2.3

$$\frac{\alpha Q}{\alpha \beta} = \sum_{i=1}^{n} 2wi(yi - \beta xi)\,(-xi) = 0$$

$$\sum_{i=1}^{n} wixi(\beta xi - yi) = 0$$

$$\sum_{i=1}^{n} wi\beta xi^2 = \sum_{i=1}^{n} wixiyi$$

$$\beta = \sum_{i=1}^{n} \frac{\sum_{i=1}^{n} wixiyi}{\sum_{i=1}^{n} wixi^2}$$

2.8

If r = 0.25
y1 – 69 = 0.25(x1 – 68) y1 = 70
y2 – 69 = 0.25(x2 – 68) y2 = 68

if r = 0.75
y1 – 69 = 0.75(x1 – 68) y1 = 72
y2 – 69 = 0.75(x2 – 68) y2 = 66

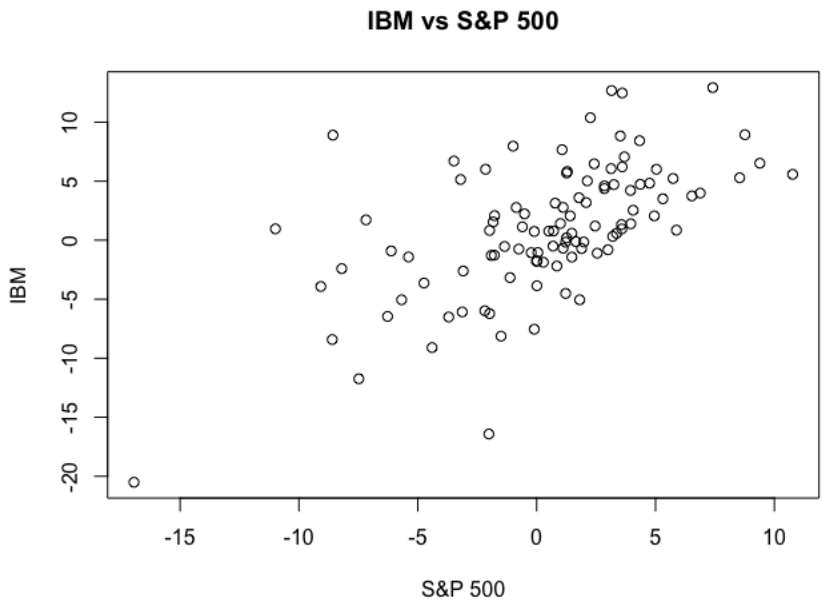
r	**0.25**	**0.5**	**0.75**
y1	70	71	72
y2	68	67	66

Conclusion: With r increasing, the degree of regression to mean decreases.

