Homework5

Yanyu Zheng yz2690

February 28, 2016

1. Chapter 7, problem 18

```
library("Sleuth3")
attach(case0702)
n = nrow(case0702)
fit1 = lm(pH ~ Time, case0702)
predict1 = predict(fit1, data.frame(Time = 5), interval = "prediction")
se.pred = (predict1[,'upr'] - predict1[,'fit']) / qt(0.975,n-3)
detach(case0702)
```

The standard error of the prediction is 0.1639102, the confidence band is 5.565463, 6.340635

2. Chapter 7, problem 24

```
attach(ex0724)
fitDenmark = lm(Denmark ~ Year, ex0724)
fitNetherlands = lm(Netherlands ~ Year, ex0724)
fitCanada = lm(Canada ~ Year, ex0724)
fitUSA = lm(USA ~ Year, ex0724)
denmark = summary(fitDenmark)$coefficients[2,3:4]
netherlands = summary(fitNetherlands)$coefficients[2,3:4]
canada = summary(fitCanada)$coefficients[2,3:4]
usa = summary(fitUSA)$coefficients[2,3:4]
table7 = cbind(denmark,netherlands,canada,usa)
```

- (a) From the summary above, we can easily confirm the estimates and standard errors.
- (b) The t value and p value
  $-2.0725983, 0.0442383, -5.7101959, 9.636921\times10^{-7}, -4.0166528, 7.3759467\times10^{-4}, -5.779212, 1.4391086\times 10^{-5}$

We can see there's evidence that the proportion of male births is truly declining.

- (c) The t-statistic is positively correlated with the slope but negetively correlated with the sum of suqres of residules. And from the plots we can see the usa data has really small sum of suqre of residules.
- (d) Because usa has smaller sum of squares of residules.

3. Chapter 7, problem 28

```r
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
attach(ex0728)
data7 = ex0728 %>%
  mutate(Group = as.factor(Years == 0))
detach(ex0728)
t.test(Activity ~ Group, data7)
```

```
##
##  Welch Two Sample t-test
##
## data:  Activity by Group
## t = 6.067, df = 11.313, p-value = 7.205e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.051473 17.170750
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            20.61111             8.00000
```

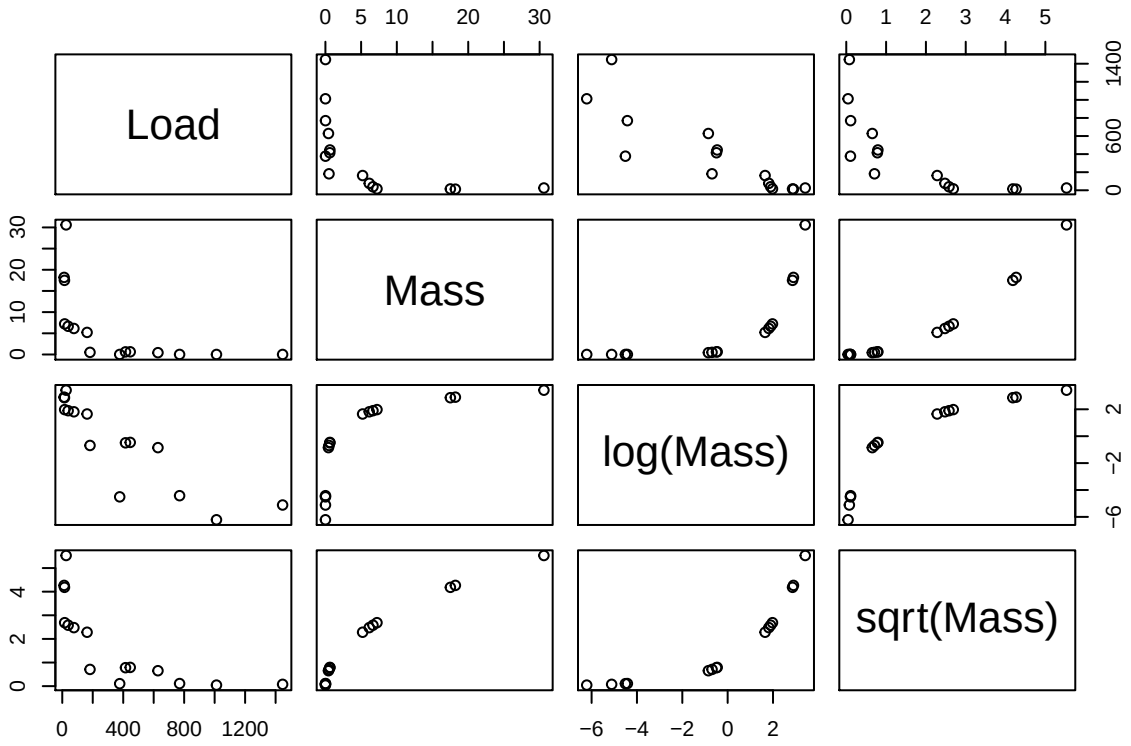
```r
summary(lm(Activity ~ Years, data7))$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 8.3872549  1.1148871 7.522963 4.354683e-06
## Years       0.9971405  0.1110454 8.979574 6.178311e-07
```

