Swetha Adike

Venu Goud Raparti

MSIS 545

---

15. a

This problem data is based on a study to investigate reproductive strategies in plants. The data is recorded by the biologists at the time of pollinating species of lily, on the time spent at sources of pollen and the proportions of pollen removed by bumblebee queens and honeybee workers.

Let's head on to the data summary

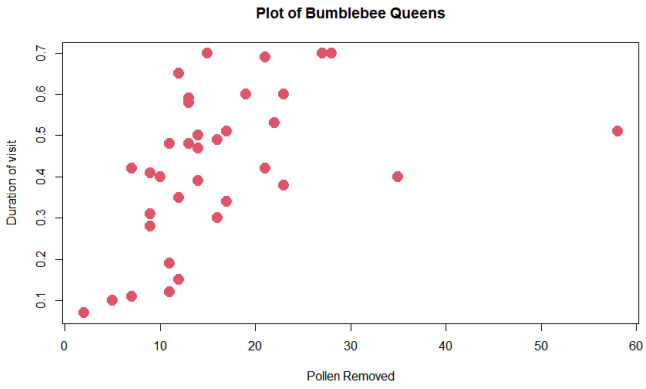
```
> summary(ex0327)
 PollenRemoved    DurationOfVisit   BeeType
 Min.   :0.0700   Min.   : 2.00   Queen :35
 1st Qu.:0.3600   1st Qu.:11.00   Worker:12
 Median :0.4900   Median :15.00
 Mean   :0.4874   Mean   :21.19
 3rd Qu.:0.6700   3rd Qu.:23.50
 Max.   :0.8900   Max.   :78.00
```

The above data shows that there are 35 bumblebee queen bees and 12 honeybee workers. The duration of visit column spread suggest a log transformation but let's see the analysis of data as it is

Let's plot the proportion of pollen removed against duration of visit for bumblebee queens, done by the below command

```
> plot(ex0327$DurationOfVisit[ex0327$BeeType == "Queen"],
+        ex0327$PollenRemoved[ex0327$BeeType == "Queen"], xlab = "Pollen Removed",
+        ylab = "Duration of visit", main = "Plot of Bumblebee Queens", col = 2, pch = 19, cex = 2)
```

And below is the scatterplot for the above command



b. Let's fit a simple linear regression of proportion of pollen removed on duration of visit by first taking the linear model command on these factors, by the below code

```
> pollen.lm <- lm( ex0327$PollenRemoved ~ ex0327$DurationOfVisit)
> pollen.lm
Call:
lm(formula = ex0327$PollenRemoved ~ ex0327$DurationOfVisit)
Coefficients:
           (Intercept)  ex0327$DurationOfVisit
              0.317851                0.008003
> summary(pollen.lm)
Call:
lm(formula = ex0327$PollenRemoved ~ ex0327$DurationOfVisit)
Residuals:
     Min       1Q   Median       3Q      Max
-0.28588 -0.11189  0.03608  0.11010  0.30805
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.317851   0.038563   8.242 1.53e-10 ***
ex0327$DurationOfVisit 0.008003   0.001436   5.574 1.33e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1624 on 45 degrees of freedom
```

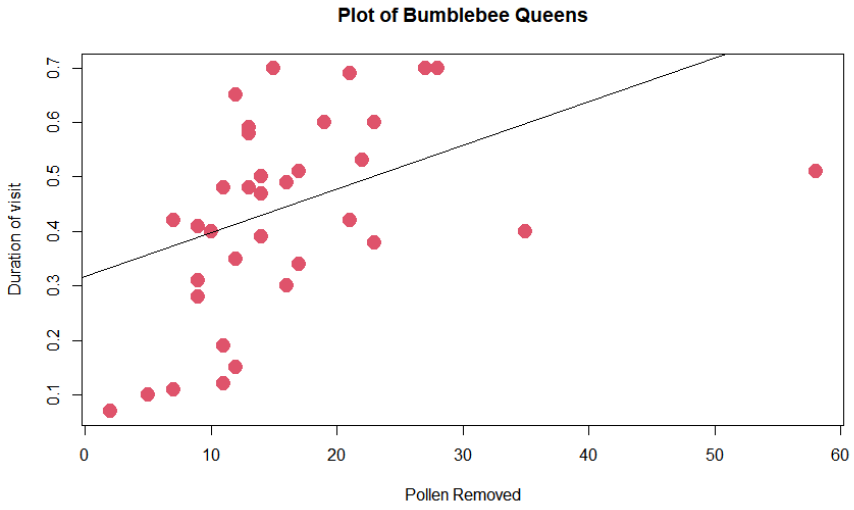
```
Multiple R-squared:  0.4084,  Adjusted R-squared:  0.3953
F-statistic: 31.07 on 1 and 45 DF,  p-value: 1.333e-06
> abline(pollen.lm)
> pollen.lm <- lm( ex0327$PollenRemoved ~ ex0327$DurationOfVisit)
> pollen.lm

Call:
lm(formula = ex0327$PollenRemoved ~ ex0327$DurationOfVisit)

Coefficients:
           (Intercept)  ex0327$DurationOfVisit
              0.317851                0.008003

> summary(pollen.lm)

Call:
lm(formula = ex0327$PollenRemoved ~ ex0327$DurationOfVisit)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28588 -0.11189  0.03608  0.11010  0.30805

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.317851   0.038563   8.242 1.53e-10 ***
ex0327$DurationOfVisit 0.008003   0.001436   5.574 1.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1624 on 45 degrees of freedom
Multiple R-squared:  0.4084,  Adjusted R-squared:  0.3953
F-statistic: 31.07 on 1 and 45 DF,  p-value: 1.333e-06
```

