In [ ]:
```
Group Member:(gac69, has175, dif24, yil210, lil131)
```

1. Data Process

We have collected several datasets, includes cocoa_consumption, alcohol_consumption, fish_consumption, smoker_number, we want to see the relation of those attributes with the nobel prize.

In [2]:
```
setwd("/Users/chengaoxiang/Desktop/18_Fall/Data_Analytics/Assignment/Assignment5")
```

In [3]:
```
getwd()
```

'/Users/chengaoxiang/Desktop/18_Fall/Data_Analytics/Assignment/Assignment5'

In [4]:
```
collection_data_ori <- read.csv("summary.csv", header = TRUE)
```

In [5]:
```
head(collection_data_ori)
```

Entity	Laureates10million	cocoaconsump2010	alcoholconsump2015	fishconsump2013	Smoker_Num2012
Algeria	0.476	0.575	0.6	3.92	3123101
Argentina	1.119	0.785	7.6	7.05	5987695
Australia	4.844	2.874	12.6	26.09	2961263
Austria	23.995	3.800	8.5	13.88	2109044
Azerbaijan	1.008	NA	2.1	2.13	1622189
Bangladesh	0.060	NA	0.2	19.21	24013742

In [6]:
```
summary(collection_data_ori)
```

```
       Entity   Laureates10million cocoaconsump2010 alcoholconsump2015
 Algeria   : 1   Min.   :  0.0470   Min.   :0.027    Min.   : 0.100    
 Argentina : 1   1st Qu.:  0.3407   1st Qu.:0.785    1st Qu.: 5.250    
 Australia : 1   Median :  2.0795   Median :1.792    Median : 9.050    
 Austria   : 1   Mean   :  7.6136   Mean   :2.055    Mean   : 8.096    
 Azerbaijan: 1   3rd Qu.:  8.6252   3rd Qu.:3.022    3rd Qu.:11.200    
 Bangladesh: 1   Max.   :111.3170   Max.   :5.883    Max.   :17.100    
 (Other)   :64                      NA's   :29                         
 fishconsump2013  Smoker_Num2012   
 Min.   : 1.290   Min.   :    14130  
 1st Qu.: 7.438   1st Qu.:   901824  
 Median :19.095   Median :  2911105  
 Mean   :20.461   Mean   : 11418155  
 3rd Qu.:25.003   3rd Qu.:  9565081  
 Max.   :91.920   Max.   :281714540  
```