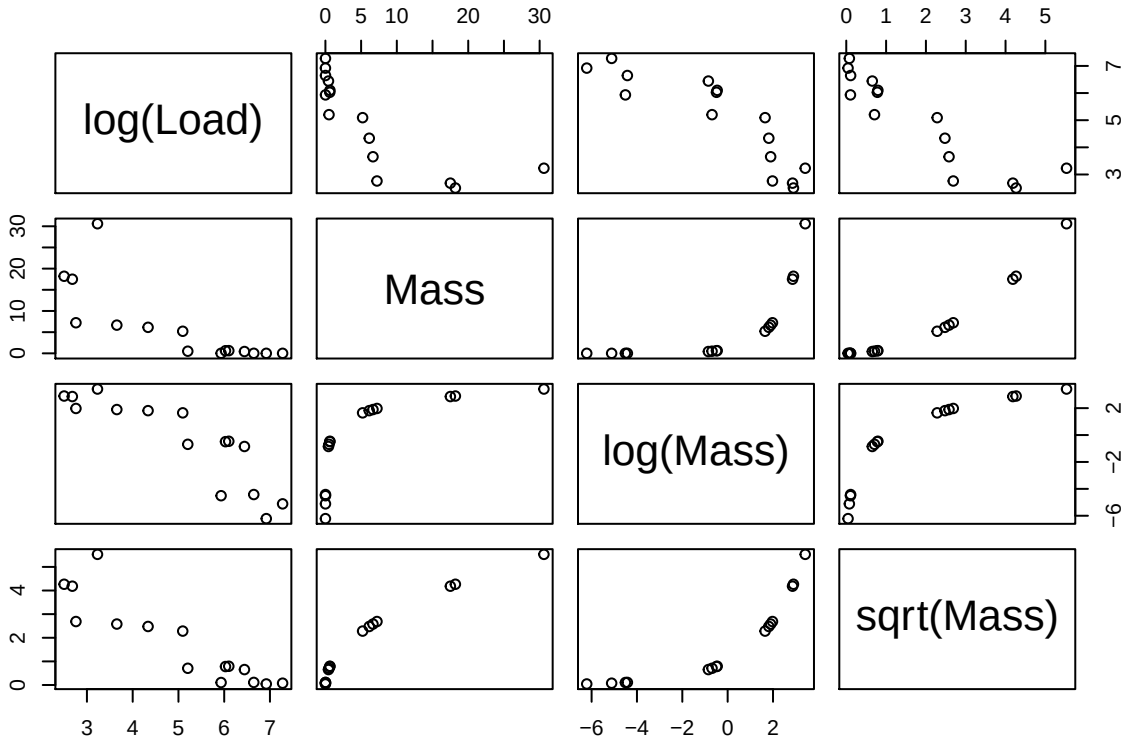
- From the two sample t test, pvalue = 7.205e 05 we can see there's significant different between string play and control.
- From the pvalue of the linear model, pvalue is 6.178311e 07 for Years, indicating significant correlation between brain activity and the year playing instrument.