From the above result output, the intercept is 0.318751 and slope is 0.008003 with standard error of 0.001436; thus forming the simple linear regression model with the equation:

Pollen removed = 0.318751 + 0.008003 * Duration of visit

For this model, lets fit a line

```
> abline(pollen.lm)
```



There are few outliers that could control the data and the line is not well fitted into the data which does not estimate the mean properly.

---

17. a

This problem consists of data to find the estimate of the time needed after the slaughter when pH reaches 6.0. During the meat processing pH in postmortem muscle decreases to 6.0 from 7.0 to 7.2 at the time of slaughter. The data consists of 10 steer carcasses for which pH is measured at one of five times after slaughter. The summary of the data goes as below

```
> summary(case0702)
      Time            pH
 Min.   :1.0   Min.   :5.360
 1st Qu.:2.0   1st Qu.:5.643
 Median :4.0   Median :6.030
 Mean   :4.2   Mean   :6.120
 3rd Qu.:6.0   3rd Qu.:6.487
 Max.   :8.0   Max.   :7.020
```

Let's establish a simple linear regression model out of this data and find the least square estimates for the equation of pH by transforming hours to log scale

```
> meat.lm <- lm(case0702$pH ~ log(case0702$Time))
> meat.lm
Call:
```

```
lm(formula = case0702$pH ~ log(case0702$Time))

Coefficients:

       (Intercept)  log(case0702$Time)

            6.9836             -0.7257

> summary(meat.lm)
```

From the above generated linear model, the simple linear regression equation is

pH = 6.9836 – 0.7257*log(Time)

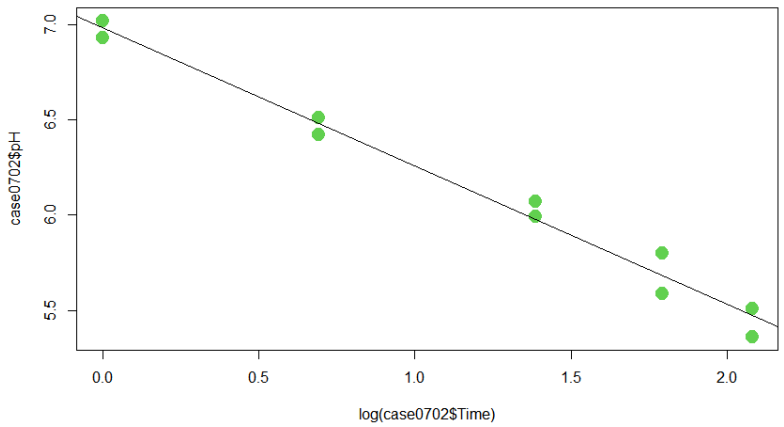
```
Call:

lm(formula = case0702$pH ~ log(case0702$Time))

Residuals:

     Min       1Q   Median       3Q      Max

-0.11466 -0.05888  0.02085  0.03612  0.11658

Coefficients:

                   Estimate Std. Error t value Pr(>|t|)

(Intercept)         6.98363    0.04853  143.90 6.08e-15 ***

log(case0702$Time) -0.72566    0.03443  -21.08 2.70e-08 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08226 on 8 degrees of freedom

Multiple R-squared:  0.9823,  Adjusted R-squared:  0.9801

F-statistic: 444.3 on 1 and 8 DF,  p-value: 2.695e-08
```

As written above, the intercept is 6.98363 with standard error of 0.04853 and slope is -0.72566 with standard error of 0.03443.

Also, the residual standard error is found to be 0.08226

For the generated model, let’s fit the regression equation line and see how well the line fits with the data

```
> plot(log(case0702$Time), case0702$pH, col = 3, pch = 19, cex = 2)

> abline(meat.lm)
```



From the above plot, the equation well fits into the data without overfitting or underfitting any points.