In [7]:
```
collection_data <- na.omit(collection_data_ori)
```

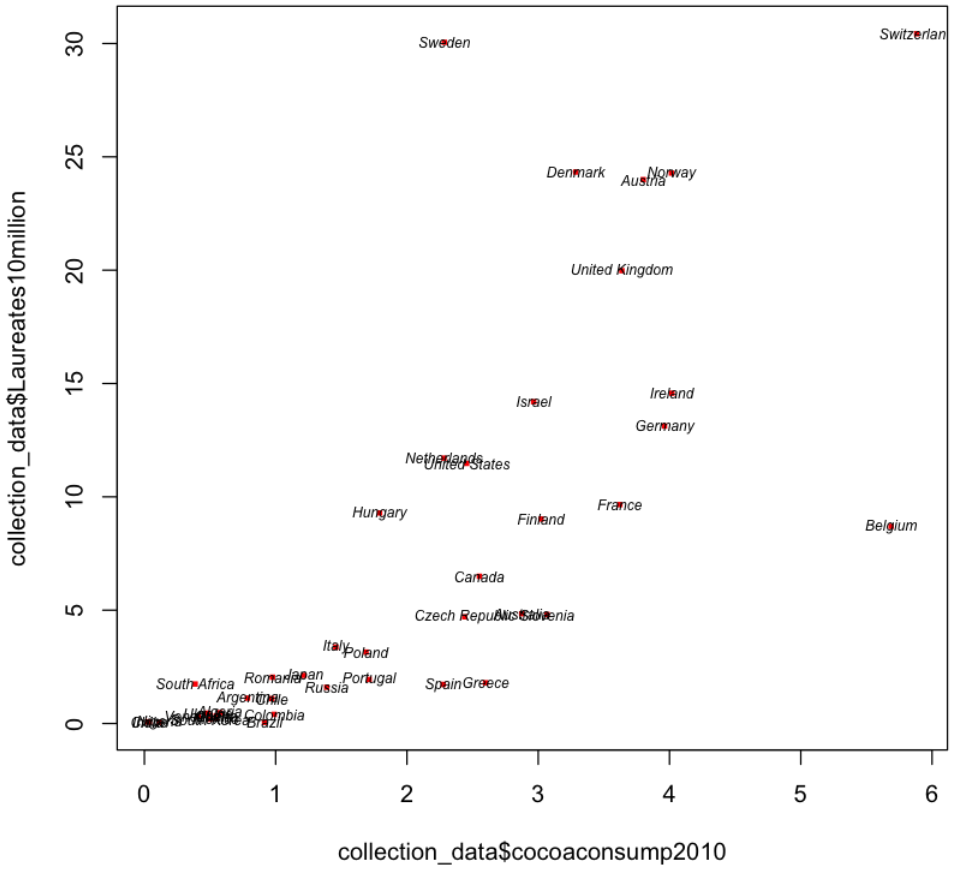
In [8]:
```
summary(collection_data)
```

```
       Entity   Laureates10million cocoaconsump2010 alcoholconsump2015
 Algeria  : 1   Min.   : 0.047     Min.   :0.027    Min.   : 0.600    
 Argentina: 1   1st Qu.: 0.476     1st Qu.:0.785    1st Qu.: 7.600    
 Australia: 1   Median : 3.149     Median :1.792    Median :10.300    
 Austria  : 1   Mean   : 7.317     Mean   :2.055    Mean   : 9.534    
 Belgium  : 1   3rd Qu.:11.476     3rd Qu.:3.022    3rd Qu.:11.500    
 Brazil   : 1   Max.   :30.431     Max.   :5.883    Max.   :14.500    
 (Other)  :35                                                         
 fishconsump2013 Smoker_Num2012   
 Min.   : 3.92   Min.   :   393481  
 1st Qu.:10.48   1st Qu.:  2109044  
 Median :20.76   Median :  3874289  
 Mean   :21.38   Mean   : 16557207  
 3rd Qu.:26.09   3rd Qu.: 10355707  
 Max.   :53.76   Max.   :281714540  
```

2. Data Analysis

2.1 Relationship between Cocoa_consumption and Nobel

In [9]:
```
plot(collection_data$cocoaconsump2010,collection_data$Laureates10million,col='red',cex=0.5,lty=1,pch=15)
text(collection_data$cocoaconsump2010,collection_data$Laureates10million,collection_data$Entity,cex=0.6, font=3)
```

