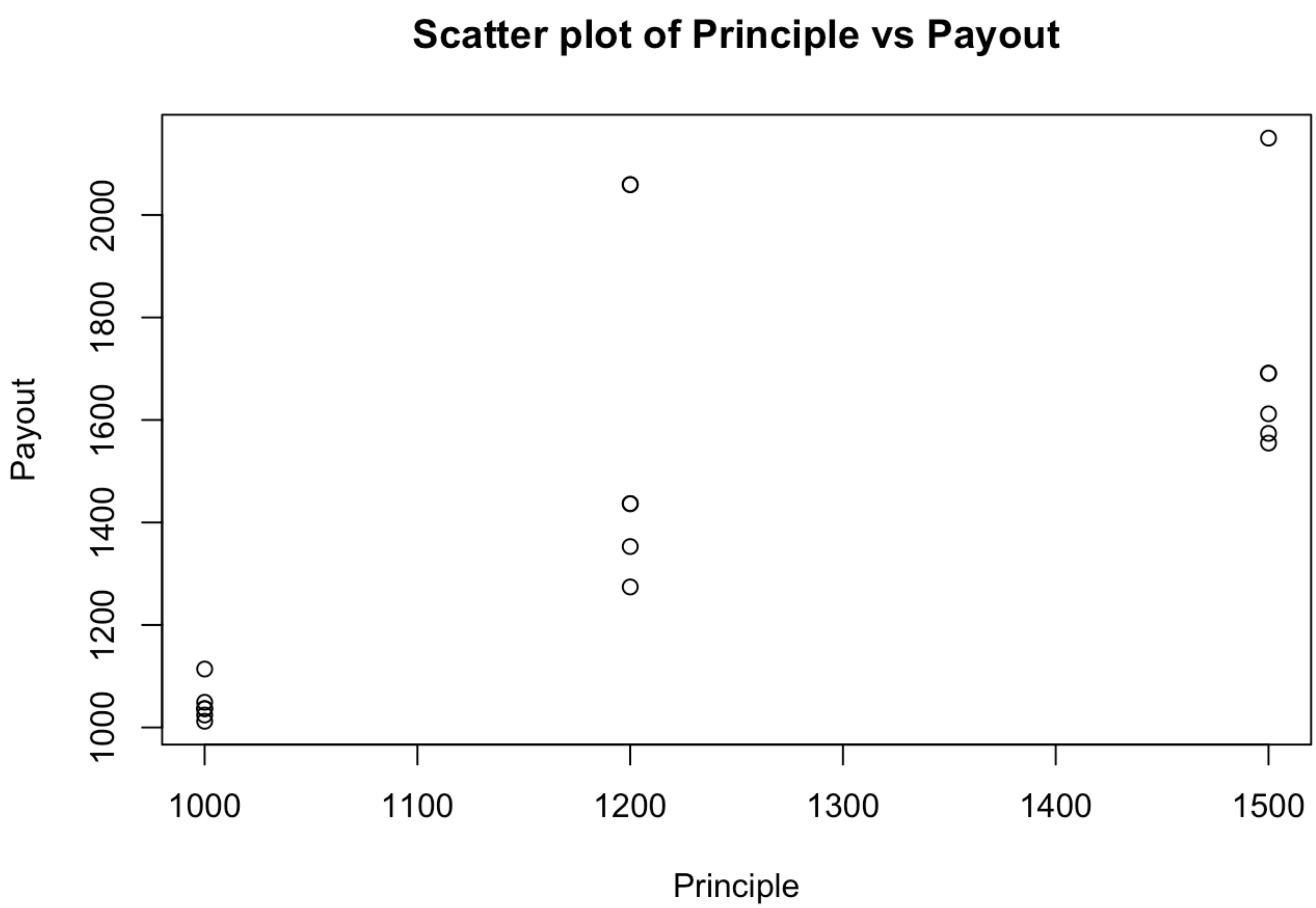
STAT 353 Assignment1

Zimeng Ming V00844078

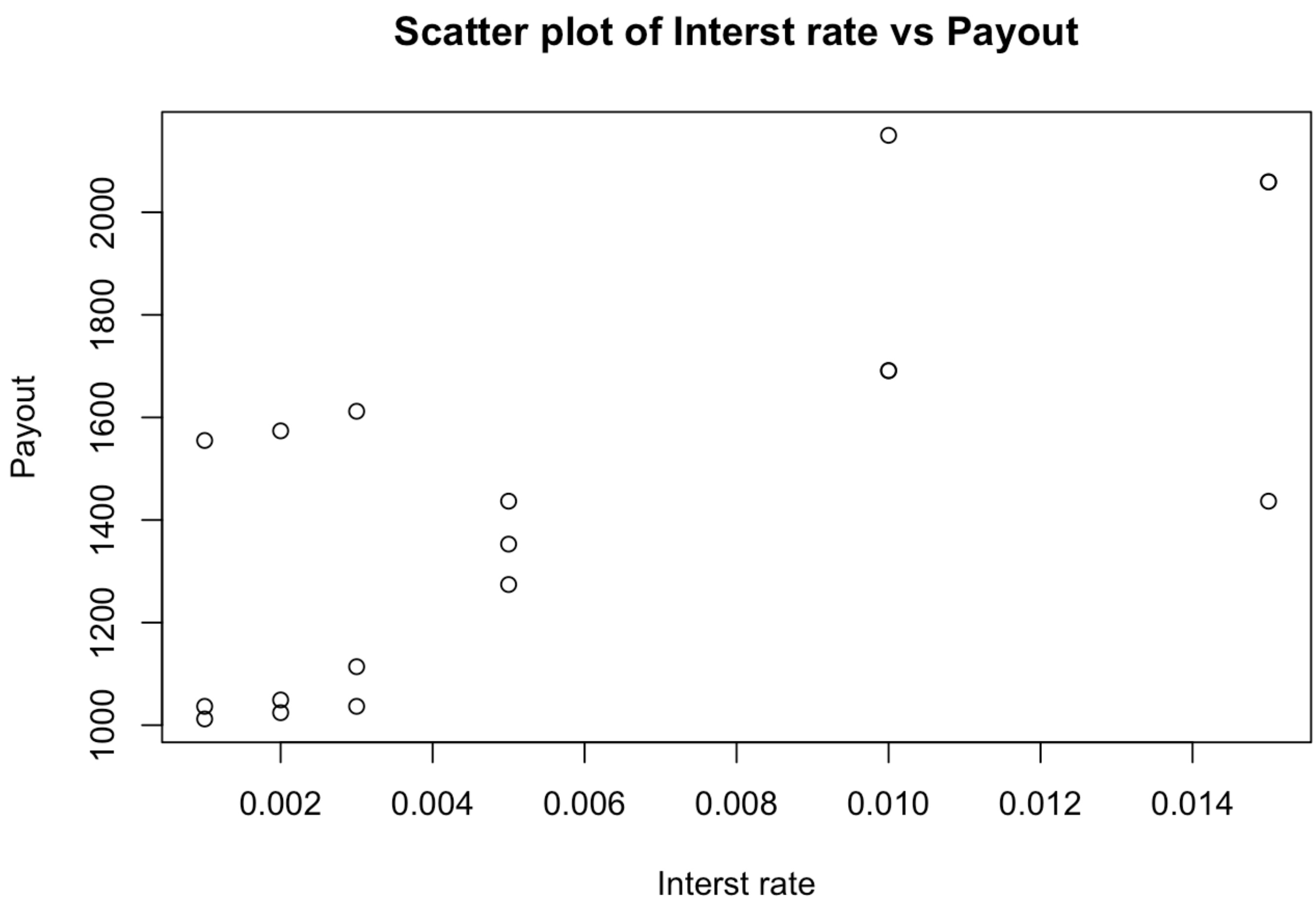
2019/9/20

Question 2 (Q1.2)

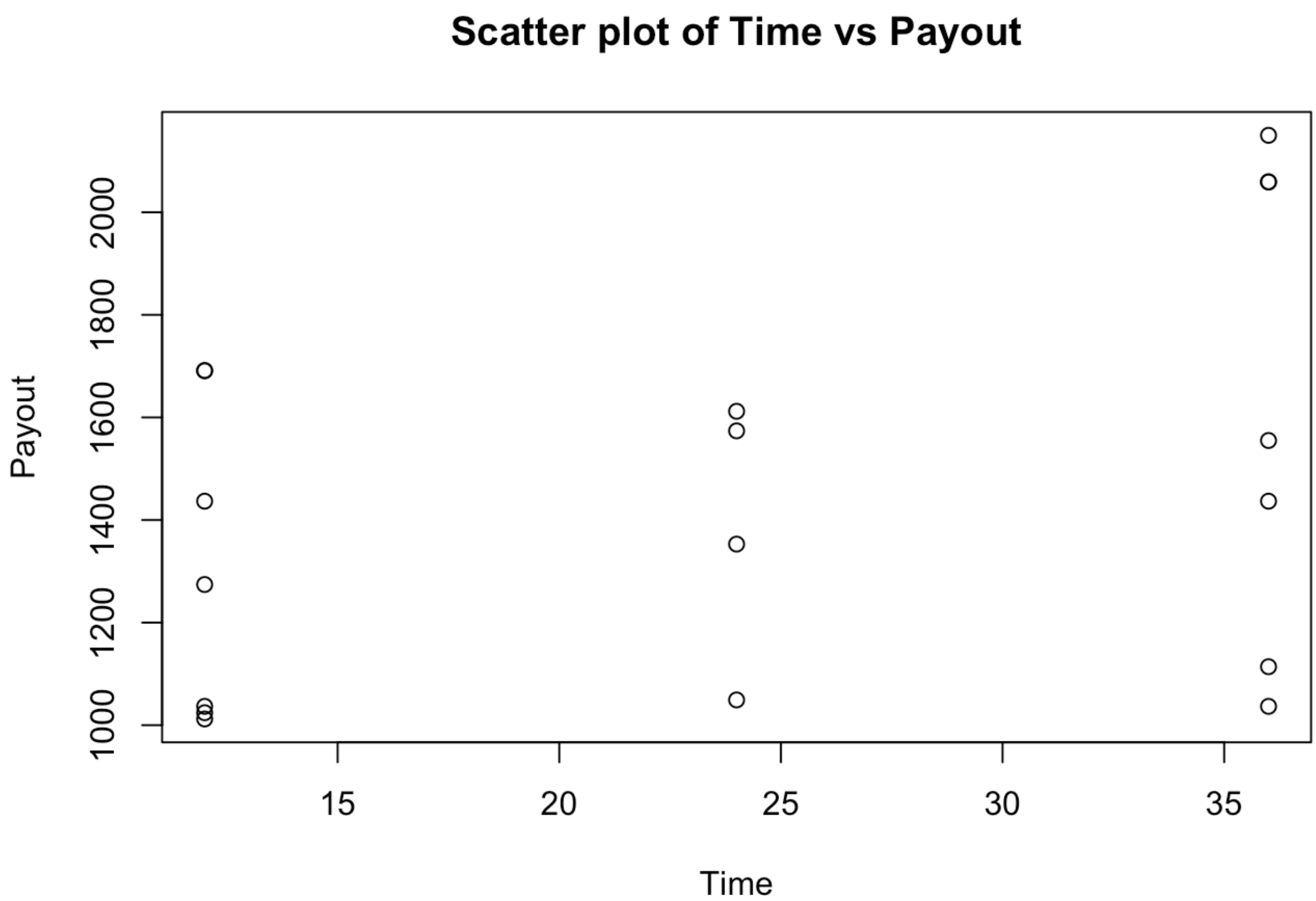
```
#get the data into the R
payout<-read.table("payout.txt",header = T)
#first we plot the scatter plots of the payout with explanatory variables.
plot(payout$Prin,payout$Payout,xlab = "Principle",ylab = "Payout",main="Scatter plot
of Principle vs Payout")
```




```
plot(payout$Int,payout$Payout,xlab = "Interst rate",ylab = "Payout",main="Scatter plo
t of Interst rate vs Payout")
```