18.b

In connection to above data, let's calculate 95% prediction interval at 5 hours after slaughter

```
> meunlog.pred <- predict(meat.lm, newdata = data.frame(Time = 5), interval = "prediction")
```

```
> head(meunlog.pred)
```

```
       fit      lwr      upr
1 5.815725 5.614009 6.01744
```

The 95% prediction interval at 5 hours after slaughter lies between 5.614 and 6.017

We thought of plotting this interval bands on the plot by the below commands

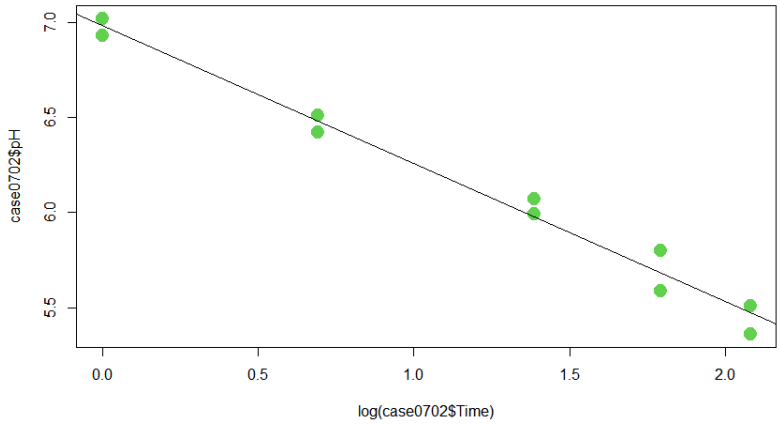
```
> plot(log(case0702$Time), case0702$pH, col = 3, pch = 19, cex = 2)
```

```
> abline(meat.lm )
```

```
> points(data.frame(Time = 5)[,1], meunlog.pred[,1], type = "l", col = "purple", lwd = 3, lty = 3)
```

```
> points(data.frame(Time = 5)[,1], meunlog.pred[,1], type = "l", col = "purple", lwd = 3, lty = 3)
```

The above code was executed without errors and warnings but the lines were not displayed on the plot. The plot after executing the above lines is as below



28.

This problem is based on the activity effect on reorganization of human central nervous system based on the fact that the part of brain associated with finger or limb activity is taken over by other purposes in individuals who lost limb or finger. The data consists of the test results obtained from the neuron activity index from MSI and the years that the one had been playing a stringed instrument.

In this current problem, magnetic source index (MSI) is used to measure neuronal activity in the brains of nine string players – six violinists, two cellists and one guitarist and six control who never played musical instrument; giving a total of 15 data points. With the help of this data points, lets try to find if there is any difference in neuronal activity between string players and control group

Let's get into the summary details

```
> summary(ex0728)
     Years          Activity
 Min.   : 0.0   Min.   : 5.00
 1st Qu.: 0.0   1st Qu.: 9.25
 Median : 6.0   Median :16.00
 Mean   : 7.2   Mean   :15.57
 3rd Qu.:12.5   3rd Qu.:24.00
 Max.   :19.0   Max.   :26.50
```

Since there is no definite control group defined, let's divide the activity into control and string group for 0 and other than 0 years respectively, done by the below code

```
> control <- ex0728$Activity[ex0728$Y == "0"]
> control
[1]  5.0  6.0  7.5  9.0  9.5 11.0
> string <- ex0728$A[ex0728$Y != "0"]
> string
[1] 16.0 16.5 11.5 16.0 25.0 25.5 25.5 23.0 26.5
```

Let's get summary statistics for the above data

```
> summary(control)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.000   6.375   8.250   8.000   9.375  11.000
> summary(string)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.50   16.00   23.00   20.61   25.50   26.50
```

The above statistics show that there is much difference in the mean and median of control group and musician group. Apart from this, the both groups are also varied to a great extent.

Let's apply t-test on this data to find how strong there is a real difference in neuronal activity and stringed musicians (string group) and control group

```
> t.test(control, string, var.equal = T)

     Two Sample t-test

data:  control and string
t = -5.1988, df = 13, p-value = 0.0001714
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -17.851677  -7.370546
sample estimates:
mean of x mean of y
  8.00000  20.61111
```