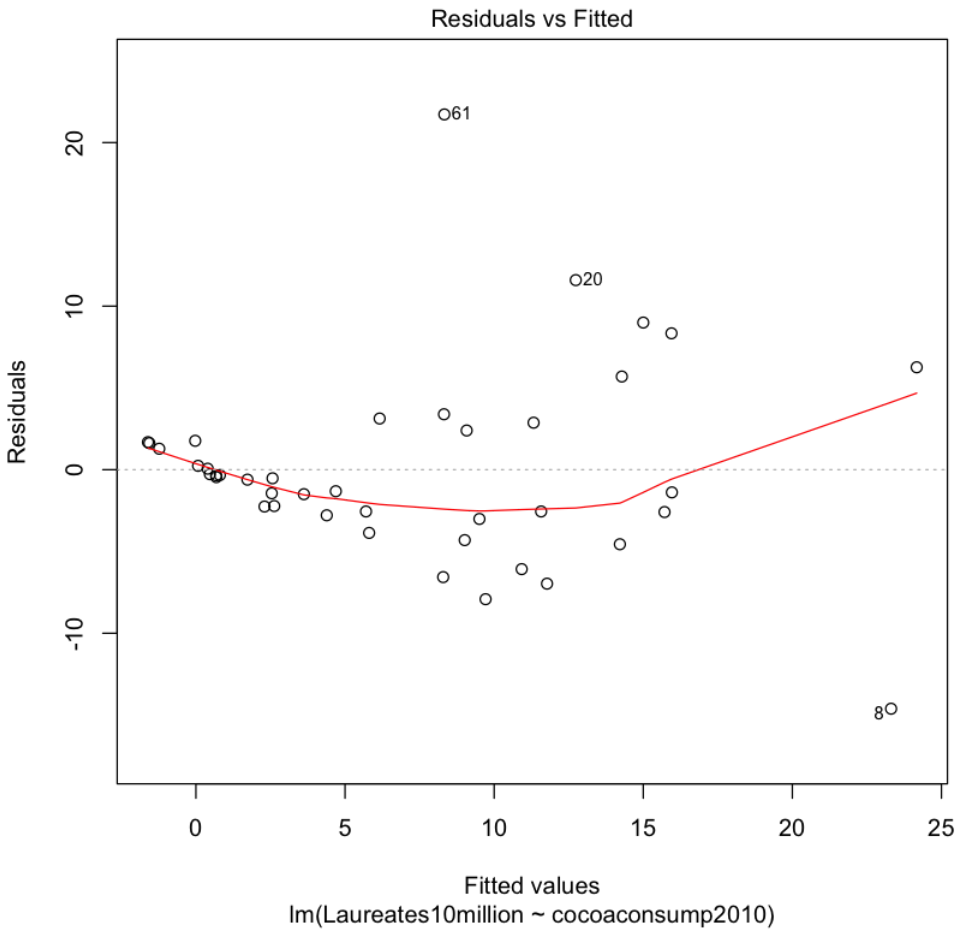


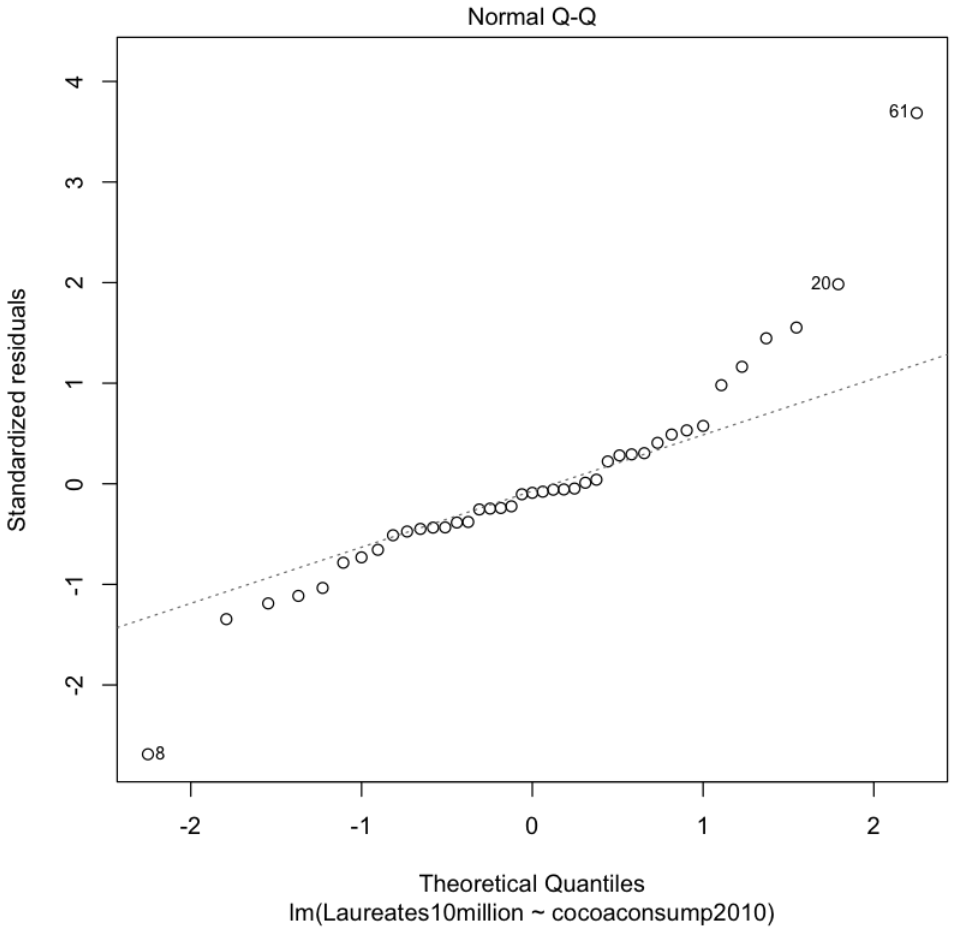
In [10]:
```
lm.cocoa_nobel <- lm(Laureates10million ~ cocoaconsump2010, data = collection_data)
```