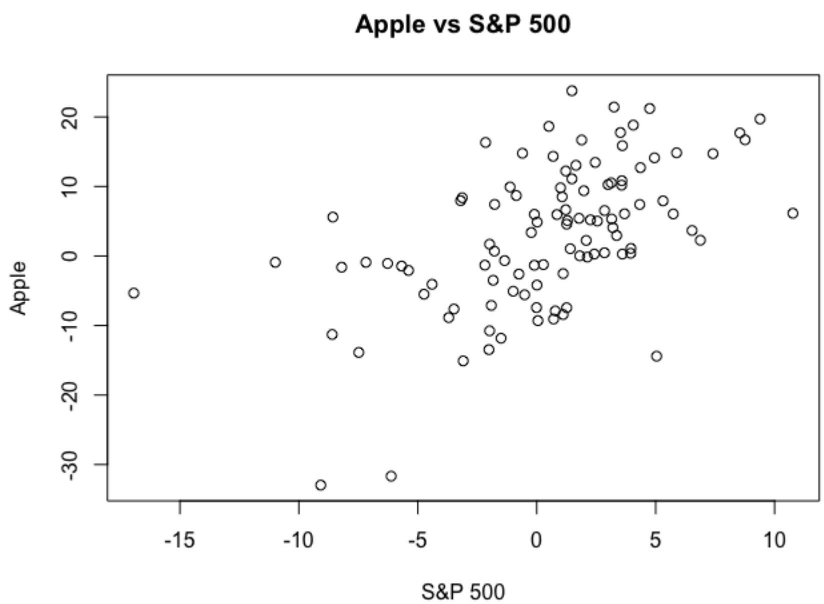
2.9

a)

```
#2.9
#a)
#read file from csv
return = read.csv("Desktop/Predictive Analytics/Assignment/1/IBM-Apple-SP500 RR Data.csv",skip = 1)

#turn the type of "date" from factor to numeric
sp = as.numeric(str_replace(return$S.P.500,"%",""))
ibm = as.numeric(str_replace(return$IBM,"%",""))
apple = as.numeric(str_replace(return$Apple,"%",""))

#plot
plot(sp,ibm,xlab = 'S&P 500',ylab = 'IBM',main = 'IBM vs S&P 500')
plot(sp,apple,xlab = 'S&P 500',ylab = 'Apple',main = 'Apple vs S&P 500')

```



Comment: The return rate between IBM and S&P 500 are positively linear correlated. Return rates mostly concentrate between 0 and 5%.



Comment: The return rate between Apple and S&P 500 are positively linear correlated. Return Rates mostly concentrate between 0 and 10%

<u>b)</u>

```
#b)
#regression for ibm
fit1=lm(ibm~sp)
summary(fit1)

```

```
Call:
lm(formula = ibm ~ sp)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.5646  -2.4261  -0.6636   2.2188  14.6414 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.64164    0.44136   1.454    0.149    
sp           0.74481    0.09898   7.525 2.15e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.478 on 102 degrees of freedom
Multiple R-squared:  0.357,	Adjusted R-squared:  0.3507 
F-statistic: 56.63 on 1 and 102 DF,  p-value: 2.15e-11
```

β for IBM and S&P 500 is 0.74481

```
#regression for apple
fit2=lm(apple~sp)
summary(fit2)
```

```
Call:
lm(formula = apple ~ sp)

Residuals:
     Min       1Q   Median       3Q      Max 
-26.5378  -5.9191   0.4677   5.5363  19.4413 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.4863     0.8606   2.889  0.00472 ** 
sp            1.2449     0.1930   6.450  3.8e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.732 on 102 degrees of freedom
Multiple R-squared:  0.2897,	Adjusted R-squared:  0.2827 
F-statistic:  41.6 on 1 and 102 DF,  p-value: 3.799e-09
```

β for Apple and S&P 500 is 1.2449

Apple had a higher expected return relative to S&P 500 because the β Apple is larger than the β of IBM.

c)

```
#c)
#need to be divided by 100 since they are percentages
sd(sp)
sd(ibm)
sd(apple)
```

```
> sd(sp)
[1] 4.457853
> sd(ibm)
[1] 5.557105
> sd(apple)
[1] 10.3104
```

```
#calculate the correlation matrix
install.packages("corrplot")
source("http://www.sthda.com/upload/rquery_cormat.r")

corre <- data.frame("sp" = sp,"ibm" = ibm,"apple" = apple)
rquery.cormat(corre)
corrplot 0.84 loaded
$r
      apple  sp ibm
apple     1
sp     0.54   1
ibm    0.41 0.6   1

$p
        apple      sp ibm
apple       0
sp    3.8e-09       0
ibm   1.2e-05 2.2e-11   0

$sym
      apple sp ibm
apple 1
sp    .     1
ibm   .     .  1
attr(,"legend")
[1] 0 ‘ ’ 0.3 ‘.’ 0.6 ‘,’ 0.8 ‘+’ 0.9 ‘*’ 0.95 ‘B’ 1
```

```
#calculate b
cor(ibm,sp)*sd(ibm)/sd(sp)
cor(apple,sp)*sd(apple)/sd(sp)
> cor(ibm,sp)*sd(ibm)/sd(sp)
[1] 0.7448088
> cor(apple,sp)*sd(apple)/sd(sp)
[1] 1.244856
```

d)

$\beta = r * Sy/Sx$

With same Sx and similar r, If Sy is larger, then β is larger, which means high return. Sy represents the volatility of dependent variable. Therefore, the higher the volatility, the higher the return.