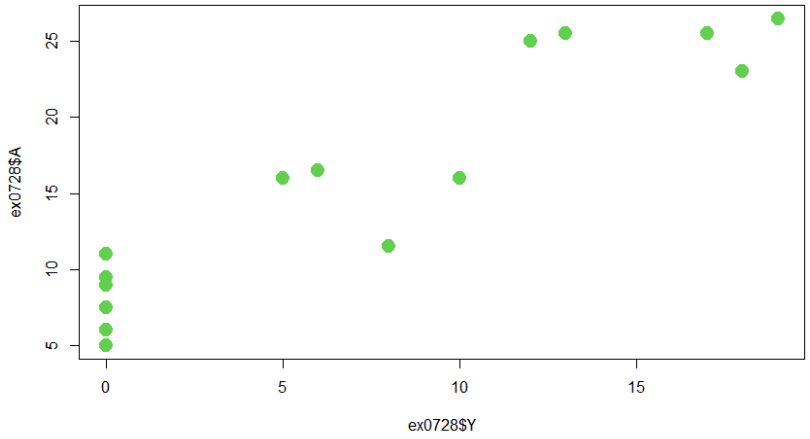
The above two sample t test shows a string convincing evidence that there is a difference in neuronal activity in controls and stringed musicians with a p value < 0.000. The estimated mean

for string musicians is in confidence interval 7.37 to 17.85; higher than for control non musician group.

Let's run a simple linear regression analysis to get amount of activity associated with the number of years the individuals have been playing the instrument

In this process, let's make a plot of the data for activity (y) Vs years (x)

```
> plot(ex0728$Y, ex0728$A, col = 3, pch = 19, cex = 2)
```

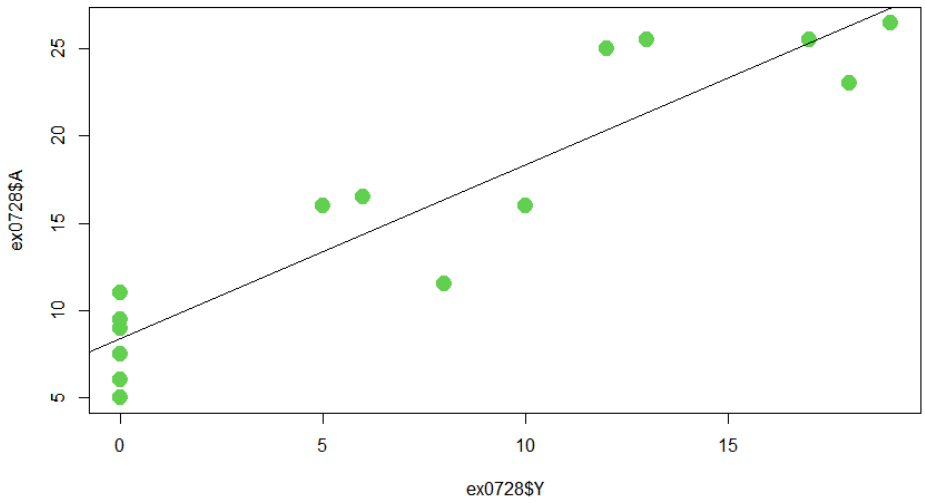


Let's develop a linear model to fit a line to the data

```
> msi.lm <- lm(ex0728$A ~ ex0728$Y)
> abline(msi.lm)
> summary(msi.lm)
Call:
lm(formula = ex0728$A ~ ex0728$Y)
Residuals:
    Min      1Q  Median      3Q     Max
-4.8644 -2.3730  0.1614  2.3713  4.6471
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.3873     1.1149   7.523 4.35e-06 ***
ex0728$Y      0.9971     0.1110   8.980 6.18e-07 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.009 on 13 degrees of freedom

Multiple R-squared:  0.8612,  Adjusted R-squared:  0.8505

F-statistic: 80.63 on 1 and 13 DF,  p-value: 6.178e-07
```

From the above linear model, a regression equation can be developed with an intercept of 8.387 and slope of 0.9971. When this line is plotted on the above plot diagram, the following plot is resulted



The above figure shows that there is a linear increase in the neuronal activity with increase in years. The estimated mean neuronal activity increases by 1 point with a standard error of 0.11 per an year increase in age.

8th Chapter 24

This problem is a study on to detect what a true high respiratory rate is. For this, the physicians should have a picture of distribution of normal respiratory rates and so a study is conducted on 618 children between the ages of 15 days and 3 years and measured respiratory rates. In this study we will make an analysis on the data and provide a statistical summary by transforming the data into different sets.

