Swetha Adike
Venu Goud Raparti
Problems for Chapter 9
MSIS 545

12.

This problem is an observatory study of brain weights and body sizes in respect to the evolutionary resistance with chance of surviving. The study is to determine which characteristics are associated with large brains and few other factors after getting effect of the body size. The data consists of the average values of brain weight, body weights, gestation length and litter size of 96 species of mammals. Since brain and body sizes are generally related to each other, the interest area is to find other variables (if any) are associated with brain size.

Before getting into further details, lets have a look at the basic summary of the data

```
> dim(case0902)
[1] 96  5
> summary(case0902[,-1])
     Brain              Body            Gestation         Litter      
 Min.   :   0.45   Min.   :   0.017   Min.   : 16.0   Min.   :1.00  
 1st Qu.:  12.60   1st Qu.:   2.075   1st Qu.: 63.0   1st Qu.:1.00  
 Median :  74.00   Median :   8.900   Median :133.5   Median :1.20  
 Mean   : 218.98   Mean   : 108.328   Mean   :151.3   Mean   :2.31  
 3rd Qu.: 260.00   3rd Qu.:  94.750   3rd Qu.:226.2   3rd Qu.:3.20  
 Max.   :4480.00   Max.   :2800.000   Max.   :655.0   Max.   :8.00
```

As discussed there are 96 data records different specie with 4 factors. The data shows an indication of lg transformations as there is wide spread.

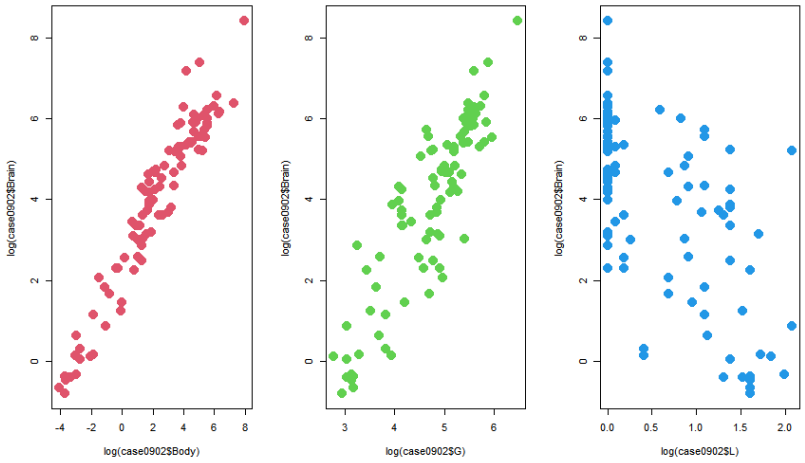
a. Let's draw a scatter plot of mammal brain weight data with all variables to transformed their logarithms

Below is the code for plot of mammal brain weight on log transformed other 3 factors Body weight, gestation and litter size

```
> par(mfrow = c(1,3))
> plot(log(case0902$Body), log(case0902$Brain), col = 2, pch = 19, cex = 2)
> plot(log(case0902$G), log(case0902$Brain), col = 3, pch = 19, cex = 2)
```

```
> plot(log(case0902$L), log(case0902$Brain), col = 4, pch = 19, cex = 2)
> par(mfrow = c(1,1))
```