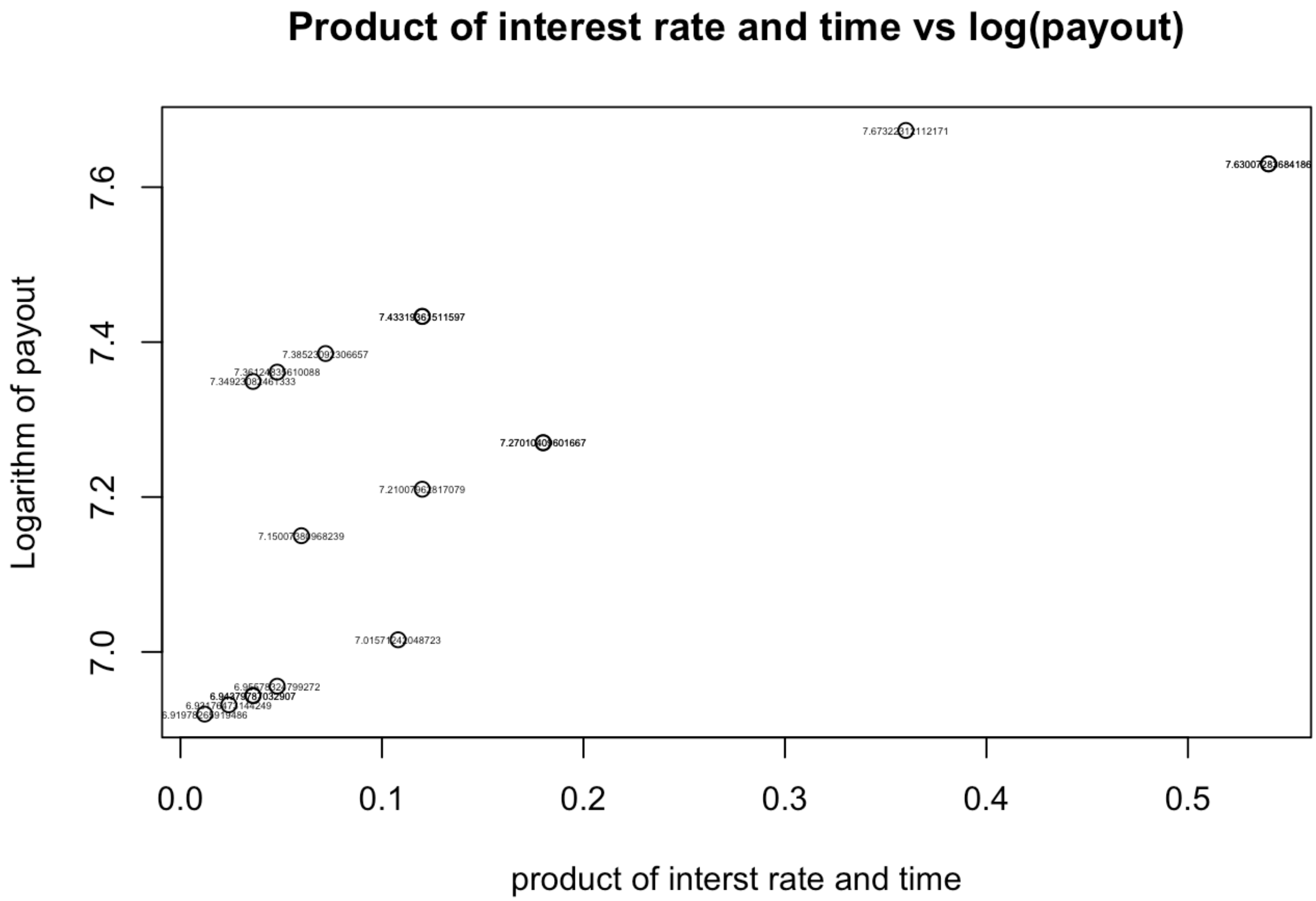



```
plot(payout$Time,payout$Payout,xlab = "Time",ylab = "Payout",main="Scatter plot of Ti
me vs Payout")
```



Comment: As we can see from the above three plot, there is only one plot that the plot of “payout verus principle” looks having linear relationships, however, the other two plot do not present the clear linear relationship from the plot.

```
#As for the logarithm of payout against the product of the interest rate and maturity
plot(payout$Int*payout$Time,log(payout$Payout),xlab = "product of interst rate and ti
me",ylab = "Logarithm of payout", main = "Product of interest rate and time vs log(pa
yout)", text(payout$Int*payout$Time,log(payout$Payout),log(payout$Payout),cex = 0.3,o
ffset = 8))
```



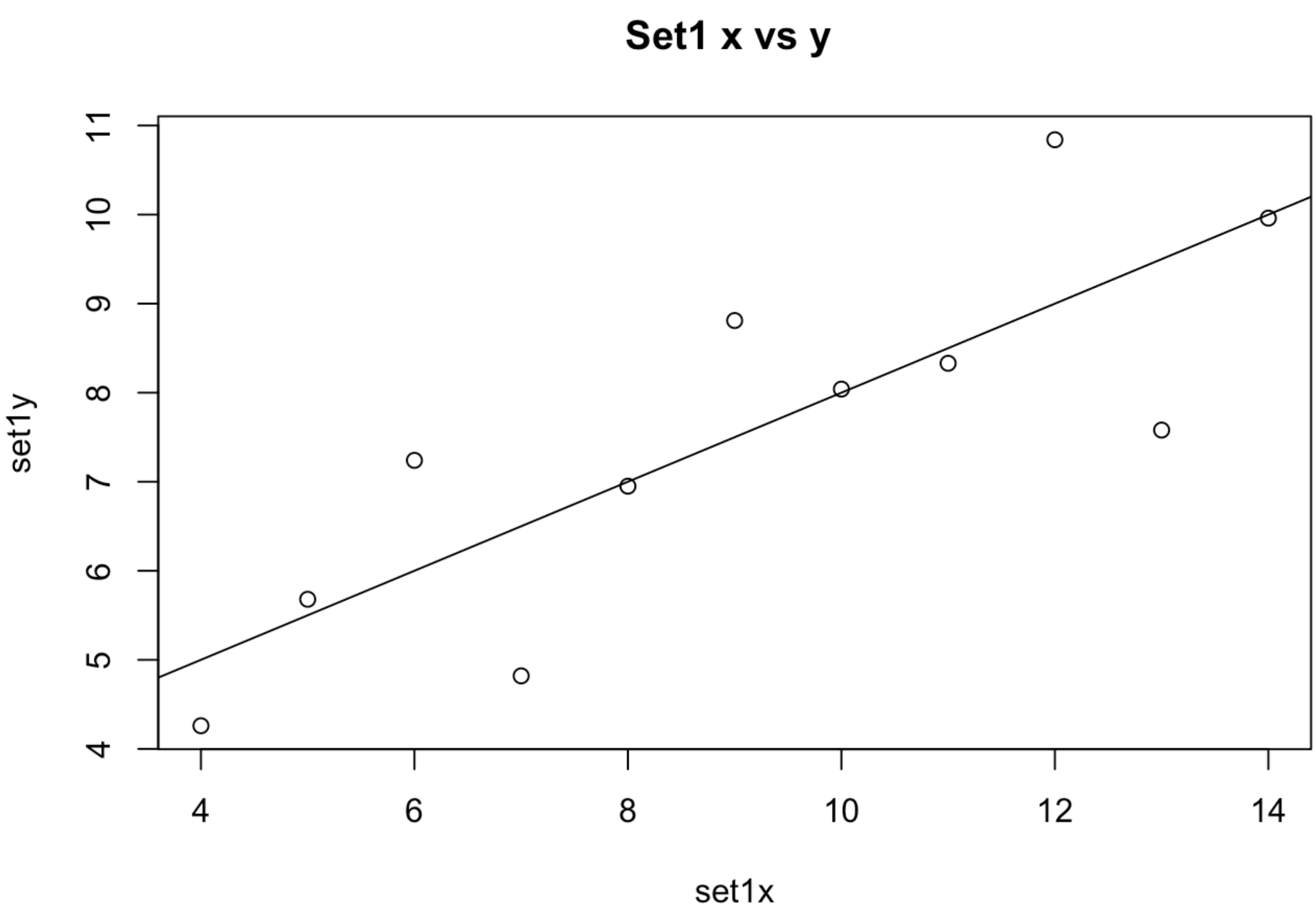
Comment: As we can see from above, the plot looks like having an linear relationship. There is three increasing linear model that indicate three logarithm payout level, log(2000), log(1500) and log(1000).