Let’s get into the summary of the data

```
> summary(ex0824)
      Age              Rate
 Min.   : 0.10   Min.   :18.00
 1st Qu.: 3.80   1st Qu.:30.00
```

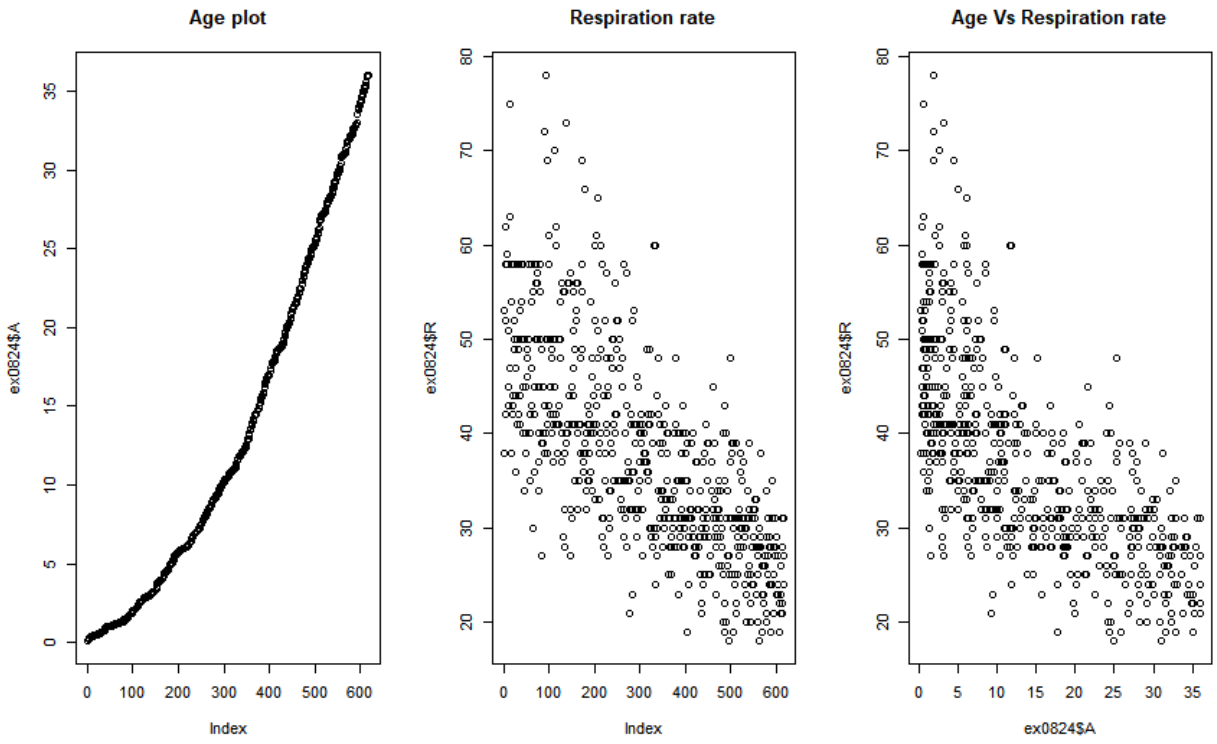
```
 Median :10.55   Median :36.50
 Mean   :13.39   Mean   :37.74
 3rd Qu.:22.00   3rd Qu.:44.00
 Max.   :36.00   Max.   :78.00
```

We can see that the data consisted of two fields of age (in months) and the child respective respiratory rate. From the summary it can be observed that there is a large variation in the data and suggests certain transformations in analysis.

Let's have the plot of data, all in one picture to have an analysis

```
> par(mfrow = c(1,3))
> plot(ex0824$A, main = "Age plot")
> plot(ex0824$R, main = "Respiration rate")
> plot(ex0824$A, ex0824$R, main = "Age Vs Respiration rate")
```

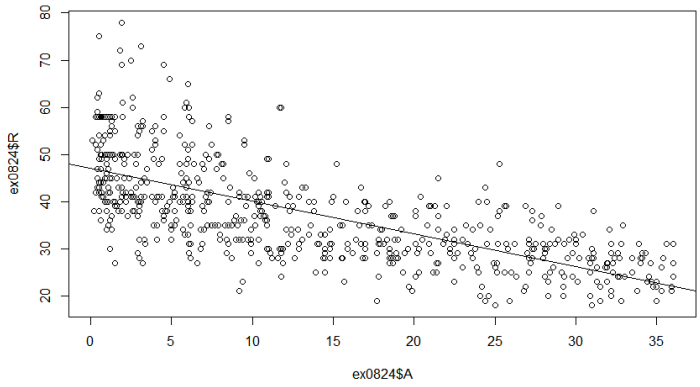
And the above code results in the plots below



We can see that there is lot of spread in the data to fit a regression equation. One comment that can be made on the Age Vs Rate plot is that as age increases, the respiration rate gradually decreased but there is no perfect uniformity in this statement.