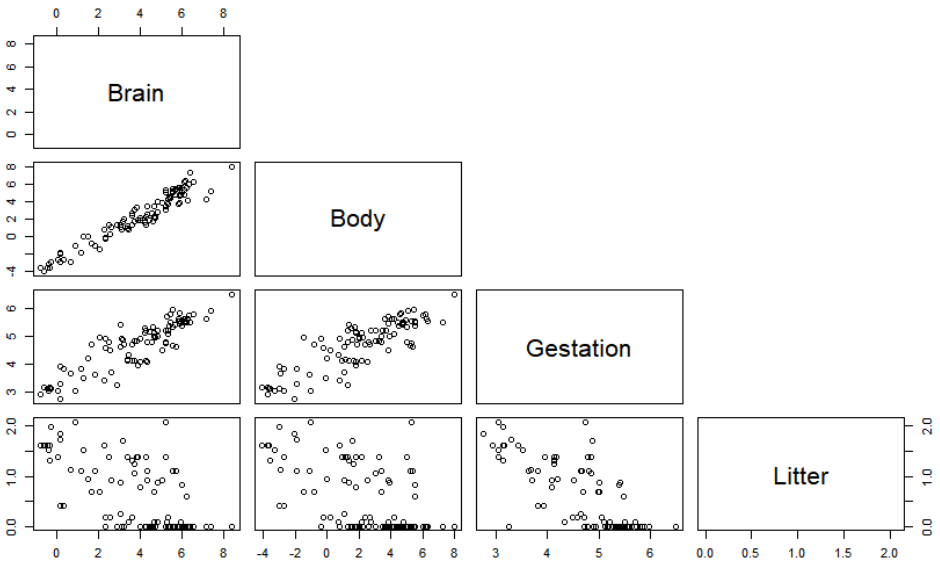
And the plot is as shown below



The above plot is less overlapped and has different patterns. The plot suggests a fit to regression model for log of brain weight and log of body weight, log gestation and log litter size.

R Code for scatter plot to depict the picture in 9.15 is as follows

```
> plot(log(case0902[,-1]), upper.panel = NULL)
```



b. Let's fit a multiple linear regression of log brain on log body weight, log gestation and log litter size.

For this, first let's make a linear model for log brain on log transformed other three variables. Below is code for the linear model

```
> bw.lm <- lm(log(Brain) ~ log(Body) + log(Gestation) + log(Litter), data = case0902)

> summary(bw.lm)

Call:

lm(formula = log(Brain) ~ log(Body) + log(Gestation) + log(Litter),

    data = case0902)

Residuals:

     Min       1Q   Median       3Q      Max

-0.95415 -0.29639 -0.03105  0.28111  1.57491

Coefficients:

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

(Intercept)     0.85482    0.66167   1.292  0.19962

log(Body)       0.57507    0.03259  17.647  < 2e-16 ***

log(Gestation)  0.41794    0.14078   2.969  0.00381 **

log(Litter)    -0.31007    0.11593  -2.675  0.00885 **

---

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

Residual standard error: 0.4748 on 92 degrees of freedom

Multiple R-squared:  0.9537,  Adjusted R-squared:  0.9522

F-statistic: 631.6 on 3 and 92 DF,  p-value: < 2.2e-16
```

The above resulted multiple linear regression coefficients confirm the estimates in Display 9.15 and the confidence intervals of above estimated coefficients are as follows

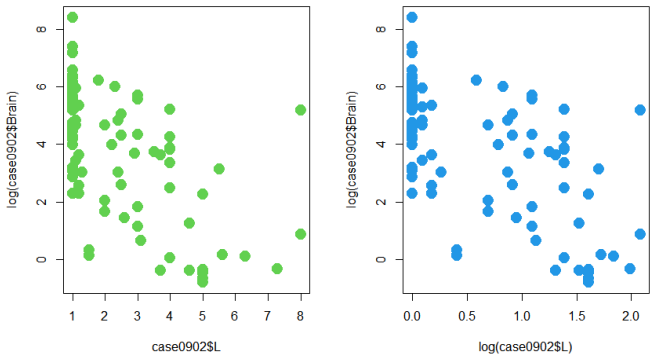
```
> confint(bw.lm)

                   2.5 %      97.5 %

(Intercept)    -0.4593167  2.16896055

log(Body)       0.5103490  0.63979373

log(Gestation)  0.1383359  0.69754827
```

```
log(Litter)     -0.5403124 -0.07982996
```

c. In this, let's make a matrix of scatterplots with litter size on natural scale and compare its relation with log transformed values with respect to log transformed brain weight.

```
> par(mfrow = c(1,2))

> plot(case0902$L, log(case0902$Brain), col = 3, pch = 19, cex = 2)

> plot(log(case0902$L), log(case0902$Brain), col = 4, pch = 19, cex = 2)

> par(mfrow = c(1,1))
```

Below is the resulting graph for the above code



As far as the above graphs are considered, there is no much difference between the two plots but the relation seems to be better when litter is taken in natural scale since the points spread is less when compared to the log transformed plot.