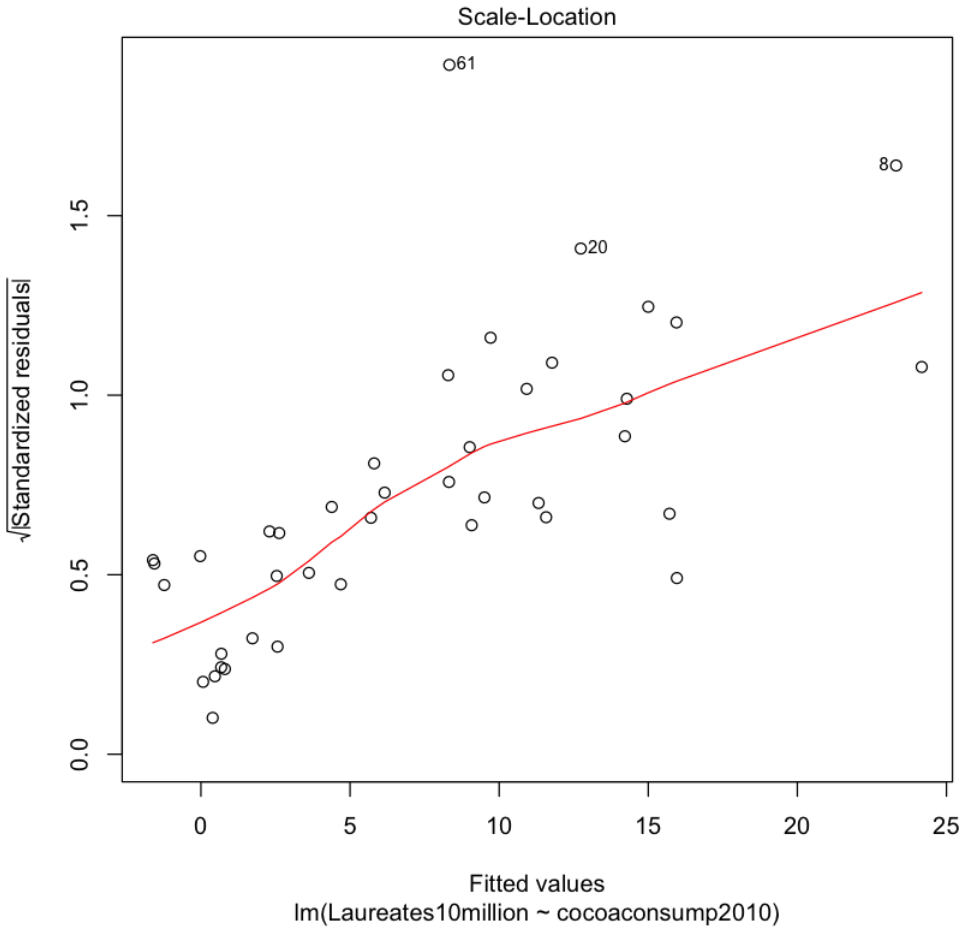
In [11]:
```
summary(lm.cocoa_nobel)
```

