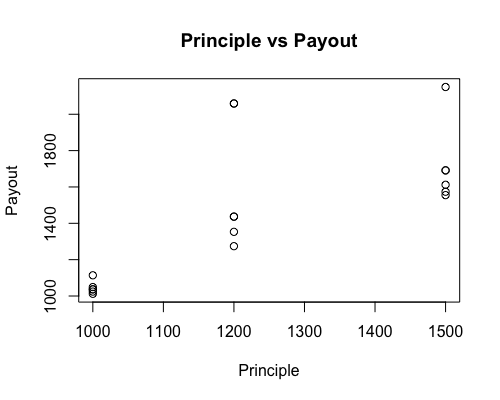
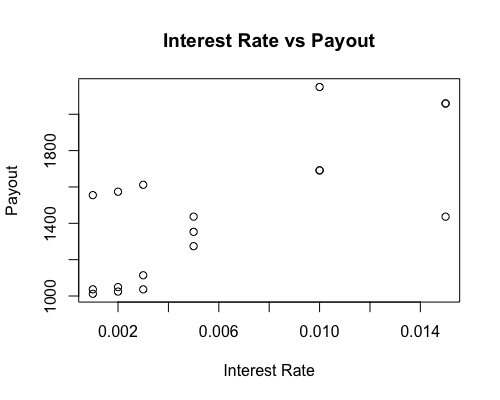
**Question 2 (1.2)**

Scatter plot of principle vs payout:



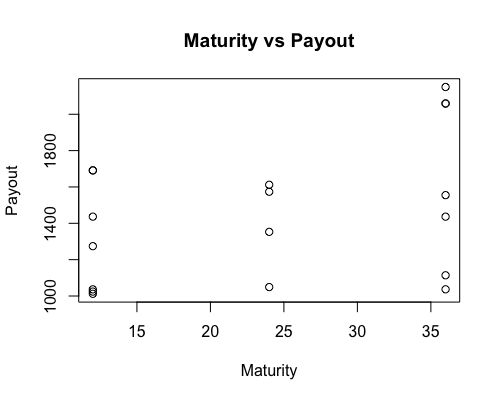
**plot(data$Prin,data$Payout,xlab='Principle',ylab='Payout',main='Principle vs Payout')**

Scatter plot of interest rate vs payout:



**plot(data$Int,data$Payout,xlab='Interest Rate',ylab='Payout',main='Interest Rate vs Payout')**

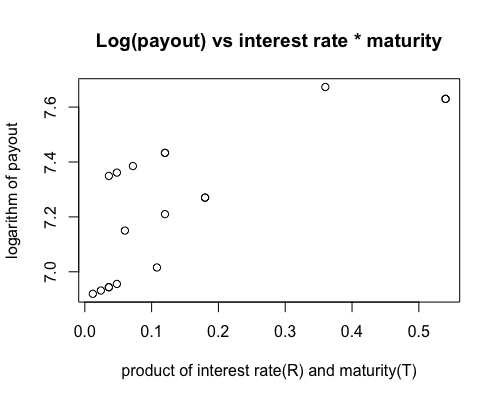
Scatter plot of maturity vs payout:



**plot(data$Time,data$Payout,xlab='Maturity',ylab='Payout',main='Maturity vs Payout')**

Despite the first plot (Principle vs Payout) looks like they have linear positive relationship, it is hard to see the relationship between payout and interest rate, maturity from the second and the third plots.

Then plot the logarithm of payout against the product of interest rate and maturity.



**plot(data$Int\*data$Time,log(data$Payout),xlab = 'product of interest rate(R) and maturity(T)',ylab='logarithm of payout',main = "Log(payout) vs interest rate \* maturity" )**

Conclusion: The scatter plot shows the linear relationship between the logarithm of payout and the product of interest rate and maturity. We can see when the product of interest rate and maturity increases, the payout also increases, which shows the positive relationship. At the same time, when the principle increases, the payout also increases. We feel confidence to our conclusion because the equation of the model is Payout = Pe^RT.