4. Chapter 8, problem 17

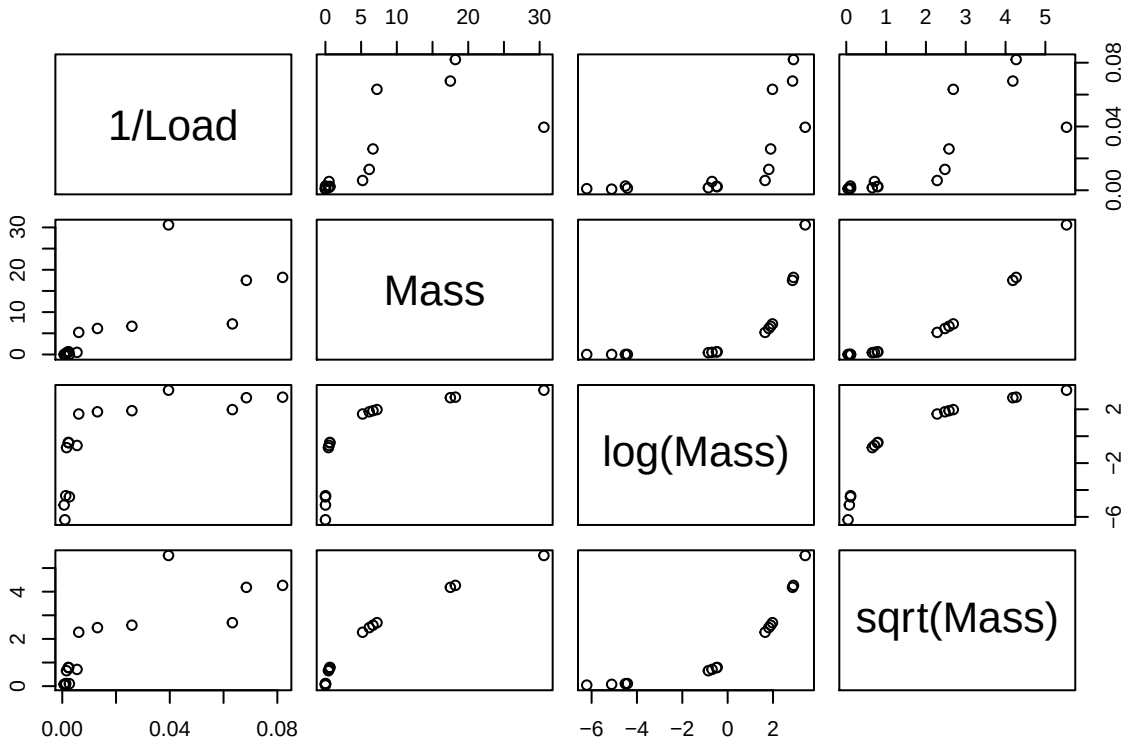
```r
attach(ex0817)
pairs(Load ~ Mass + log(Mass)  + 1/Mass + sqrt(Mass))
```