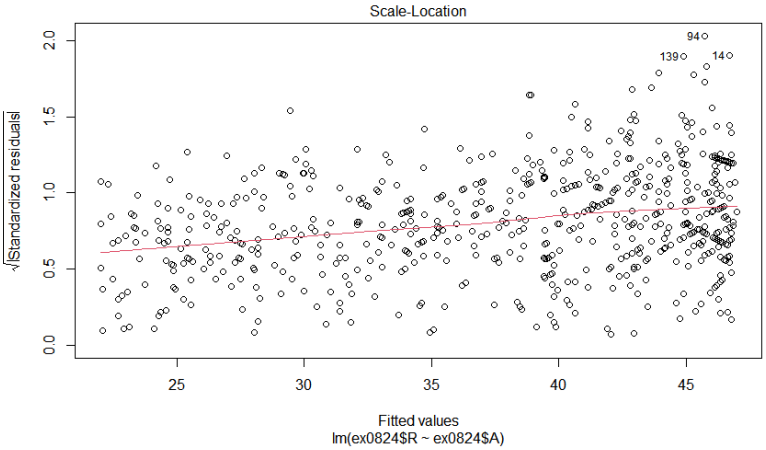
Let’s try generating a regression model on this

```
> no.lm <- lm(ex0824$R ~ ex0824$A)
> no.lm <- lm(ex0824$R ~ ex0824$A)
> plot(ex0824$R ~ ex0824$A)
> abline(no.lm, 1)
> summary(no.lm)

Call:
lm(formula = ex0824$R ~ ex0824$A)

Residuals:
    Min      1Q  Median      3Q     Max
-19.652  -5.432  -0.608   4.589  32.270

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.05216    0.50422   93.32   <2e-16 ***
ex0824$A    -0.69571    0.02938  -23.68   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.842 on 616 degrees of freedom
Multiple R-squared:  0.4766,  Adjusted R-squared:  0.4758
F-statistic: 560.9 on 1 and 616 DF,  p-value: < 2.2e-16
```



The above plot seems ok, let's have a look at the residual plot

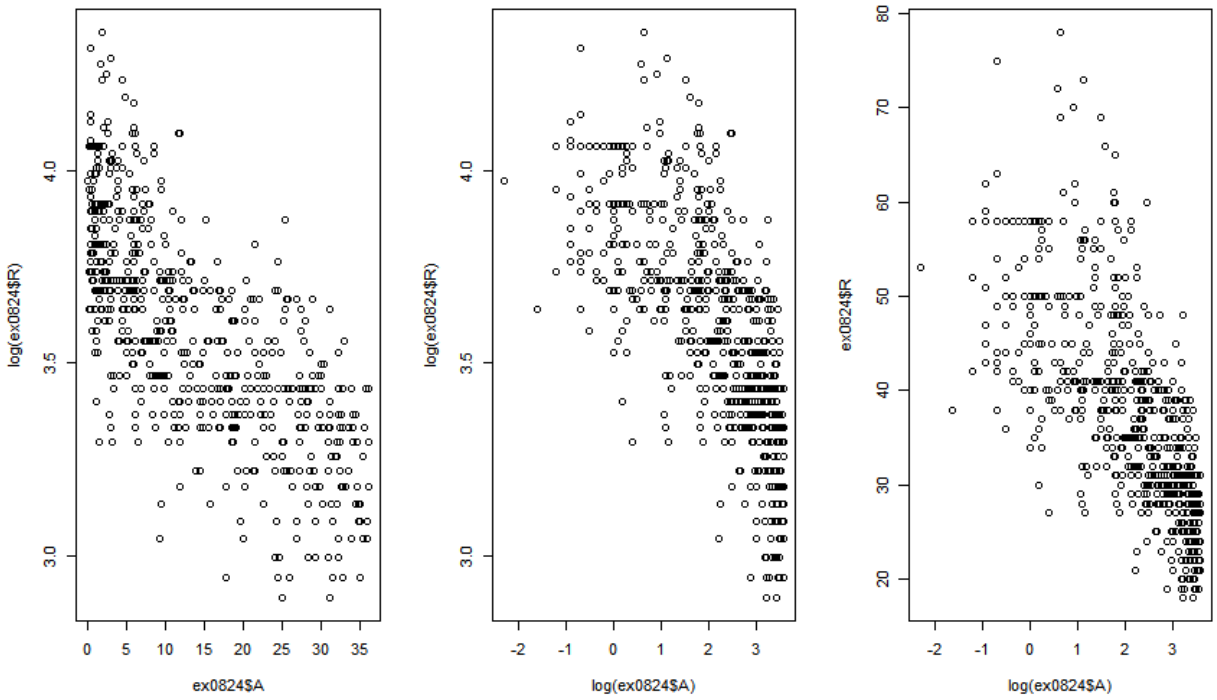
> plot(no.lm, 3)



The residual plot has few diverse residuals

Let's try few transformations on the data to see which is better. Let's try log transformation combinations on the data

```
> par(mfrow = c(1,3))
> plot(ex0824$A, log(ex0824$R))
> plot(log(ex0824$A), log(ex0824$R))
> plot(log(ex0824$A), ex0824$R)
> par(mfrow = c(1,1))
```