**Question 5 (2.3)**

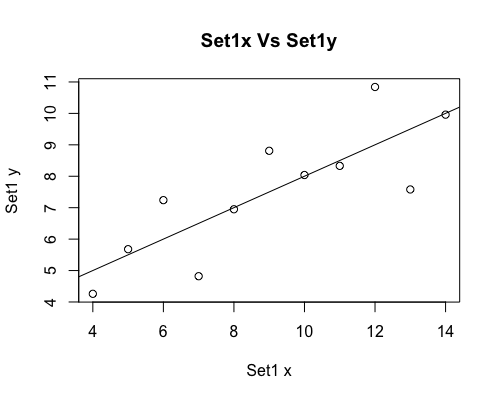
Set1:

**data<-read.table('anscombe.txt', header=T)**

**plot(data$Set1x,data$Set1y,xlab='Set1 x',ylab='Set1 y',main='Set1x Vs Set1y')**

**set1xyfit <- lm(data$Set1y~ data$Set1x)**

**abline(set1xyfit)**



**>summary(set1xyfit)**

Call:

lm(formula = data$Set1y ~ data$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 \*

data$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(set1xyfit)**

Analysis of Variance Table

Response: data$Set1y

Df Sum Sq Mean Sq F value Pr(>F)

data$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

**> cor(data$Set1x,data$Set1y)**

[1] 0.8164205

**> summary(set1xyfit)$r.squared**

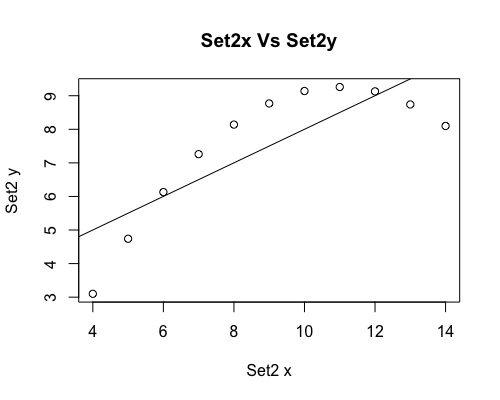
[1] 0.6665425

Set2:

**plot(data$Set2x,data$Set2y,xlab='Set2 x',ylab='Set2 y',main='Set2x Vs Set2y')**

**set2xyfit <- lm(data$Set2y~ data$Set2x)**

**abline(set2xyfit)**

****

**> summary(set2xyfit)**

Call:

lm(formula = data$Set2y ~ data$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 \*

data$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(set2xyfit)**

Analysis of Variance Table

Response: data$Set2y

Df Sum Sq Mean Sq F value Pr(>F)

data$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

**> cor(data$Set2x,data$Set2y)**

[1] 0.8162365

**> summary(set2xyfit)$r.squared**

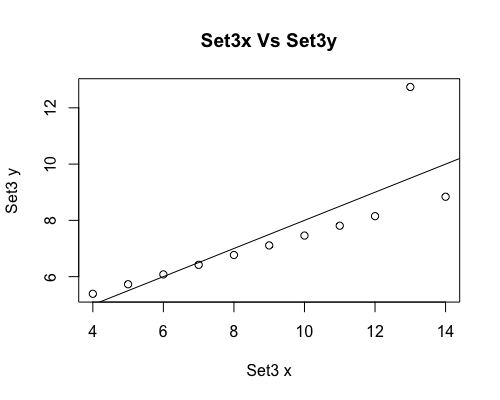
[1] 0.666242

Set3:

**plot(data$Set3x,data$Set3y,xlab='Set3 x',ylab='Set3 y',main='Set3x Vs Set3y')**

**set3xyfit <- lm(data$Set3y~ data$Set3x)**

**abline(set3xyfit)**

****

**> summary(set3xyfit)**

Call:

lm(formula = data$Set3y ~ data$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 \*

data$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(set3xyfit)**

Analysis of Variance Table

Response: data$Set3y

Df Sum Sq Mean Sq F value Pr(>F)

data$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

**> cor(data$Set3x,data$Set3y)**

[1] 0.8162867

**> summary(set3xyfit)$r.squared**

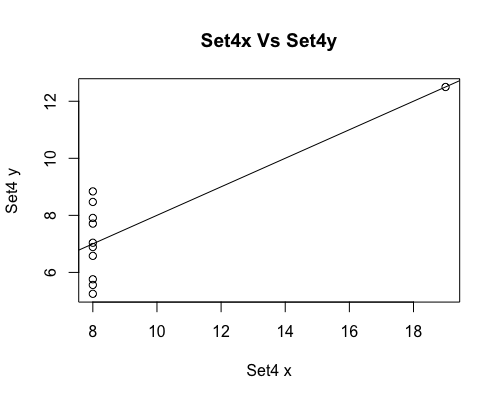
[1] 0.666324

Set4:

**plot(data$Set4x,data$Set4y,xlab='Set4 x',ylab='Set4 y',main='Set4x Vs Set4y')**

**set4xyfit <- lm(data$Set4y~ data$Set4x)**

**abline(set4xyfit)**



**> summary(set4xyfit)**

Call:

lm(formula = data$Set4y ~ data$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 \*

data$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(set4xyfit)**

Analysis of Variance Table

Response: data$Set4y

Df Sum Sq Mean Sq F value Pr(>F)

data$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

**> cor(data$Set4x,data$Set4y)**

[1] 0.8165214

**> summary(set4xyfit)$r.squared**

[1] 0.6667073

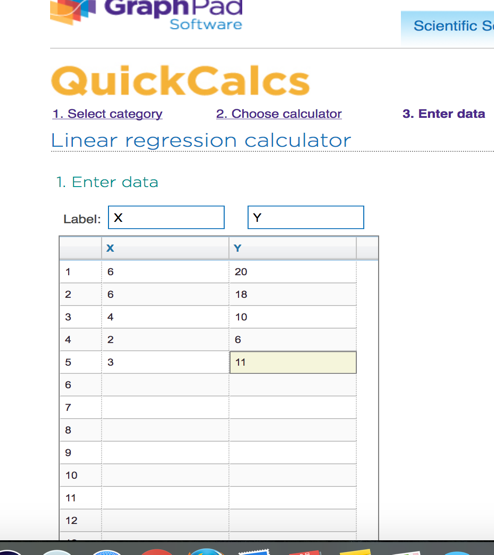
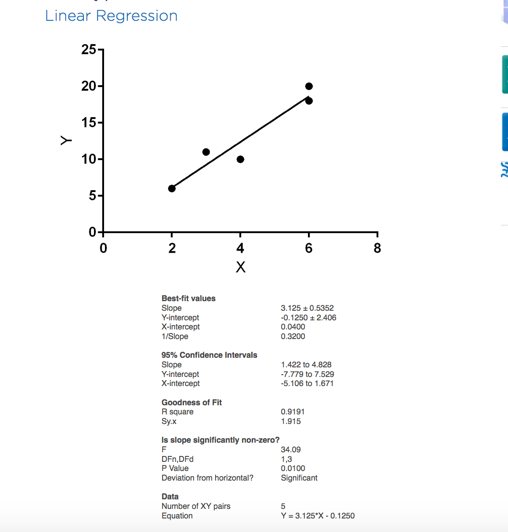
Comments:

According to the images of scatter plots and the regressions lines on each set, we find that the linear regression of y on x in set1 is appropriate. The regression equation is mostly the same in all sets. The regression lines are best fitting in set1. However, in set2, set3 and set4, the regression line does not fit the data appropriately. The residual is large except in set1. The correlation coefficient of both sets is approximately 0.816, which means they both have strong positive relationship between two variables.

**Question 8 (2.28):**

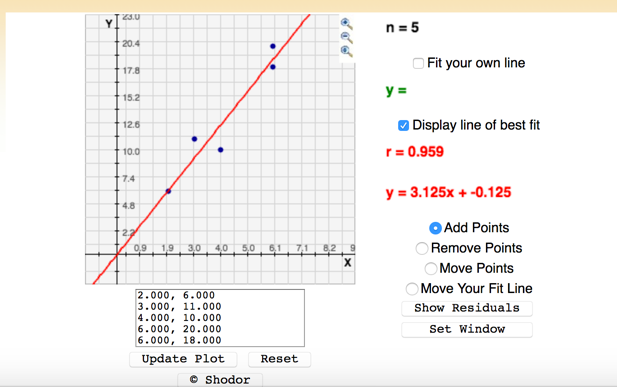
I searched the Google for regression applets, which gave me a lot of results. Here I chose two of them as an example.