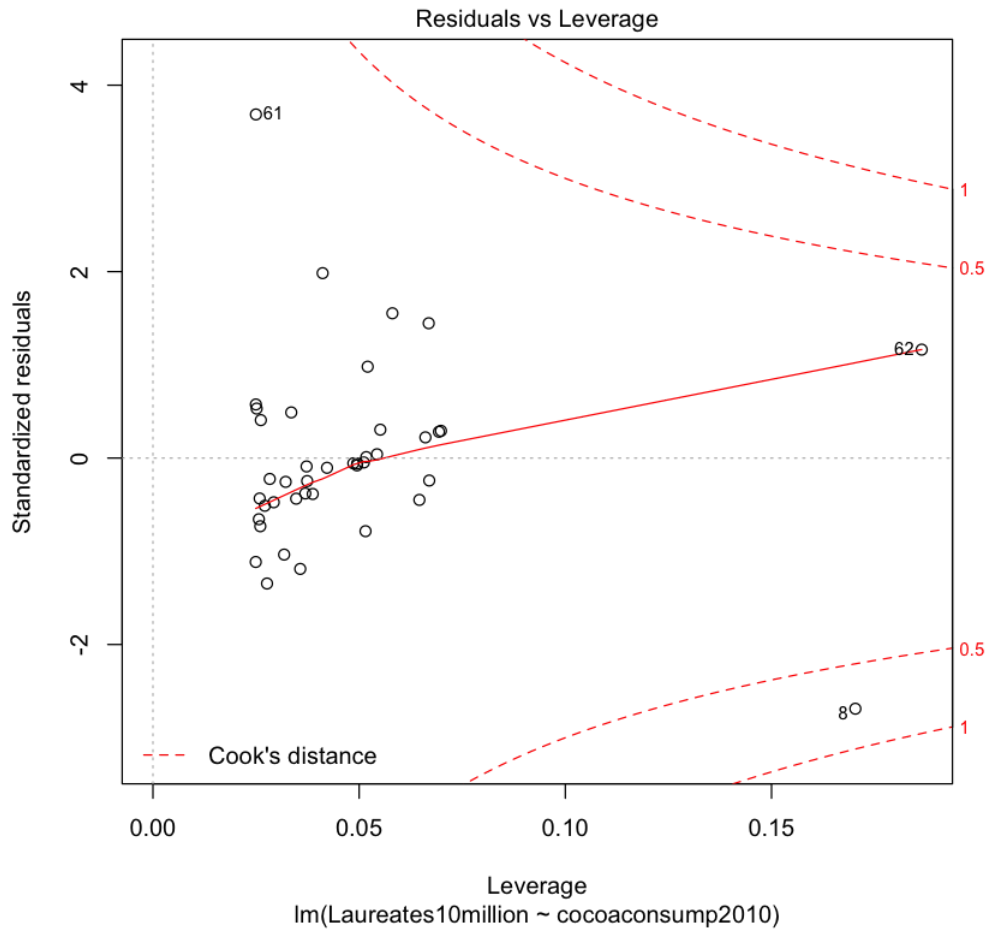
```
Call:
lm(formula = Laureates10million ~ cocoaconsump2010, data = collection_data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6144  -2.5895  -0.5259   1.7659  21.7208

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.7275     1.5907  -1.086    0.284
cocoaconsump2010   4.4021     0.6274   7.016 2.04e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.967 on 39 degrees of freedom
Multiple R-squared:  0.558,     Adjusted R-squared:  0.5466
F-statistic: 49.23 on 1 and 39 DF,  p-value: 2.037e-08
```

In [12]:
```
plot(lm.cocoa_nobel)
```


Residuals vs Fitted
61
20
8
Residuals
20
10
0
-10
0
5
10
15
20
25
Fitted values
lm(Laureates10million ~ cocoaconsump2010)