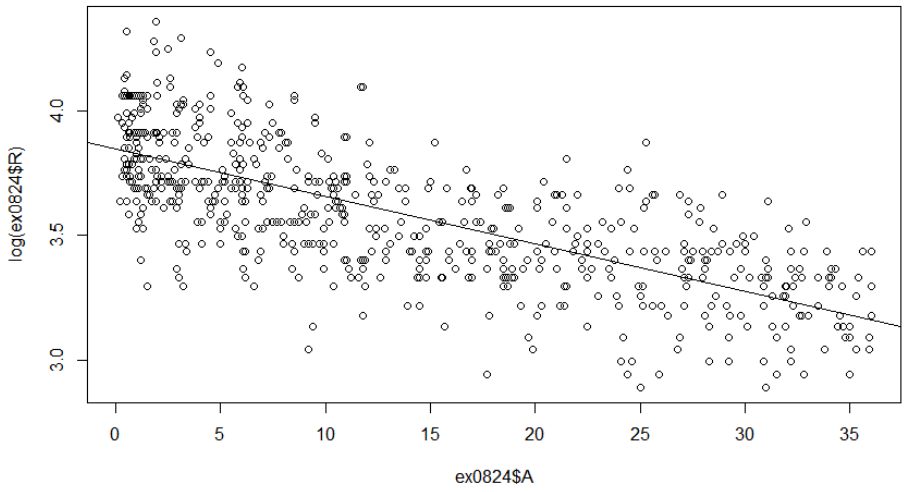


From the above plot, the first plot can be selected rather than second and third.

Let's try generating a regression model on this first transformed data by taking log scale on rate field

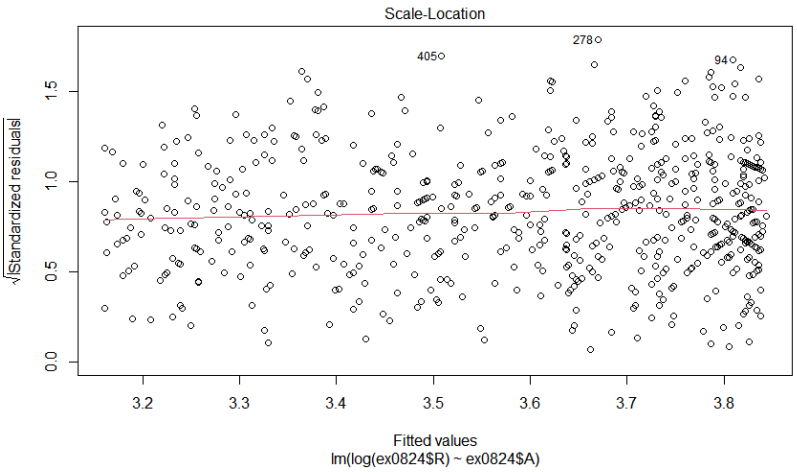
```
> logr.lm <- lm(log(ex0824$R) ~ ex0824$A)
> plot(log(ex0824$R) ~ ex0824$A)
> abline(logr.lm, 1)
> summary(logr.lm)

Call:
lm(formula = log(ex0824$R) ~ ex0824$A)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62571 -0.13201 -0.00402  0.13489  0.54771

Coefficients:
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.8451185  0.0126277  304.50   <2e-16 ***
ex0824$A     -0.0190090  0.0007357  -25.84   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1964 on 616 degrees of freedom
Multiple R-squared:  0.5201,  Adjusted R-squared:  0.5193
F-statistic: 667.6 on 1 and 616 DF,  p-value: < 2.2e-16
```