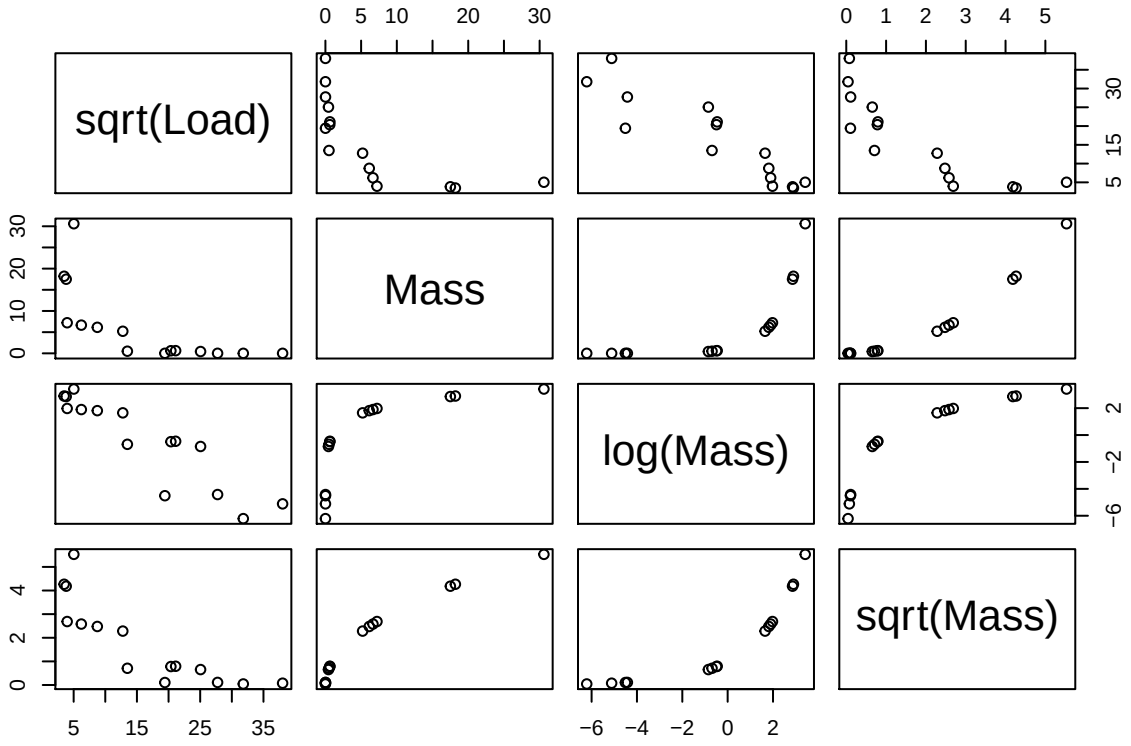

```
pairs(log(Load) ~ Mass + log(Mass)  + 1/Mass + sqrt(Mass))
```




```
pairs(1/Load ~ Mass + log(Mass)  + 1/Mass + sqrt(Mass))
```




```
pairs(sqrt(Load) ~ Mass + log(Mass)  + 1/Mass + sqrt(Mass))
```