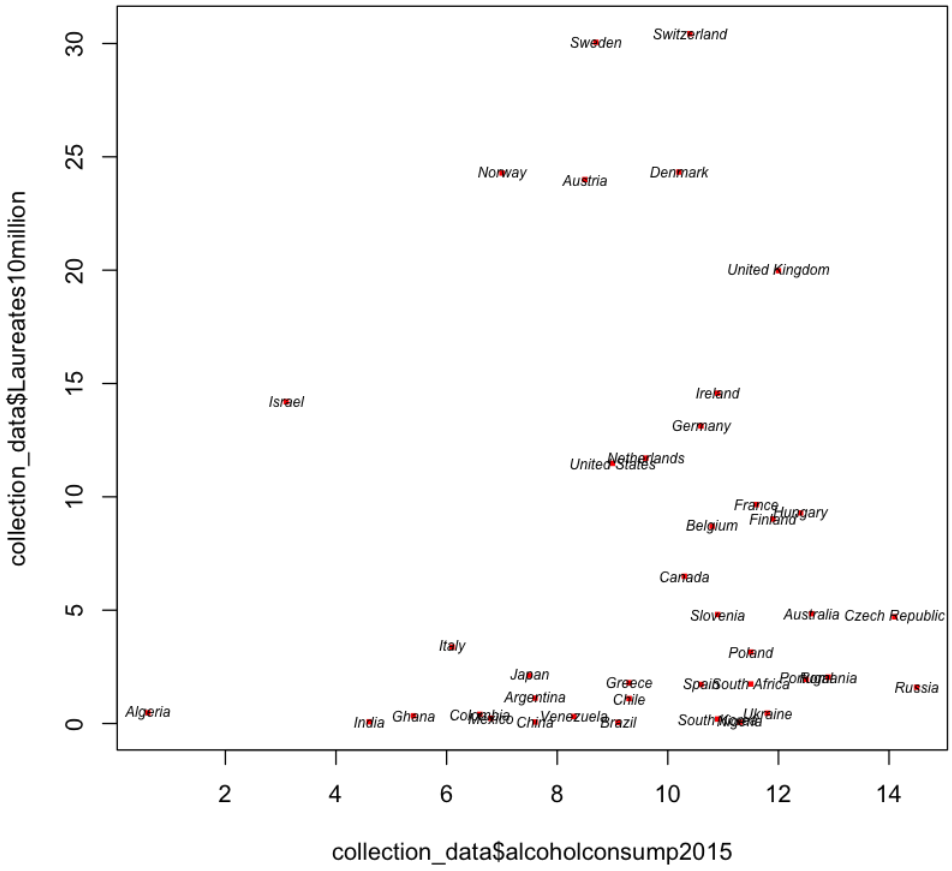

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Laureates10million ~ cocoaconsump2010)
61
20
8
-2
-1
0
1
2
3
4


Scale-Location
61
8
20
√|Standardized residuals|
0.0
0.5
1.0
1.5
0
5
10
15
20
25
Fitted values
lm(Laureates10million ~ cocoaconsump2010)



2.2 Relationship between Alcohol_consumption and Nobel

In [13]:
```
plot(collection_data$fishconsump2013,collection_data$Laureates10million,col='red',cex=0.5,lty=1,pch=15)
text(collection_data$fishconsump2013,collection_data$Laureates10million,collection_data$Entity,cex=0.6, font=3)
```



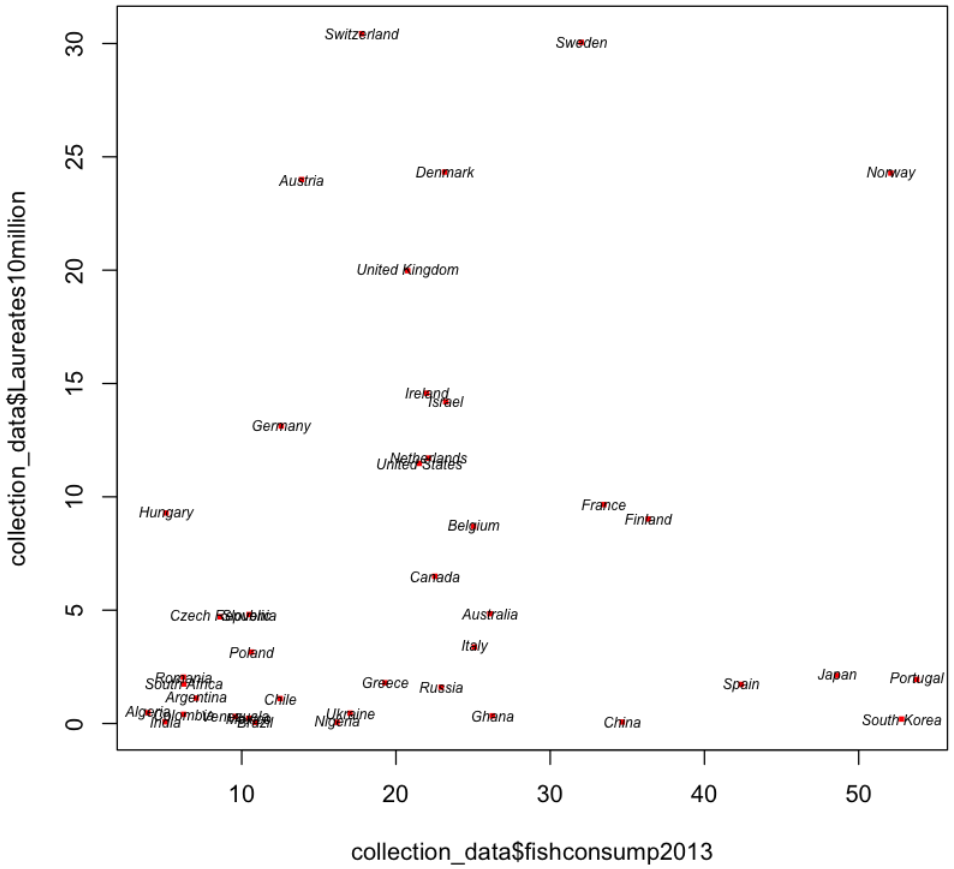
In [14]:
```
lm.alcohol_nobel <- lm(Laureates10million ~ alcoholconsump2015, data = collection_data)
```

