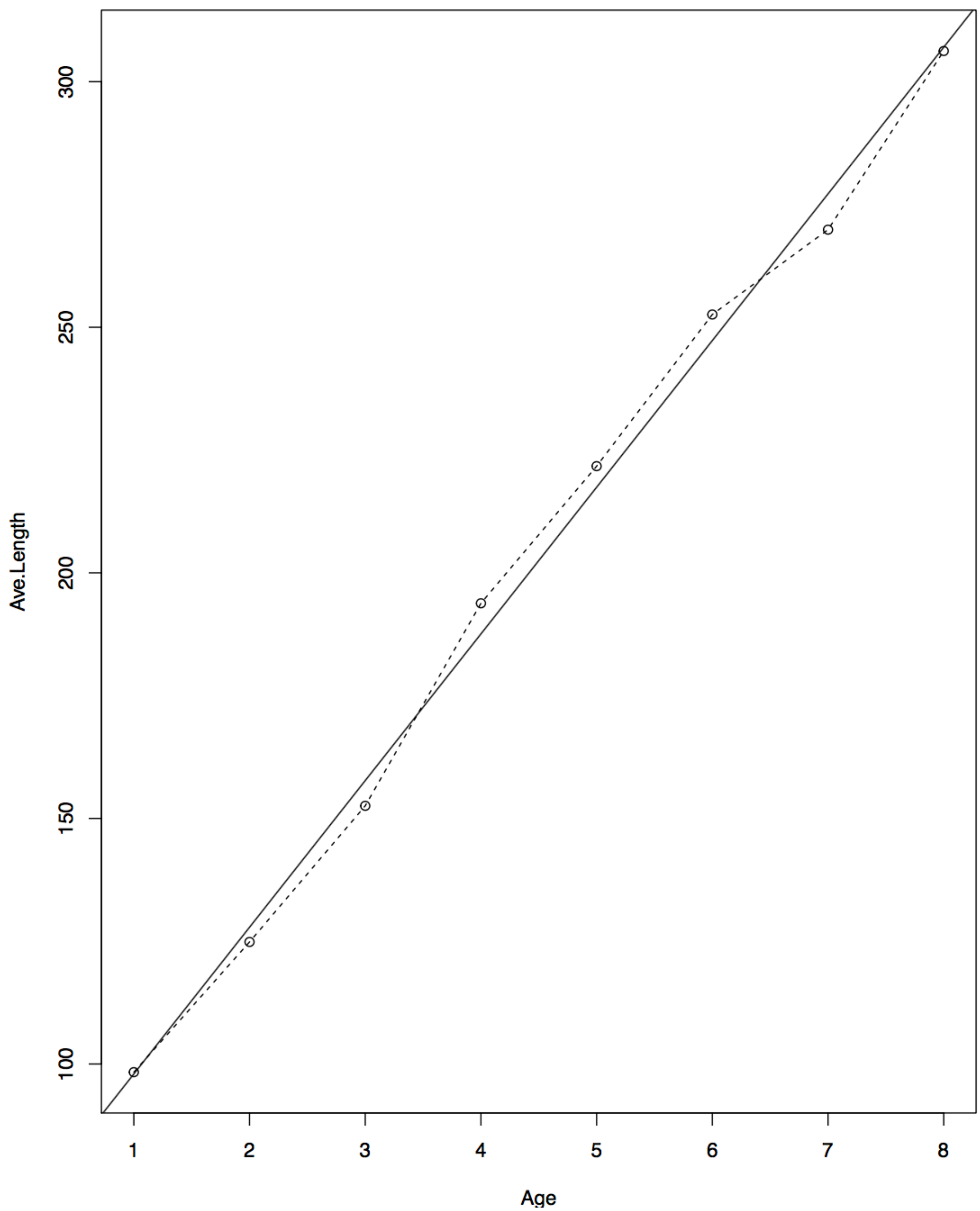
3.

```
> head(wblake)
  Age Length   Scale
1   1     71 1.90606
2   1     64 1.87707
3   1     57 1.09736
4   1     68 1.33108
5   1     72 1.59283
6   1     80 1.91602
```

```
> tapply(wblake$Length,wblake$Age,mean)
        1         2         3         4         5         6 
 98.34211 124.84722 152.56383 193.80000 221.72059 252.59770 
        7         8 
269.86885 306.25000 
> tapply(wblake$Length,wblake$Age,var)
        1         2         3         4         5         6 
 808.2312  697.2862  411.6679  867.4571  985.6969 1105.0805 
        7         8 
 869.3825 1802.9167 
```