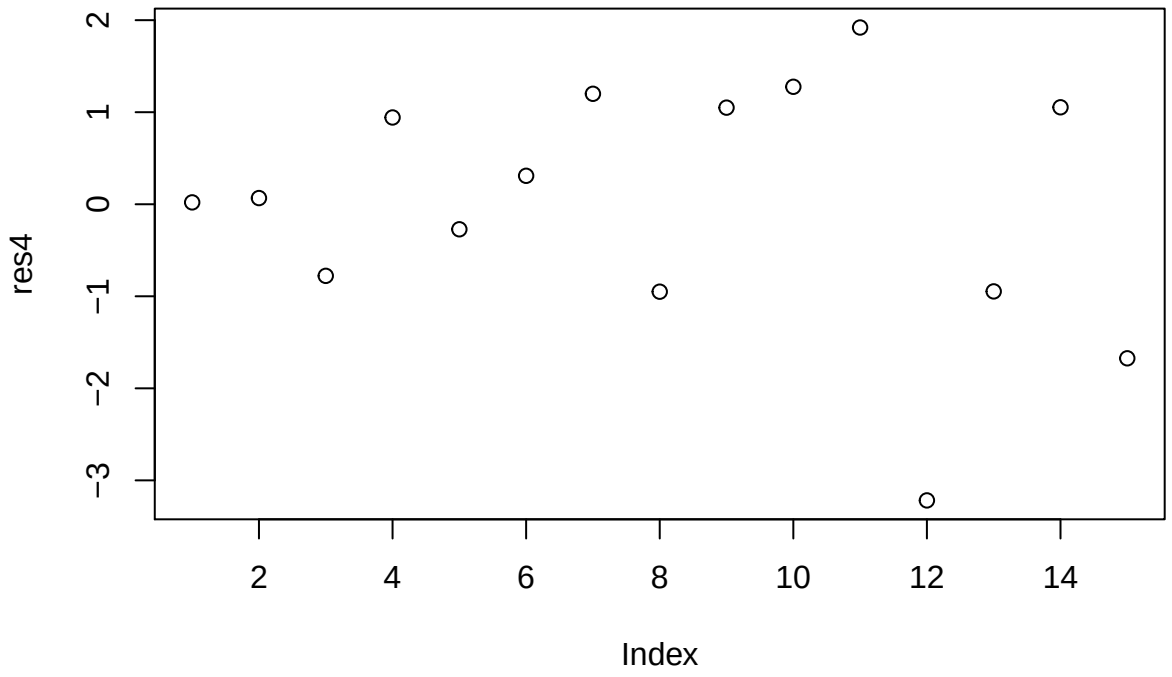


- We can see that the most detectable linear correlation is log(Mass) versus sqrt(Load).

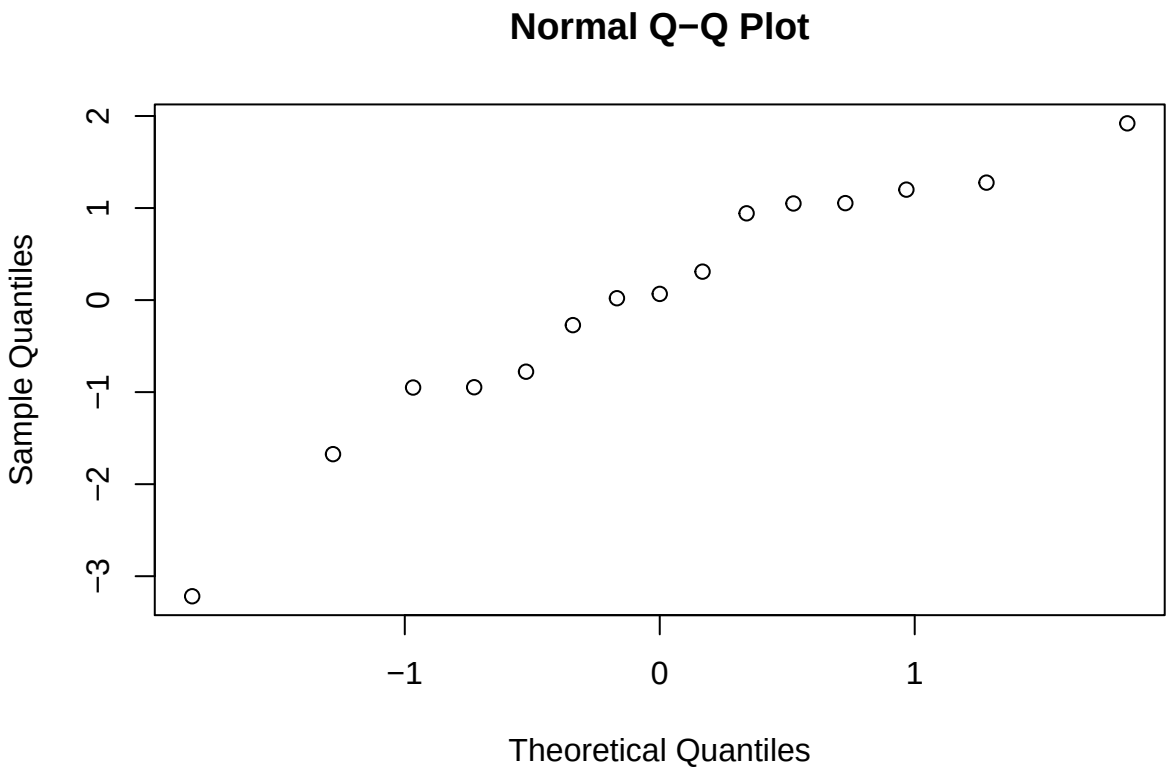
```r
fit4 = lm(log(Mass) ~ sqrt(Load))
fit4
```

```
##
## Call:
## lm(formula = log(Mass) ~ sqrt(Load))
##
## Coefficients:
## (Intercept)   sqrt(Load)
##      3.7965      -0.2621
```

```r
res4 = fit4$residuals
fitted4 = fit4$fitted.values
plot(res4)
```




```r
qqnorm(res4)
```




```
detach(ex0817)
```