In [15]:
```
summary(lm.alcohol_nobel)
```

```
Call:
lm(formula = Laureates10million ~ alcoholconsump2015, data = collection_data)

Residuals:
   Min     1Q Median     3Q    Max
-7.530 -6.360 -4.462  4.239 22.984

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.8910     4.8493   1.215    0.232
alcoholconsump2015  0.1496     0.4870   0.307    0.760

Residual standard error: 8.964 on 39 degrees of freedom
Multiple R-squared:  0.002414,  Adjusted R-squared:  -0.02317
F-statistic: 0.09437 on 1 and 39 DF,  p-value: 0.7603
```

2.3 Relationship between Fish_consumption and Nobel

In [16]:
```
plot(collection_data$fishconsump2013,collection_data$Laureates10million,col='red',cex=0.5,lty=1,pch=15)
text(collection_data$fishconsump2013,collection_data$Laureates10million,collection_data$Entity,cex=0.6, font=3)
```



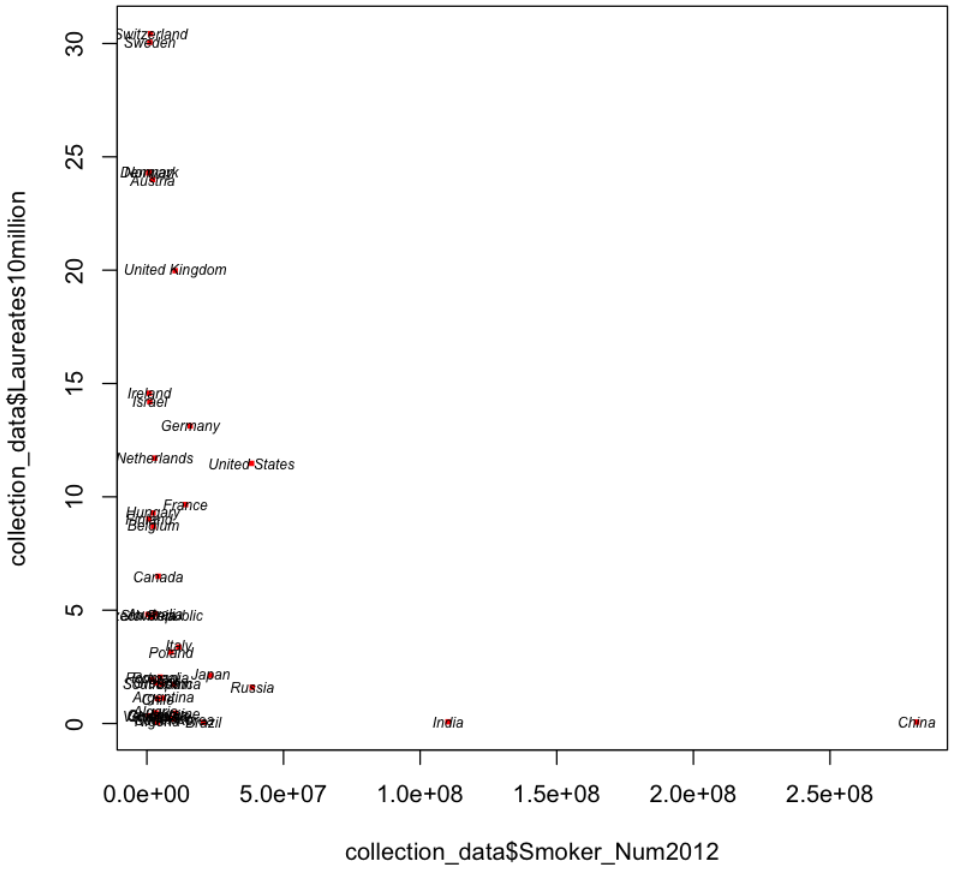
In [17]:
```
lm.fish_nobel <- lm(Laureates10million ~ fishconsump2013, data = collection_data)
```

In [18]:
```
summary(lm.fish_nobel)
```

```
Call:
lm(formula = Laureates10million ~ fishconsump2013, data = collection_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.486  -5.886  -3.651   4.144  23.499

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.0275     2.5630   1.962    0.057 .
fishconsump2013   0.1071     0.1010   1.061    0.295
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.848 on 39 degrees of freedom
Multiple R-squared:  0.02804,   Adjusted R-squared:  0.003121
F-statistic: 1.125 on 1 and 39 DF,  p-value: 0.2953
```