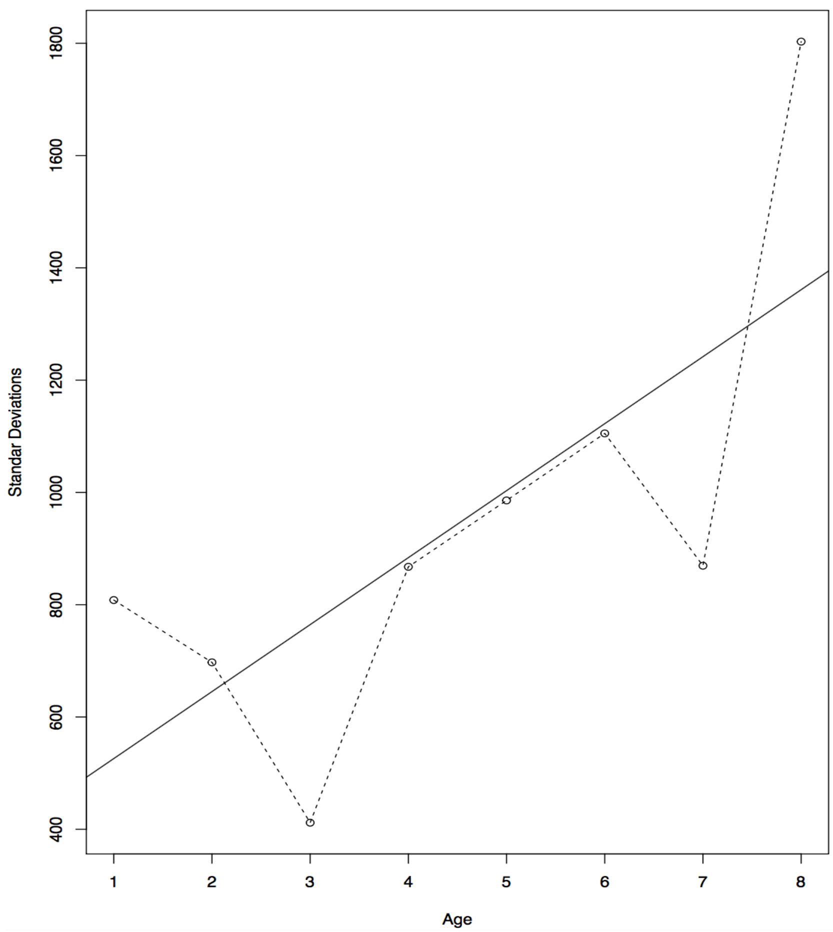
```
> Ave.Length=tapply(wblake$Length,wblake$Age,mean)
> Ave.Length
        1         2         3         4         5         6 
 98.34211 124.84722 152.56383 193.80000 221.72059 252.59770 
        7         8 
269.86885 306.25000 
```

```
> AAge=c(1:8)
> AAge
[1] 1 2 3 4 5 6 7 8
> plot(Ave.Length,xlab="Age")
> lines(1:8,Ave.Length, lty=2)
> abline(lm(Ave.Length~AAge))
```



Comparing the graph with Figure 1.5, two lines(solid line and dashed line) are same relatively since the procedure of graphing the lines in Figure 1.5 are same as using the average length in each age subpopulations.

```
> Var1=tapply(wblake$Length,wblake$Age,var)
> Var1
        1         2         3         4         5         6 
 808.2312  697.2862  411.6679  867.4571  985.6969 1105.0805 
        7         8 
 869.3825 1802.9167 
> plot(Var1,xlab = "Age", ylab = "Standar Deviations")
> lines(1:8,Var1,lty=2)
> abline(lm(Var1~AAge))
```



In the graph, it indicates that the variance is not constant, therefore, it is not a null plot.