15.

This problem consists of data containing corn yield and respective rainfall details in six US corn producing states for a period of time from 1890 to 1927. Let's head into the data details

```
> dim(ex0915)

[1] 38  3
```

The above code result says that there are 38 records of corn yield on certain recorded rainfall with respect to the year, corresponding to 3 fields and 38 attributes. Let's get into the summary details

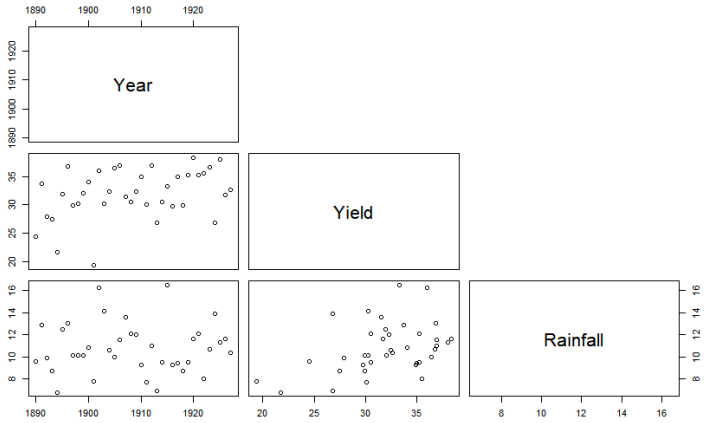
```
> summary(ex0915)

      Year          Yield          Rainfall

 Min.   :1890   Min.   :19.40   Min.   : 6.800
```

```
 1st Qu.:1899   1st Qu.:29.95   1st Qu.: 9.425
 Median :1908   Median :32.15   Median :10.500
 Mean   :1908   Mean   :31.92   Mean   :10.784
 3rd Qu.:1918   3rd Qu.:35.20   3rd Qu.:12.075
 Max.   :1927   Max.   :38.30   Max.   :16.500
```

The summary details says that the spread is not much, suggesting no requirement of any transformations.

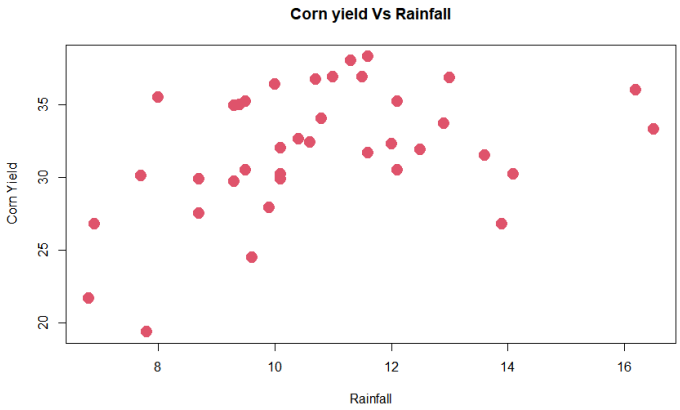
Let's make a scatterplot matrix of the variables

```
> plot(ex0915, upper.panel = NULL)
```



a. Let's plot corn yield Vs rainfall to see how they are related individually.

```
> plot(ex0915$Rainfall, ex0915$Yield, col = 2, pch = 19, cex = 2, xlab = "Rainfall", + ylab = "Corn Yield", main = "Corn yield Vs Rainfall")
```



The above plot shows that the corn yield increases with rainfall up to a rainfall of 12 inches and then slightly decreases with higher rainfall. To fit these data points, a straight-line regression model is not satisfactory.