2.10

a)

```r
#2.10
#a)
price = read.csv("Desktop/Predictive Analytics/Assignment/1/Steak+Prices.csv")


chuck = as.numeric(str_replace(price$Chuck.Price,"\\$",""))
porthse = as.numeric(str_replace(price$PortHse.Price,"\\$",""))
ribeye = as.numeric(str_replace(price$RibEye.Price,"\\$",""))

fit1=lm(log(price$Chuck.Qty)~log(chuck))
summary(fit1)

fit2=lm(log(price$PortHse.Qty)~log(porthse))
summary(fit2)

fit3=lm(log(price$RibEye.Qty)~log(ribeye))
summary(fit3)
```

```
> fit1=lm(log(price$Chuck.Qty)~log(chuck))
> summary(fit1)

Call:
lm(formula = log(price$Chuck.Qty) ~ log(chuck))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.32463 -0.12036 -0.01714  0.09430  0.49725 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.8899     0.2871  20.513  < 2e-16 ***
log(chuck)   -1.3687     0.3199  -4.278 9.44e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1812 on 46 degrees of freedom
Multiple R-squared:  0.2846,	Adjusted R-squared:  0.2691 
F-statistic:  18.3 on 1 and 46 DF,  p-value: 9.441e-05

> 
> fit2=lm(log(price$PortHse.Qty)~log(porthse))
> summary(fit2)

Call:
lm(formula = log(price$PortHse.Qty) ~ log(porthse))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.57655 -0.23544  0.00317  0.23511  0.49991 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    9.1123     0.5136  17.742  < 2e-16 ***
log(porthse)  -2.6565     0.2752  -9.654 1.23e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.283 on 46 degrees of freedom
Multiple R-squared:  0.6695,	Adjusted R-squared:  0.6624 
F-statistic:  93.2 on 1 and 46 DF,  p-value: 1.233e-12
```

```
>
> fit3=lm(log(price$RibEye.Qty)~log(ribeye))
> summary(fit3)

Call:
lm(formula = log(price$RibEye.Qty) ~ log(ribeye))

Residuals:
     Min       1Q   Median       3Q      Max
-0.54075 -0.21801  0.03995  0.20328  0.70950

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.6627     0.7537  10.167 2.39e-13 ***
log(ribeye)  -1.4460     0.3731  -3.876 0.000335 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2943 on 46 degrees of freedom
Multiple R-squared:  0.2462,	Adjusted R-squared:  0.2298
F-statistic: 15.02 on 1 and 46 DF,  p-value: 0.0003352
```

The absolute values of correlation coefficients are in following orders:
Chuck < ribeye < porthse
If chuck is the cheapest beef, then its price elasticity should be the most volatile one. But it is not. So, the price elasticities are not in the expected order.

b)

$$\ln y1 = \alpha + \beta \ln x1$$
$$\ln y2 = \alpha + \beta \ln x2$$
$$ln\frac{y2}{y1} = \beta ln\frac{x2}{x1}$$
$$\frac{y2}{y1} = \left(\frac{x2}{x1}\right)^{\beta}$$
$$y2 = 1.1^{\beta}$$
$$\frac{y2 - y1}{y1} = 1.1^{\beta} - 1$$