The above plot seems ok, let’s have a look at the residual plot

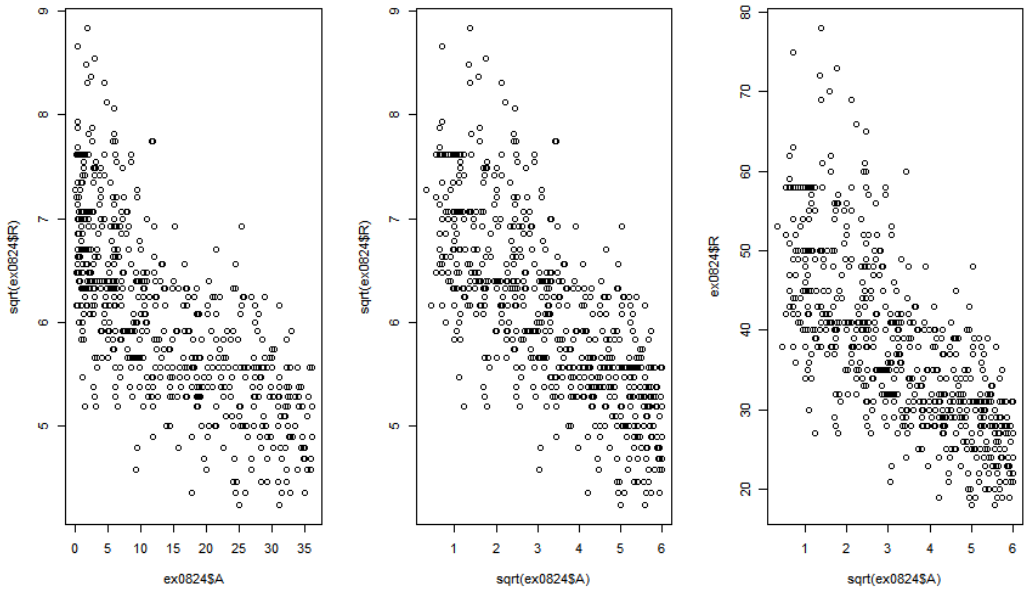
```
> plot(logr.lm, 3)
```



The residual plot has few diverse residuals. This is a better plot than before and could be better than the graphs which will come hereby

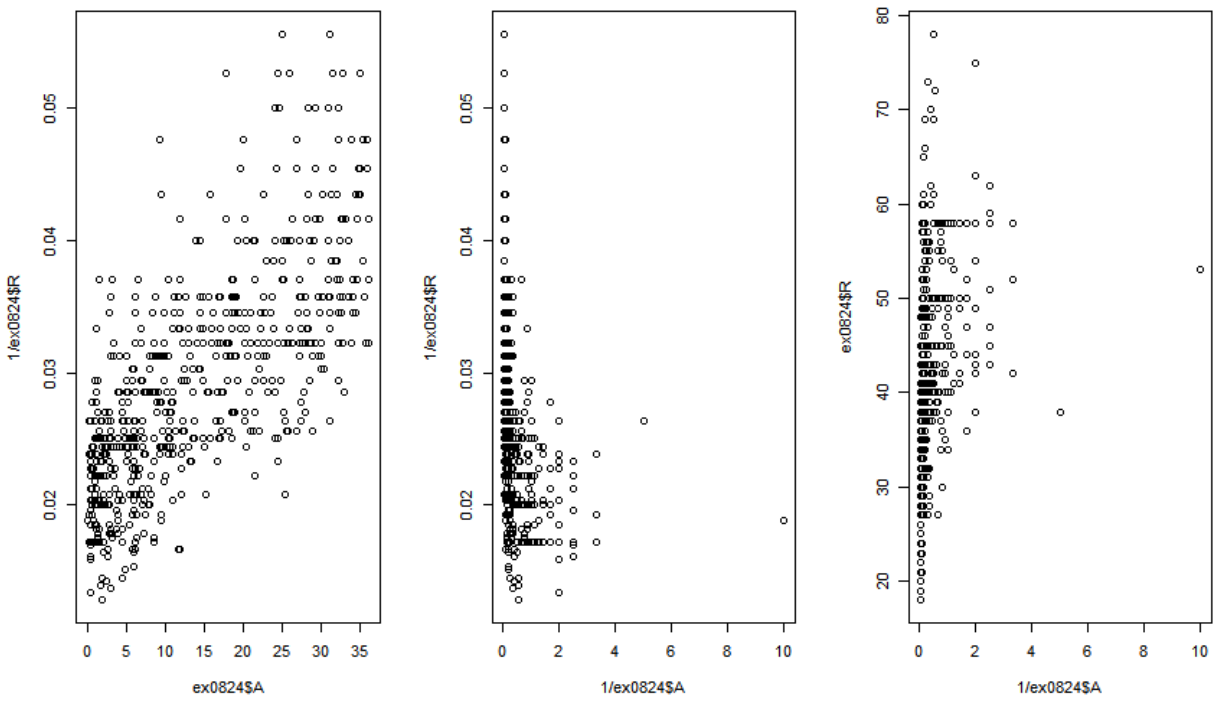
Let's try transforming the data by imposing square root combinations

```
> par(mfrow = c(1,3))
> plot(ex0824$A, sqrt(ex0824$R))
> plot(sqrt(ex0824$A), sqrt(ex0824$R))
> plot(sqrt(ex0824$A), ex0824$R)
> par(mfrow = c(1,1))
```



All the plots seems good to fit an equation but let's try transforming the data by reciprocal combinations

```
> par(mfrow = c(1,3))
> plot(ex0824$A, 1/ex0824$R)
> plot(1/ex0824$A, 1/ex0824$R)
> plot(1/ex0824$A, ex0824$R)
> par(mfrow = c(1,1))
```



Though the first plot seem to be better than second and thirds (which could be directly rejected), the square root transformations look better than these inversion transformations