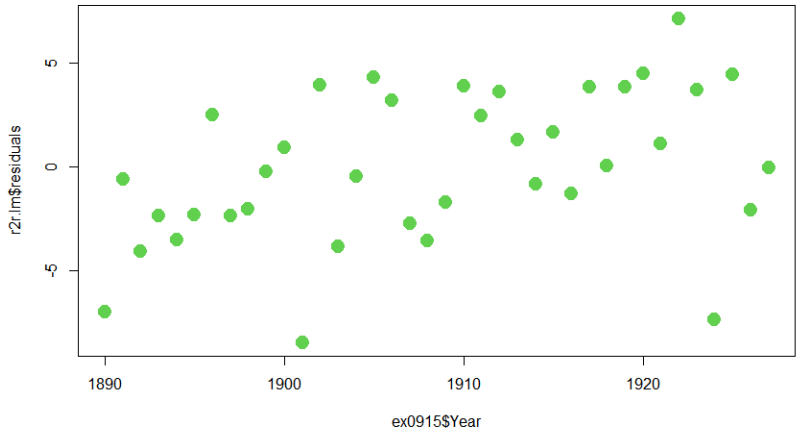
b. Let's fit a regression model for corn yield on rain and rain^2 with below code

```
> r2 <- ex0915$Rainfall^2
> r2r.lm <- lm(ex0915$Yield ~ ex0915$Rainfall + r2)
> summary(r2r.lm)
Call:
lm(formula = ex0915$Yield ~ ex0915$Rainfall + r2)
Residuals:
    Min      1Q  Median      3Q     Max
-8.4642 -2.3236 -0.1265  3.5151  7.1597
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.01467   11.44158  -0.438  0.66387
ex0915$Rainfall  6.00428    2.03895   2.945  0.00571 **
r2             -0.22936    0.08864  -2.588  0.01397 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.763 on 35 degrees of freedom
Multiple R-squared:  0.2967,  Adjusted R-squared:  0.2565
F-statistic: 7.382 on 2 and 35 DF,  p-value: 0.002115
```

c. Now let's plot the residuals for the above model Vs year and see if there is any pattern connected to advancements in technology

```
> plot(ex0915$Year, r2r.lm$residuals, col = 3, pch = 19, cex = 2)
```

Below is the plot for this code



The above plot shows a trend that residuals increased with increase in years which can be inferred that the yield is larger than what is predicted from the above regression model of yield on rainfall

d. Let's fit a multiple regression of corn yield on rain, rain^2 and year, from the below code

```
> r2ry.lm <- lm(ex0915$Yield ~ ex0915$Rainfall + r2 + ex0915$Year)
> summary(r2ry.lm)
Call:
lm(formula = ex0915$Yield ~ ex0915$Rainfall + r2 + ex0915$Year)
Residuals:
    Min      1Q  Median      3Q     Max
-9.3995 -1.8086 -0.0479  2.4050  5.1839
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -263.30324   98.24094  -2.680  0.01126 *
ex0915$Rainfall    5.67038    1.88824   3.003  0.00499 **
r2                -0.21550    0.08207  -2.626  0.01286 *
ex0915$Year        0.13634    0.05156   2.644  0.01229 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 3.477 on 34 degrees of freedom
Multiple R-squared:  0.4167,  Adjusted R-squared:  0.3652
```

```
F-statistic: 8.095 on 3 and 34 DF,  p-value: 0.0003339
```

From the above fitted regression model, lets extract coefficients and compare rain and rain^2 coefficients with the r2r model in 15.c. The below are the coefficients for the regression model in this section, as per the above results

```
> r2ry.lm$coefficients
    (Intercept) ex0915$Rainfall              r2     ex0915$Year
   -263.3032448       5.6703818      -0.2154981        0.1363414
```

Below is the regression model coefficients from the regression model of corn yield on rain and rain^2, from the previous section

```
> r2r.lm$coefficients
    (Intercept) ex0915$Rainfall              r2
     -5.0146670       6.0042835      -0.2293639
```

From the above, there is not much difference in the coefficients. Let's take a look at standard errors.

Std error {yield | rain, rain^2} = 3.763

Std error {yield | rain, rain^2, year} = 3.477

Below table is a summary comparison of both the models. The standard error of the coefficients decreased when compared to rain and rain^2 model.