2.4 Relationship between Smoker and Nobel

In [19]:
```
plot(collection_data$Smoker_Num2012,collection_data$Laureates10million,col='red',cex=0.5,lty=1,pch=15)
text(collection_data$Smoker_Num2012,collection_data$Laureates10million,collection_data$Entity,cex=0.6, font=3)
```



In [20]:
```
lm.smoke_nobel <- lm(Laureates10million ~ Smoker_Num2012, data = collection_data)
```

In [21]:
```
summary(lm.smoke_nobel)
```

```
Call:
lm(formula = Laureates10million ~ Smoker_Num2012, data = collection_data)

Residuals:
   Min     1Q Median     3Q    Max
-7.771 -6.615 -3.549  3.849 22.513

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.970e+00  1.459e+00   5.463 2.88e-06 ***
Smoker_Num2012 -3.945e-08  3.002e-08  -1.314    0.197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.783 on 39 degrees of freedom
Multiple R-squared:  0.0424,    Adjusted R-squared:  0.01784
F-statistic: 1.727 on 1 and 39 DF,  p-value: 0.1965
```

2.5 Relationship between All_data and Nobel

In [22]:
```
lm.all_nobel <- lm(Laureates10million ~ ., data = collection_data[,2:6])
```

In [24]:
```
summary(lm.all_nobel)
```

```
Call:
lm(formula = Laureates10million ~ ., data = collection_data[,
    2:6])

Residuals:
    Min      1Q  Median      3Q     Max
-14.795  -2.821  -1.144   2.061  21.094

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.730e+00  3.616e+00   0.478    0.635
cocoaconsump2010    4.591e+00  7.008e-01   6.551 1.28e-07 ***
alcoholconsump2015 -4.507e-01  3.424e-01  -1.317    0.196
fishconsump2013     2.056e-02  7.155e-02   0.287    0.775
Smoker_Num2012      7.489e-10  2.217e-08   0.034    0.973
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.06 on 36 degrees of freedom
Multiple R-squared:  0.5792,    Adjusted R-squared:  0.5325
F-statistic: 12.39 on 4 and 36 DF,  p-value: 1.955e-06
```

2.5.1 Choose good attributes

In [25]:
```
final.lm <- step(lm.all_nobel)
```

```
Start:  AIC=152.41
Laureates10million ~ cocoaconsump2010 + alcoholconsump2015 +
    fishconsump2013 + Smoker_Num2012

                     Df Sum of Sq    RSS    AIC
- Smoker_Num2012      1      0.04 1322.0 150.41
- fishconsump2013     1      3.03 1325.0 150.50
- alcoholconsump2015  1     63.65 1385.6 152.33
<none>                             1322.0 152.41
- cocoaconsump2010    1   1576.08 2898.1 182.59

Step:  AIC=150.41
Laureates10million ~ cocoaconsump2010 + alcoholconsump2015 +
    fishconsump2013

                     Df Sum of Sq    RSS    AIC
- fishconsump2013     1      3.24 1325.3 148.51
- alcoholconsump2015  1     64.74 1386.8 150.37
<none>                             1322.0 150.41
- cocoaconsump2010    1   1727.99 3050.0 182.68

Step:  AIC=148.51
Laureates10million ~ cocoaconsump2010 + alcoholconsump2015

                     Df Sum of Sq    RSS    AIC
- alcoholconsump2015  1     63.48 1388.8 148.43
<none>                             1325.3 148.51
- cocoaconsump2010    1   1808.79 3134.1 181.80

Step:  AIC=148.43
Laureates10million ~ cocoaconsump2010

                   Df Sum of Sq    RSS    AIC
<none>                          1388.8 148.43
- cocoaconsump2010  1    1752.9 3141.6 179.90
```

In [38]:
```
summary(final.lm)
```

```
Call:
lm(formula = Laureates10million ~ cocoaconsump2010, data = collection_data[,
    2:6])

Residuals:
     Min       1Q   Median       3Q      Max
-14.6144  -2.5895  -0.5259   1.7659  21.7208

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.7275     1.5907  -1.086    0.284
cocoaconsump2010   4.4021     0.6274   7.016 2.04e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.967 on 39 degrees of freedom
Multiple R-squared:  0.558,     Adjusted R-squared:  0.5466
F-statistic: 49.23 on 1 and 39 DF,  p-value: 2.037e-08
```

2.6 Conclusion

So it seems that only cocoa consumption has good linear relation with nobel prize among all the attributes.