Question 3(1.12) (on hand page)

Question 4(2.2)(on hand page)

Question 5(2.3) a):

```
anscombe<-read.table("anscombe.txt",header = T)
#plot the 1x 1y
plot(anscombe$Set1x,anscombe$Set1y,xlab = "set1x",ylab = "set1y",main = "Set1 x vs y
")
#get the linear model of set 1
set1_lm<-lm(anscombe$Set1y~anscombe$Set1x)
abline(set1_lm)
```




```
#summary the model
summary(set1_lm)
```

```
##
## Call:
## lm(formula = anscombe$Set1y ~ anscombe$Set1x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.92127 -0.45577 -0.04136  0.70941  1.83882
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.0001     1.1247   2.667  0.02573 *
## anscombe$Set1x   0.5001     0.1179   4.241  0.00217 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6665, Adjusted R-squared:  0.6295
## F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```

```
#anova of the model
anova(set1_lm)
```

```
## Analysis of Variance Table
##
## Response: anscombe$Set1y
##                Df Sum Sq Mean Sq F value  Pr(>F)
## anscombe$Set1x  1 27.510 27.5100   17.99 0.00217 **
## Residuals       9 13.763  1.5292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#correlation
cor(anscombe$Set1x,anscombe$Set1y)
```

```
## [1] 0.8164205
```

```
#r square
cor(anscombe$Set1x,anscombe$Set1y)^2
```

```
## [1] 0.6665425
```

Comment: Fitted line y=3.0001+0.5001x

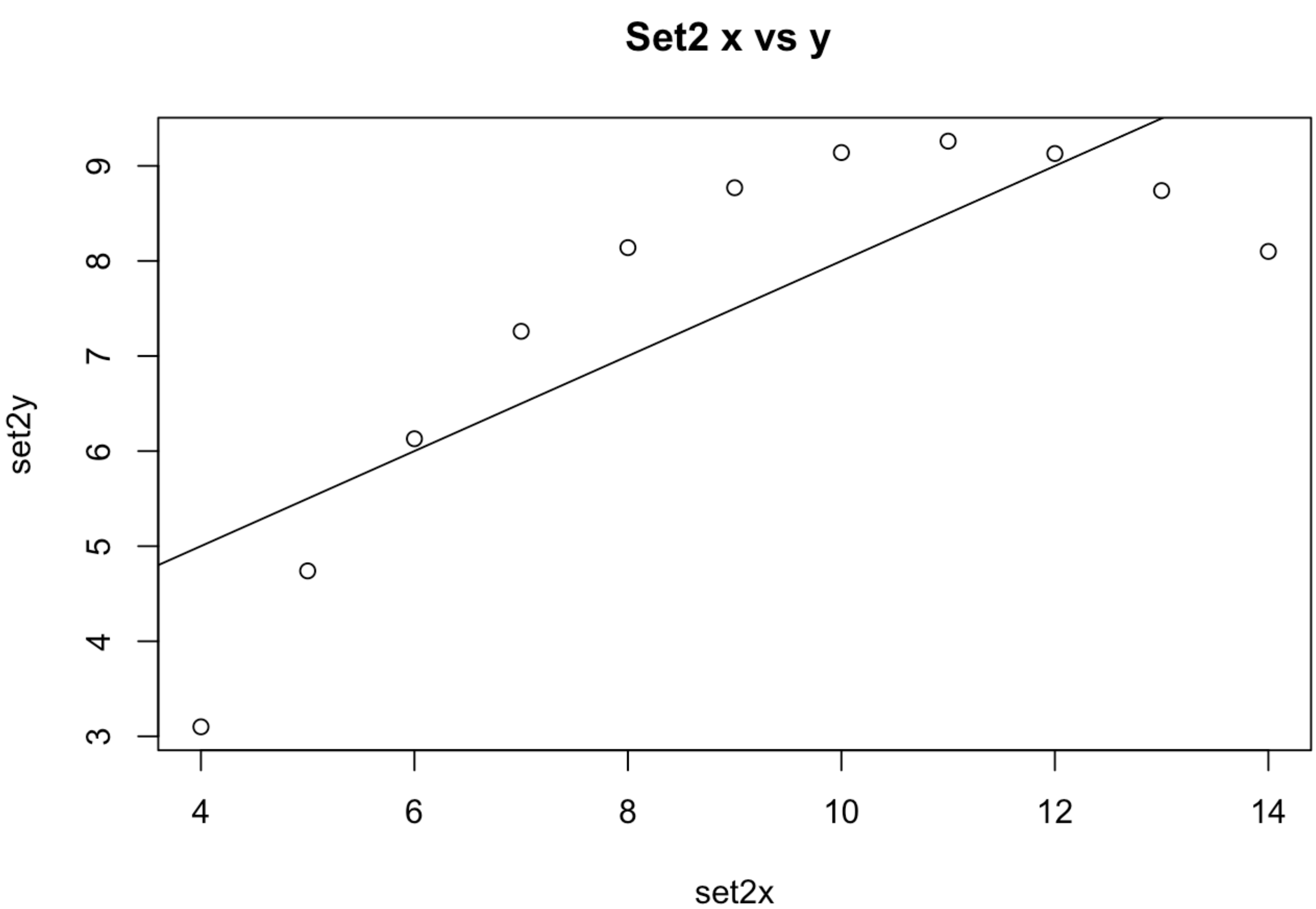
R_square: 0.6665425

Correlation corfficient: 0.8164205

We can conclude that there is an linear model appropriate for set1, there is 66.65% of variation in y is explained by variation of x.

b):

```
#plot the 1x 1y
plot(anscombe$Set2x,anscombe$Set2y,xlab = "set2x",ylab = "set2y",main = "Set2 x vs y
")
#get the linear model of set 1
set2_lm<-lm(anscombe$Set2y~anscombe$Set2x)
abline(set2_lm)
```




```
#summary the model
summary(set2_lm)
```

```
##
## Call:
## lm(formula = anscombe$Set2y ~ anscombe$Set2x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9009 -0.7609  0.1291  0.9491  1.2691
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.001      1.125   2.667  0.02576 *
## anscombe$Set2x    0.500      0.118   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 9 degrees of freedom
## Multiple R-squared:  0.6662, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002179
```