Estimates	Intercept	Rain	Rain^2	year
Coefficients(r2r)	-5.015	6.004	-0.229	
R2ry	-263.303	5.670	-0.216	0.136
Std Error of coef(r2r)	11.442	2.039	0.089	
R2ry	98.241	1.889	0.082	0.0512
P value(r2r)	0.66	0.005	0.01	
R2ry	0.01	0.005	0.01	0.01
Standard error(r2r)	3.763			
R2ry	3.477			

On contrary to the estimated coefficient, the standard error decreased when year is added to rain and rain^2.

As in the question, an inch increase in rainfall

e. Let's fit multiple regression of corn yield on rain, rain^2, year and year * rain

```
> ry <- ex0915$Year * ex0915$Rainfall
```

```
> r2ryry.lm <- lm(ex0915$Yield ~ ex0915$Rainfall + r2 + ex0915$Year
+ ry)

> summary(r2ryry.lm)

Call:

lm(formula = ex0915$Yield ~ ex0915$Rainfall + r2 + ex0915$Year +
ry)

Residuals:

    Min      1Q  Median      3Q     Max

-6.2969 -2.5471  0.6011  1.9923  5.0204


Coefficients:

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

(Intercept)     -1.909e+03  4.862e+02  -3.927 0.000414 ***

ex0915$Rainfall  1.588e+02  4.457e+01   3.564 0.001138 **

r2              -1.862e-01  7.198e-02  -2.588 0.014257 *

ex0915$Year      1.001e+00  2.555e-01   3.919 0.000423 ***

ry              -8.064e-02  2.345e-02  -3.439 0.001599 **

---

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

Residual standard error: 3.028 on 33 degrees of freedom

Multiple R-squared:  0.5706,  Adjusted R-squared:  0.5185

F-statistic: 10.96 on 4 and 33 DF,  p-value: 9.127e-06
```

The p value for rain*year says that the effect of rainfall on yield is smaller as the years increase (to 1927). It can be said that due in technology, the yield becomes independent from the rainfall.

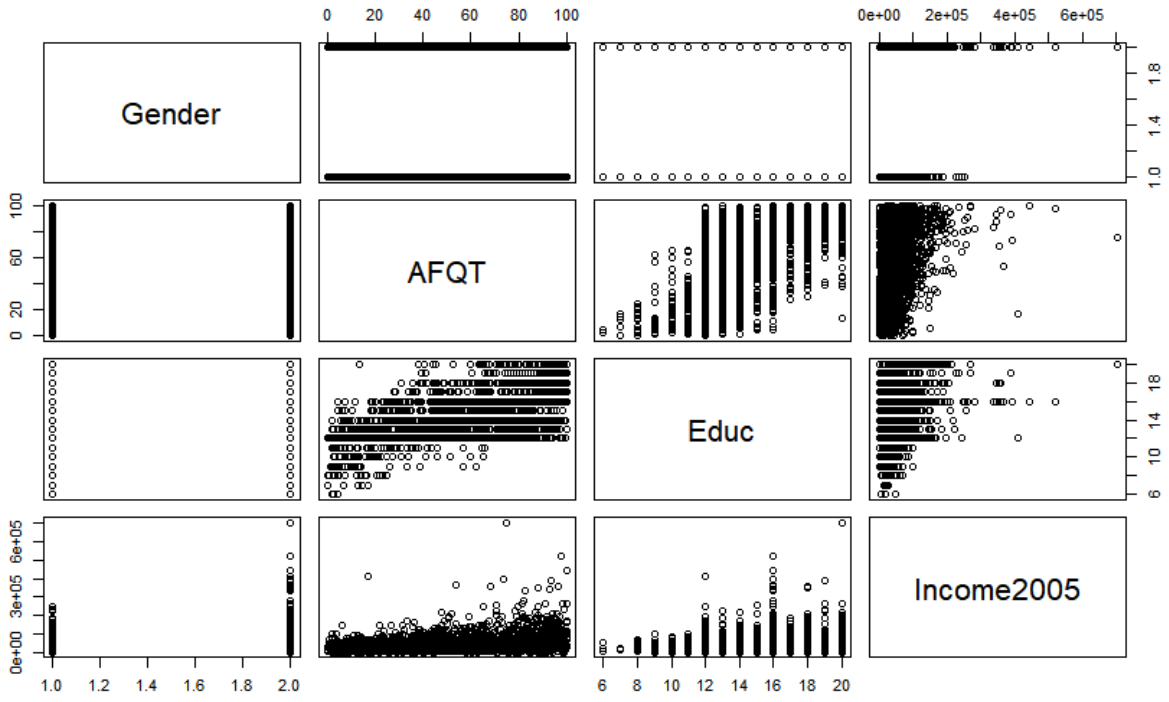
23.