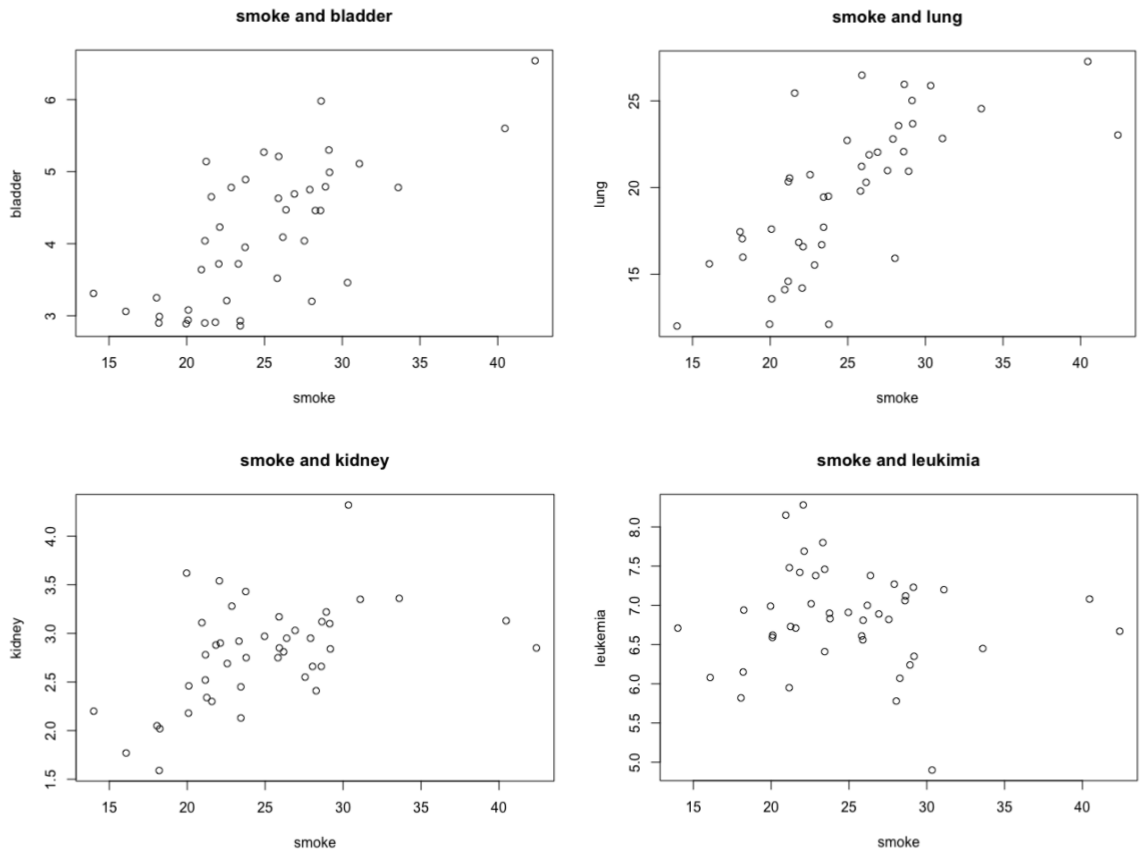
```
#b)
(1.1)^-1.3687 - 1 #demand change of chuck
(1.1)^-2.6565 - 1 #demand change of porthse
(1.1)^-1.4460 - 1 #demand change of ribeye
```

```
> (1.1)^-1.3687 - 1
[1] -0.1223005
> (1.1)^-2.6565 - 1
[1] -0.2236808
> (1.1)^-1.4460 - 1
[1] -0.1287432
>
```

2.11

a)

```
#2.11
#a)
cig = read.csv("Desktop/Predictive Analytics/Assignment/1/smoking-cancer.csv")
plot(cig$Smoke,cig$Bladder,xlab = 'smoke',ylab = 'bladder',main = 'smoke and bladder')
plot(cig$Smoke,cig$Lung,xlab = 'smoke',ylab = 'lung',main = 'smoke and lung')
plot(cig$Smoke,cig$Kidney,xlab = 'smoke',ylab = 'kidney',main = 'smoke and kidney')
plot(cig$Smoke,cig$Leukemia,xlab = 'smoke',ylab = 'leukemia',main = 'smoke and leukimia')
```



Smoking and deaths of bladder, lung have apparent linear relationships. Smoking and deaths of kidney have a tiny linear relationship. Smoking and leukemia have no linear relationships. Outliers exists in all of the charts.

b)

```
#b)
cor.test(cig$Bladder,cig$Smoke)
cor.test(cig$Lung,cig$Smoke)
cor.test(cig$Kidney,cig$Smoke)
cor.test(cig$Leukemia,cig$Smoke)
```

```
> cor.test(cig$Bladder,cig$Smoke)

        Pearson's product-moment correlation

data:  cig$Bladder and cig$Smoke
t = 6.4173, df = 42, p-value = 9.964e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5141412 0.8276195
sample estimates:
      cor
0.7036219

> cor.test(cig$Lung,cig$Smoke)

        Pearson's product-moment correlation

data:  cig$Lung and cig$Smoke
t = 6.3064, df = 42, p-value = 1.439e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5051008 0.8237330
sample estimates:
      cor
0.6974025

> cor.test(cig$Kidney,cig$Smoke)

        Pearson's product-moment correlation

data:  cig$Kidney and cig$Smoke
t = 3.6174, df = 42, p-value = 0.0007922
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2227387 0.6851336
sample estimates:
      cor
0.4873896

> cor.test(cig$Leukemia,cig$Smoke)

        Pearson's product-moment correlation

data:  cig$Leukemia and cig$Smoke
t = -0.44485, df = 42, p-value = 0.6587
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3580815  0.2331390
sample estimates:
        cor
-0.06848123
```

From the results of correlation test, it can be seen that the p-value of the leukemia is larger than 0.5, which means there is not enough evidence to prove that smoke has relationships with leukemia. For the rest of the three, test of bladder has the largest correlation coefficient. So, deaths of bladder is mostly significant to smoking.