```
#anova of the model
anova(set2_lm)
```

```
## Analysis of Variance Table
##
## Response: anscombe$Set2y
##                Df Sum Sq Mean Sq F value   Pr(>F)
## anscombe$Set2x  1 27.500 27.5000  17.966 0.002179 **
## Residuals       9 13.776  1.5307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#correlation
cor(anscombe$Set2x,anscombe$Set2y)
```

```
## [1] 0.8162365
```

```
#r square
cor(anscombe$Set2x,anscombe$Set2y)^2
```

```
## [1] 0.666242
```

Fitted line y=3.001+0.500x

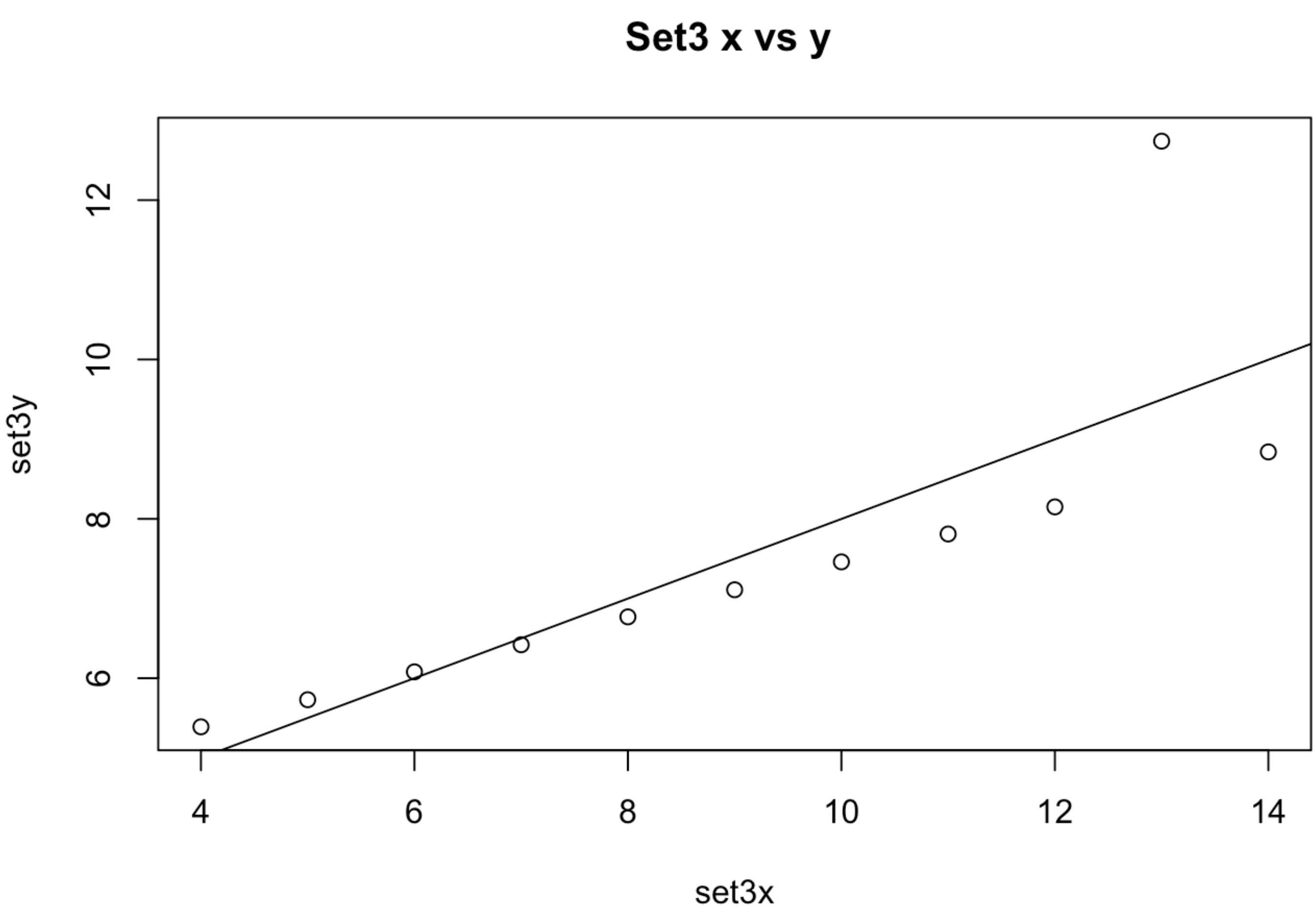
R_square: 0.666242

Correlation corfficient: 0.8162365

There is 66.62% of variation in y is explained by variation of x but we can clearly see from the graph, there is not looks like an linear relationship between x and y, the quadratic model might more suitable for the plot.

c):

```
#plot the 1x 1y
plot(anscombe$Set3x,anscombe$Set3y,xlab = "set3x",ylab = "set3y",main = "Set3 x vs y
")
#get the linear model of set 1
set3_lm<-lm(anscombe$Set3y~anscombe$Set3x)
abline(set3_lm)
```




```
#summary the model
summary(set3_lm)
```

```
##
## Call:
## lm(formula = anscombe$Set3y ~ anscombe$Set3x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1586 -0.6146 -0.2303  0.1540  3.2411
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.0025     1.1245   2.670  0.02562 *
## anscombe$Set3x   0.4997     0.1179   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176
```

```
#anova of the model
anova(set3_lm)
```

```
## Analysis of Variance Table
##
## Response: anscombe$Set3y
##                Df Sum Sq Mean Sq F value   Pr(>F)
## anscombe$Set3x  1 27.470 27.4700  17.972 0.002176 **
## Residuals       9 13.756  1.5285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#correlation
cor(anscombe$Set3x,anscombe$Set3y)
```