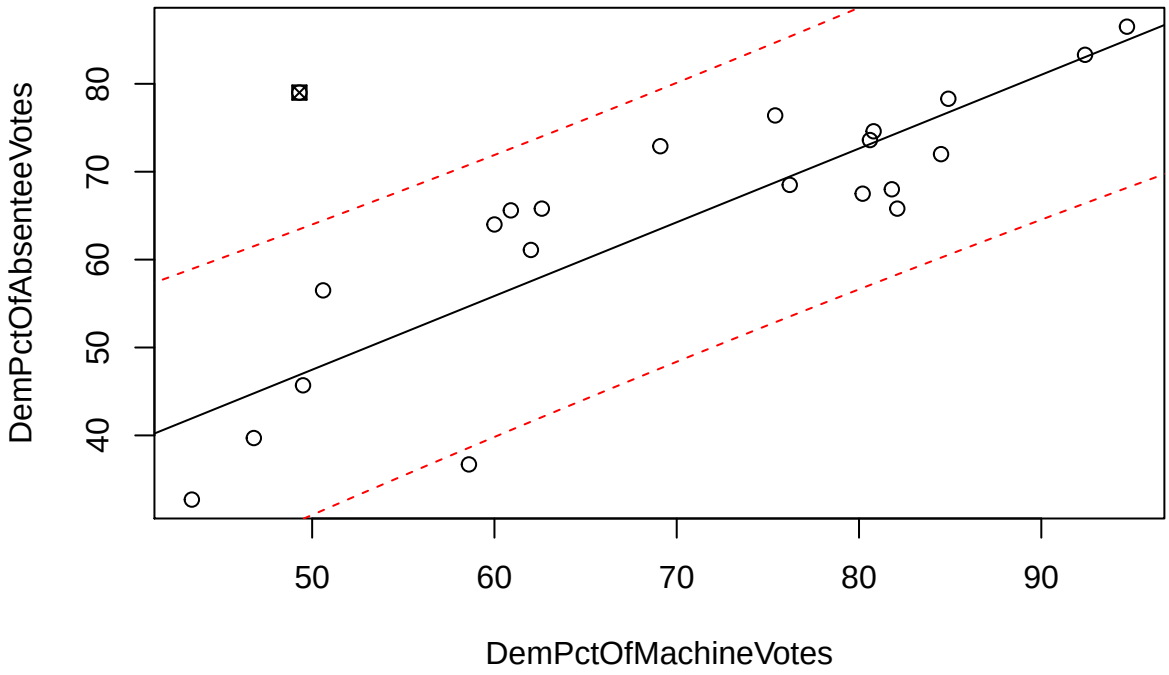
- From the plot, we can see the residuals are pretty random and is near to a normal distribution.

5. Chapter 8, problem 20

```
attach(ex0820)
```

```
## The following object is masked from ex0724:
##
##     Year
```

```
plot(DemPctOfMachineVotes, DemPctOfAbsenteeVotes)
y = ex0820[Year==93, "DemPctOfAbsenteeVotes"]
x = ex0820[Year==93, "DemPctOfMachineVotes"]
points(x, y, pch=7)
x = ex0820[Year!=93, "DemPctOfMachineVotes"]
y = ex0820[Year!=93, "DemPctOfAbsenteeVotes"]
fit8 = lm(y~x)
abline(fit8)
newx = seq(0,100)
prd = predict(fit8,newdata = data.frame(x=newx),interval = "prediction",type = "response")
lines(newx,prd[,2],col="red",lty=2)
lines(newx,prd[,3],col="red",lty=2)
```