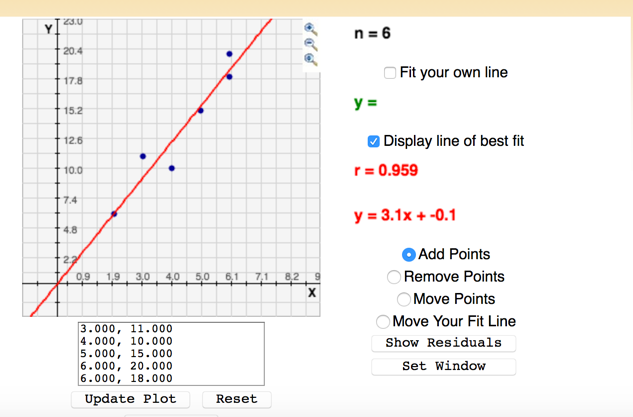
The first one is <https://www.graphpad.com/quickcalcs/linear1/>. It is a beautiful and easy understanding online regression tools. I used the data of exercise 2.4 as an example. It was easy for me to type in the data. Then I could compute the regression model immediately, which was straightforward.

The results provided me all the information I needed such as slope, 95% Confidence Interval etc.

The second one is <http://www.shodor.org/interactivate/activities/Regression/>. One of the advantages is allowing me to add more points and they illustrate the effect of such changes. For example, the first screen shot is the same data I used in the first example. Then I add a point like (5.00,16.00), it will change the regression line right away.





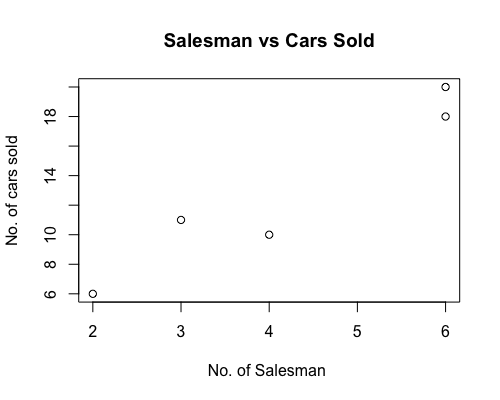
In conclusion, the online regression tools are very useful and convenient when we are doing regression and analyzing the data. But the online tools also have limitation, when the data sets are huge, it is not efficient to type in the data one by one. Using R studio with input data files will be more convenient.

**Question 7 (2.5)**

a)

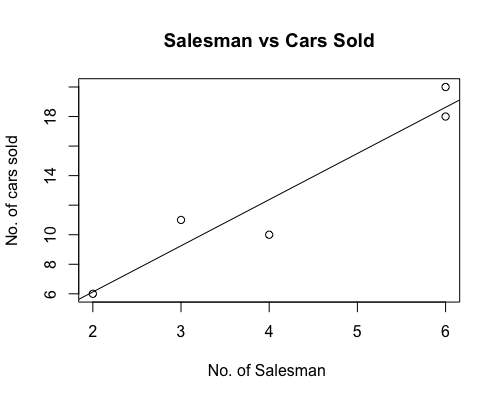
**> data<-read.table('CarAndSales.txt', header=T)**

**> plot(data$NumPeo, data$NumCars,xlab='No. of Salesman ',ylab='No. of cars sold',main ='Salesman vs Cars Sold')**

****

b) c) **> xyfit <- lm(data$NumCars~ data$NumPeo)**

**> abline(xyfit)**



**> summary(xyfit)**

Call:

lm(formula = data$NumCars ~ data$NumPeo)

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

data$NumPeo 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

**Estimate of the slope is = 3.1250**

**Estimate of the intercept is = -0.1250**

**The fitted line is: y = 3.125x – 0.125**

f)

**> anova(xyfit)**

Analysis of Variance Table

Response: data$NumCars

Df Sum Sq Mean Sq F value Pr(>F)

data$NumPeo 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

**Estimate of σ^2 is 3.667**

g)

**> confint(xyfit,'data$NumPeo', level = 0.95)**

2.5 % 97.5 %

data$NumPeo 1.421697 4.828303

**95% Confidence interval for β1 is (1.421697, 4.828303)**