```
## [1] 0.8162867
```

```
#r square
cor(anscombe$Set3x,anscombe$Set3y)^2
```

```
## [1] 0.666324
```

Comment: Fitted line y=3.0025+0.4997x

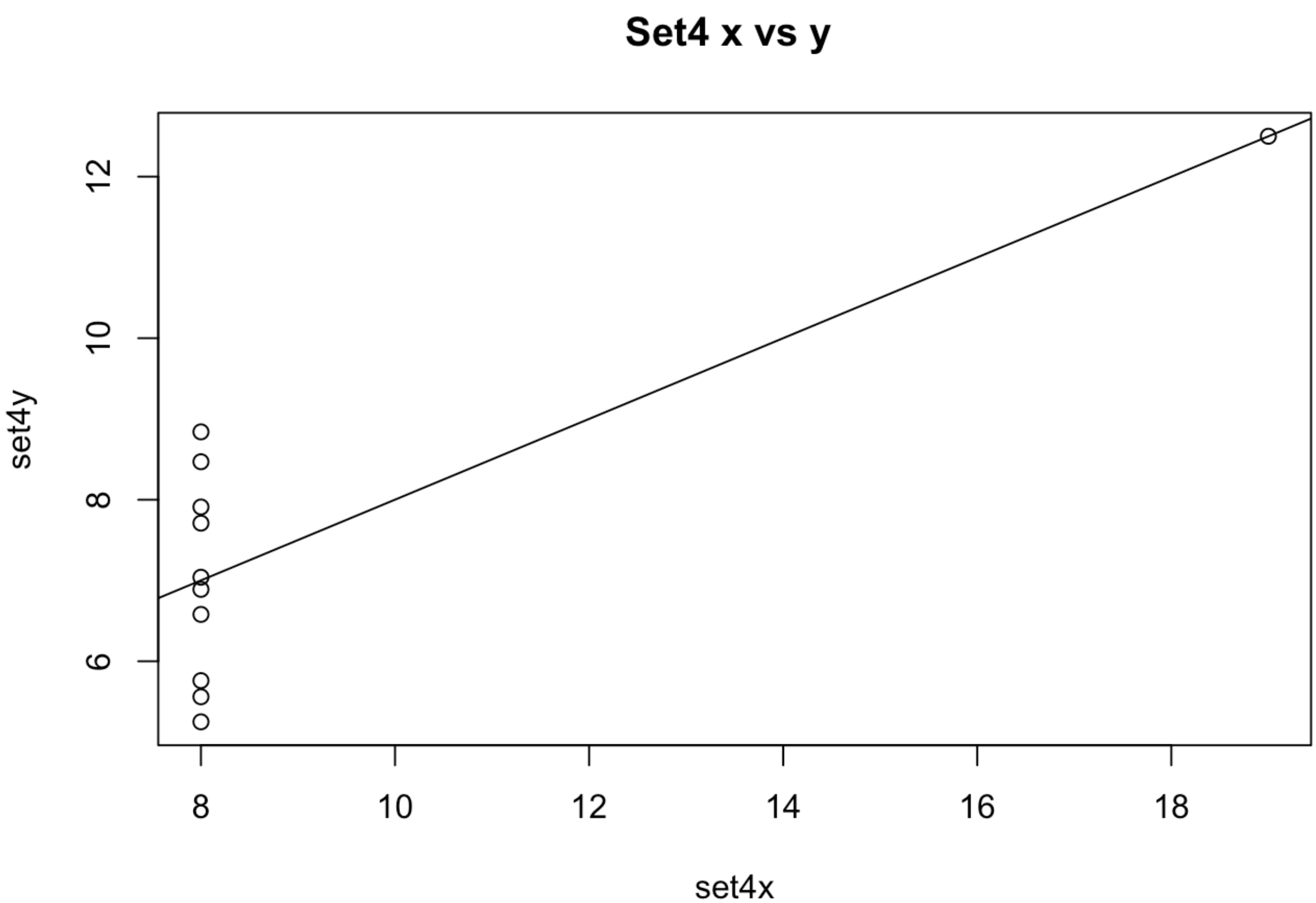
R_square: 0.666324

Correlation corfficient: 0.8162867

We can conclude that there is an linear model appropriate for set3, there is 66.63% of variation in y is explained by variation of x.

d):

```
#plot the 1x 1y
plot(anscombe$Set4x,anscombe$Set4y,xlab = "set4x",ylab = "set4y",main = "Set4 x vs y
")
#get the linear model of set 4
set4_lm<-lm(anscombe$Set4y~anscombe$Set4x)
abline(set4_lm)
```




```
#summary the model
summary(set4_lm)
```

```
##
## Call:
## lm(formula = anscombe$Set4y ~ anscombe$Set4x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.751 -0.831  0.000  0.809  1.839
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.0017     1.1239   2.671  0.02559 *
## anscombe$Set4x   0.4999     0.1178   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.236 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

```
#anova of the model
anova(set4_lm)
```

```
## Analysis of Variance Table
##
## Response: anscombe$Set4y
##                Df Sum Sq Mean Sq F value   Pr(>F)
## anscombe$Set4x  1 27.490 27.4900  18.003 0.002165 **
## Residuals       9 13.742  1.5269
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#correlation
cor(anscombe$Set4x,anscombe$Set4y)
```

```
## [1] 0.8165214
```

```
#r square
cor(anscombe$Set4x,anscombe$Set4y)^2
```

```
## [1] 0.6667073
```

Comment: Fitted line y=3.0017+0.4999x

R_square: 0.6667073

Correlation corfficient: 0.8165214

There is 66.67% of variation in y is explained by variation of x but we can clearly see from the graph, there is not looks like an linear relationship between x and y.

Problem6 (On hand)