This problem is to find if there is any intelligence difference on gender, associated with income. A test is made on a sample with respect to different quotients on mind to know the capability. The samples were tested for arithmetic, word knowledge, comprehension and mathematical knowledge. AFQT gives an overall score depicting all these four terms. This AFQT is measured in 1981 and years of education completed by 2006 for 1306 males and 1278 females between 14 and 22 years of age in 1979 and re interviewed in 2006.

The present exercise is to find if there is any evidence that mean salary for males exceeding the mean salary for females with same years of education and AFQT scores; by how many dollars or by what percent.

Before getting into the data, let's look at the summary

```
> summary(ex0923)
    Subject            Gender             AFQT                Educ
Income2005
 Min.    :     2    female:1278    Min.    :  0.00    Min.    : 6.00
Min.    :    63
 1st Qu.: 1586    male  :1306    1st Qu.: 31.48    1st Qu.:12.00    1st
Qu.: 23000
 Median : 3108                    Median : 56.80    Median :13.00
Median : 38231
Mean    : 3494                    Mean    : 54.44    Mean    :13.89    Mean
: 49417
 3rd Qu.: 4636                    3rd Qu.: 78.07    3rd Qu.:16.00    3rd
Qu.: 61000
 Max.    :12140                    Max.    :100.00    Max.    :20.00
Max.    :703637
> plot(ex0923[,-1])
```



The above summary statistics and scatter plot suggest log transformations on income since the spread is wide when compared to other variables.

Let's fit a regression model for income on education, AFQT scores and gender

```
> ex0923.lm <- lm(log(ex0923$I) ~ ex0923$Educ + ex0923$AFQT +
ex0923$Gender)
> summary(ex0923.lm)
Call:
lm(formula = log(ex0923$I) ~ ex0923$Educ + ex0923$AFQT +
ex0923$Gender)
Residuals:
    Min      1Q  Median      3Q     Max
-7.0906 -0.3301  0.1404  0.5091  2.5452

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.7312115  0.1026287  85.076  < 2e-16 ***
ex0923$Educ       0.0769506  0.0084888   9.065  < 2e-16 ***
ex0923$AFQT       0.0059139  0.0007657   7.724  1.6e-14 ***
ex0923$Gendermale 0.6245093  0.0341748  18.274  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.8661 on 2580 degrees of freedom
Multiple R-squared:  0.2101,  Adjusted R-squared:  0.2092
F-statistic: 228.7 on 3 and 2580 DF,  p-value: < 2.2e-16
```

From the above result with p value < 0.00001, there is convincing evidence that median salary of males exceeds the median salary of females with a positive relationship. To know the confidence interval of this statement, let's find the 95% confidence interval.

```
> exp(confint(ex0923.lm))
                        2.5 %      97.5 %
(Intercept)       5064.287091 7573.831779
ex0923$Educ          1.062161    1.098116
ex0923$AFQT          1.004422    1.007443
ex0923$Gendermale    1.746295    1.996753
```

From the above, with 95% confidence, the median income of male is between 1.7463 (74.63%) and 1.9967 or (99.67%) times the median income of female, with same years of education and AFQT scores.