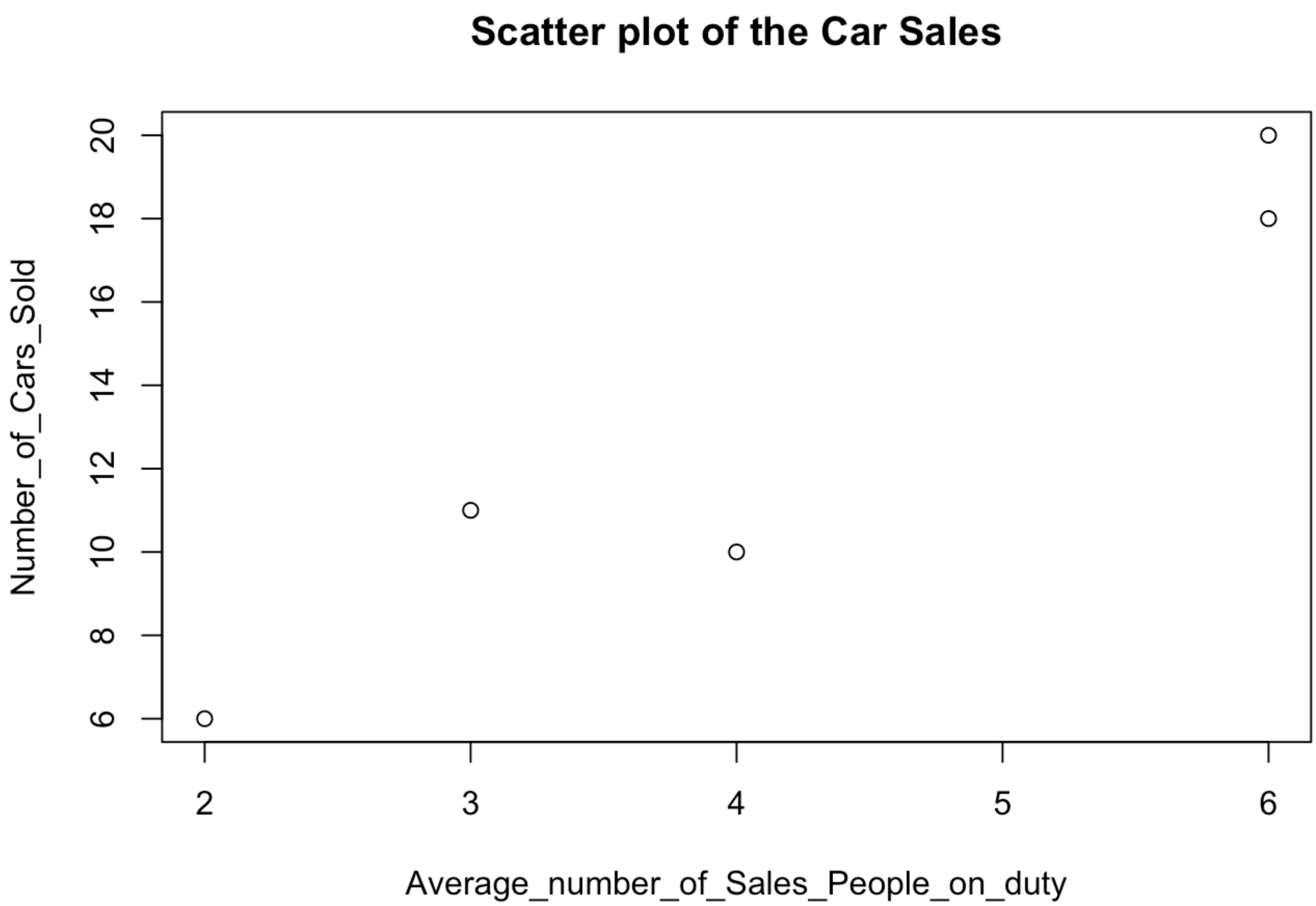
Problem7

```
#As there is no data provide from course space, I create a csv file.
cars<- read.csv("Cars.csv", header = T)
cars
```

```
##        Week_of Number_of_Cars_Sold Average_number_of_Sales_People_on_duty
## 1 January 30th                  20                                      6
## 2    june 29th                  18                                      6
## 3    March 2nd                  10                                      4
## 4 October 26th                   6                                      2
## 5 February 7th                  11                                      3
```

a):

```
plot(cars$Average_number_of_Sales_People_on_duty,cars$Number_of_Cars_Sold,xlab = "Ave
rage_number_of_Sales_People_on_duty",ylab = "Number_of_Cars_Sold",main = "Scatter plo
t of the Car Sales")
```



b):

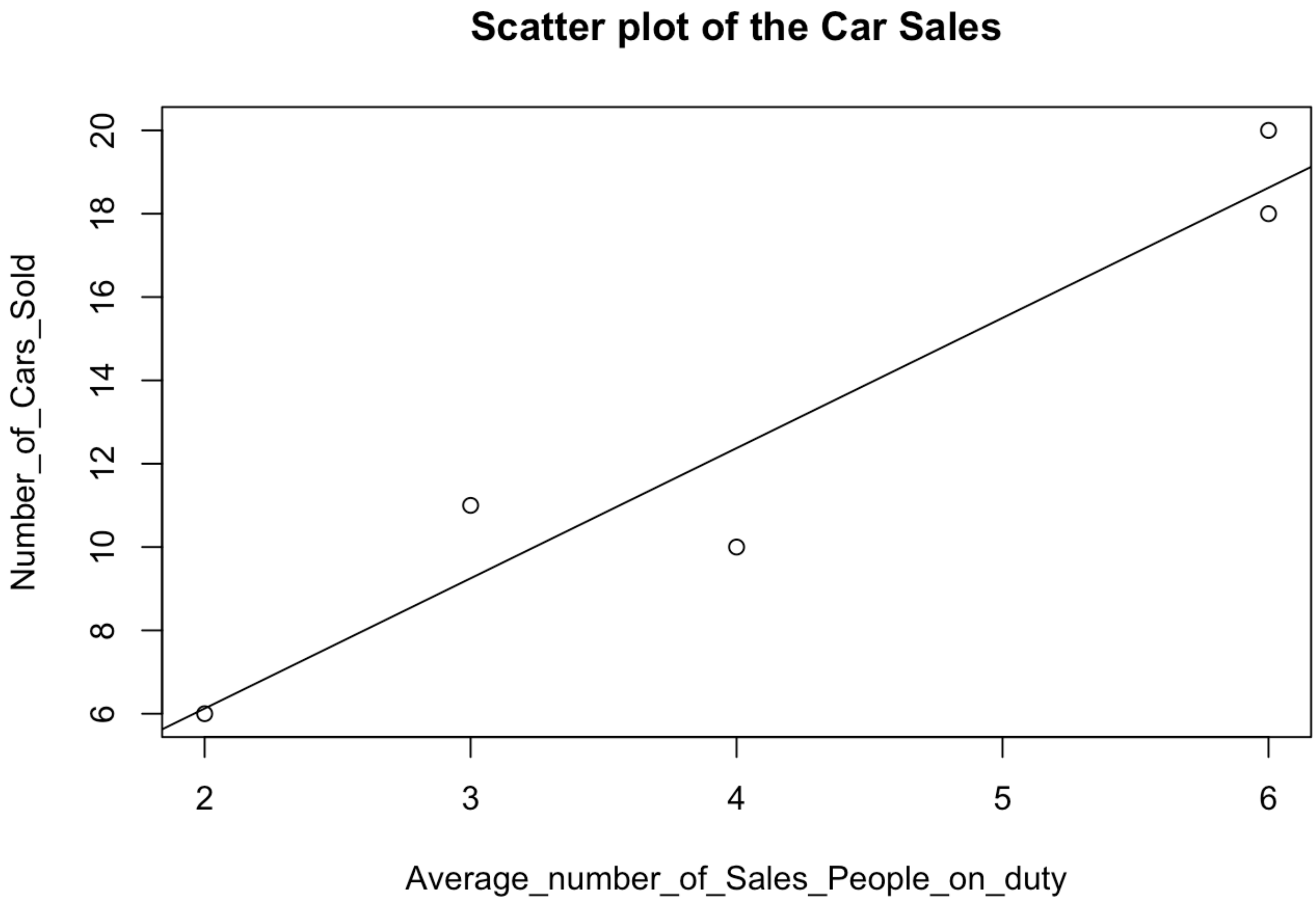
```
y<-cars$Number_of_Cars_Sold
x<-cars$Average_number_of_Sales_People_on_duty
cars_lm<-lm(y~x)
summary(cars_lm)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      1      2      3      4      5
##  1.375 -0.625 -2.375 -0.125  1.750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1250     2.4055  -0.052    0.962
## x             3.1250     0.5352   5.839    0.010 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.915 on 3 degrees of freedom
## Multiple R-squared:  0.9191, Adjusted R-squared:  0.8922
## F-statistic: 34.09 on 1 and 3 DF,  p-value: 0.01001
```

Comment: The estimate Y intercept is -0.1250, and the slope of the line is 3.1250

c):

```
plot(cars$Average_number_of_Sales_People_on_duty,cars$Number_of_Cars_Sold,xlab = "Ave
rage_number_of_Sales_People_on_duty",ylab = "Number_of_Cars_Sold",main = "Scatter plo
t of the Car Sales")
abline(cars_lm)
```



d):

```
predict(cars_lm, newdata = data.frame(x=5))
```

```
##    1
## 15.5
```

Comment: According to the output, we can say that approximately 15 cars (15.5 cars so round to 15) should the dealer expected to sell in a week

e.

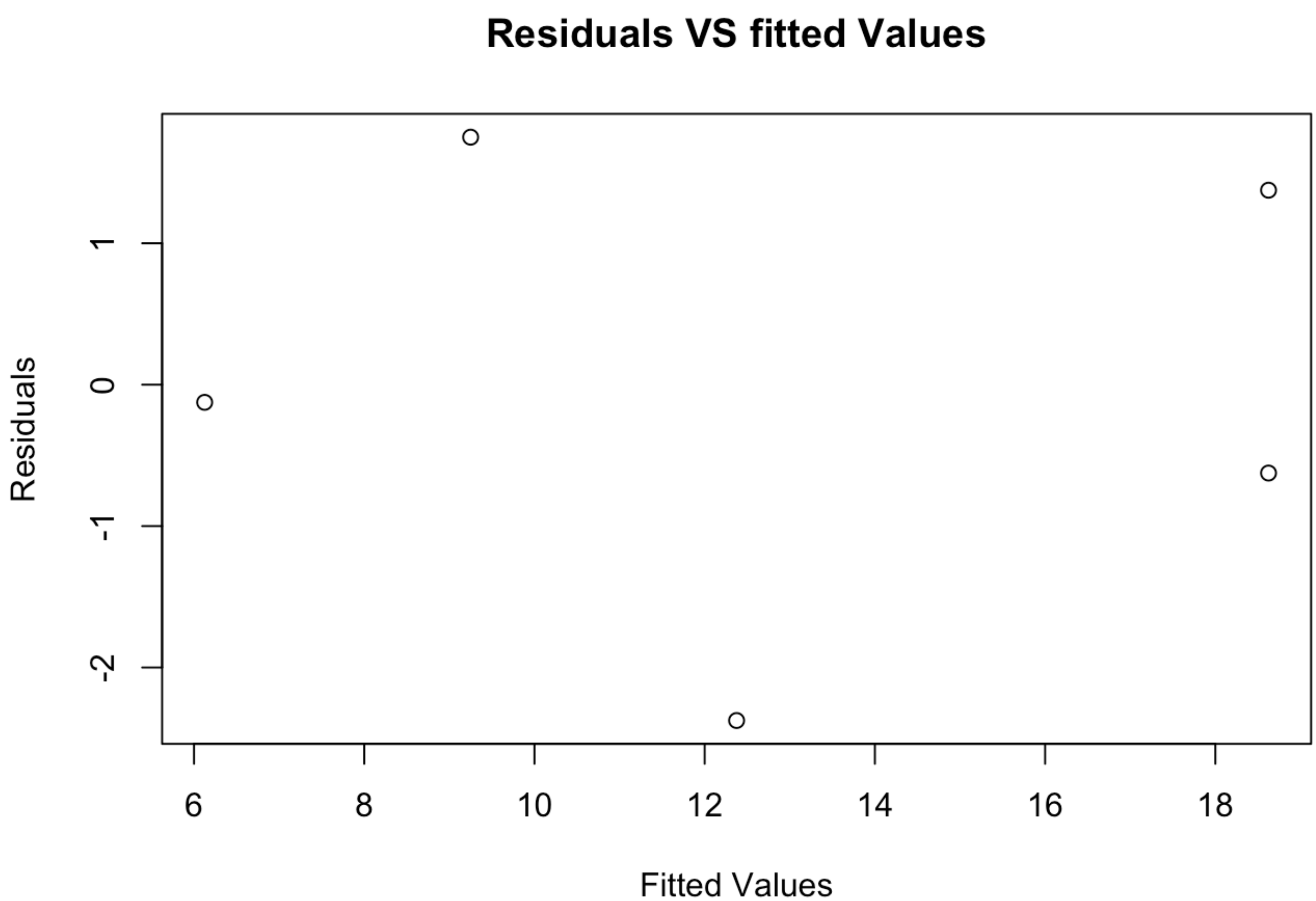
```
cars_lm$fitted.values
```

```
##      1      2      3      4      5
## 18.625 18.625 12.375  6.125  9.250
```

```
cars_lm$residuals
```

```
##      1      2      3      4      5
##  1.375 -0.625 -2.375 -0.125  1.750
```

```
plot(cars_lm$fitted.values,cars_lm$residuals,xlab = "Fitted Values", ylab = "Residual
s",main = "Residuals VS fitted Values")
```



Comment: Residuals are above or blowe the zero with no clearly indicates, There is one residuals taht is -2.375 which are the most far away point from the zero, this point might not satisfied with the fit. However, there are only five data point in the plot and the R_square value are 0.9191 which is very large, thus we can kindly say the fit is satisfied.

f):

```
anova(cars_lm)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## x          1    125 125.000  34.091 0.01001 *
## Residuals  3     11   3.667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments: sigama_Square = 3.667

g):

```
confint(cars_lm)
```

```
##                  2.5 %   97.5 %
## (Intercept) -7.780393 7.530393
## x            1.421697 4.828303
```

Comment: The confidents interval for Beta1 is (1.421697, 4.828303) which do not include 0, so the beta1 is significant.

h):

Since the R_Square is 0.9191, which indicates that 91.91% of the variation in the number of cars sales can be explained by variation in number of sales people on duty. Therefore we can conclude that there exist a linear relationship between those two variables.

i): The data period is from January 30th to June 29th, it do not have enough data to preict the data for next year. So i will not use this model to determine the number of sales people to have on duty next year.

---

proplem 8

As I can search from the web, some applet are very useful in the real problem application. It is very easy to see how the regression inetercept and slope changes when we adding or removing the points. Some applet can see the change in R_square dirctly, for example, if a data has a outlier that would reduce the r_square and r_square will increase when we remove it. It is very useful to help us to see the fitted model for the problems.