- We can see the absentee percentage in teh disputed election is clearly out of the 95 prediction band.

```
prd = predict(fit8,data.frame(x=49.3),se.fit = TRUE,interval = "prediction")
pred = prd$fit[1]
se = prd$se.fit
sePrd = (prd$fit[1]-prd$fit[2])/qt(0.975,19)
(79-pred)/sePrd
```

```
## [1] 4.057263
```

```
(1-pt(4.057263,19))*2
```

```
## [1] 0.0006722556
```

Prediction is 46.88664. Standard error is 2.788739. Prediction error is 7.915031, which means the real value is 4.057263 time of prediction error from the prediction mean. And the two sided pvalue is 0.0006722556.

```
22*(1-pt(4.057263,19))*2
```

```
## [1] 0.01478962
```

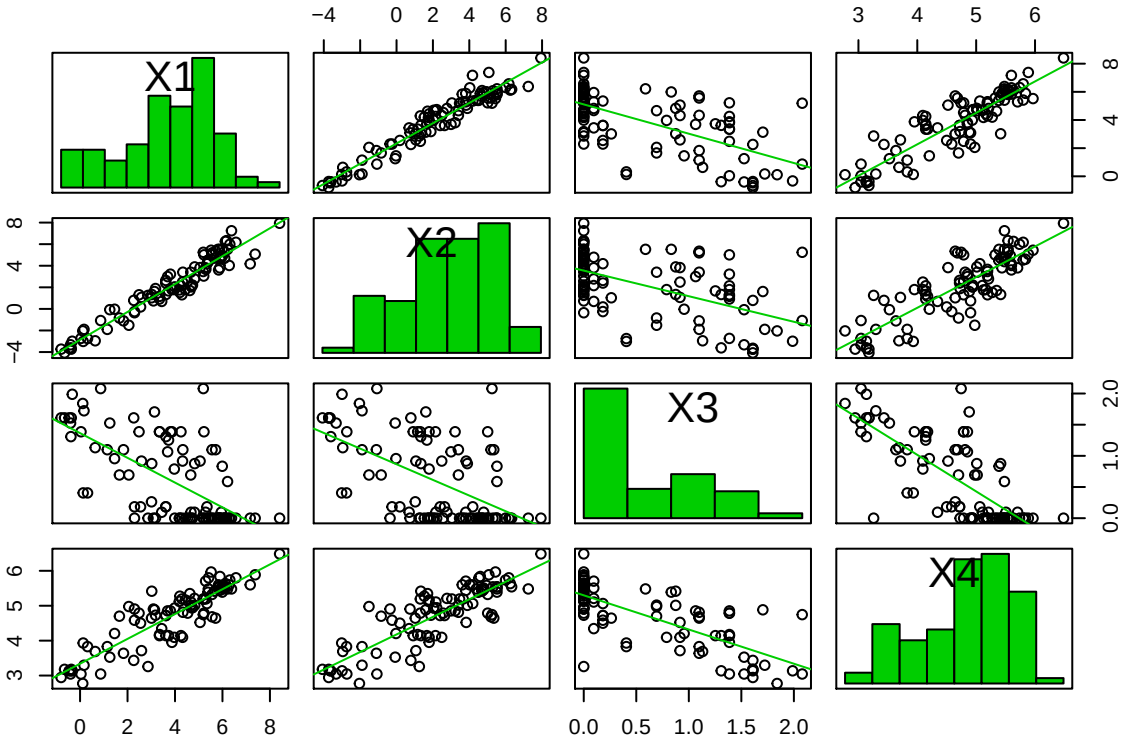
Bonferonni adjusted pValue is 0.01478962.

6. Chapter 9, problem 12

- a

```
attach(case0902)
myMatrix=cbind(log(Brain), log(Body), log(Litter), log(Gestation))
    if(require(car)){
    scatterplotMatrix(myMatrix,
      smooth=FALSE,
      diagonal="histogram")
      }
```

```
## Loading required package: car
```



- b

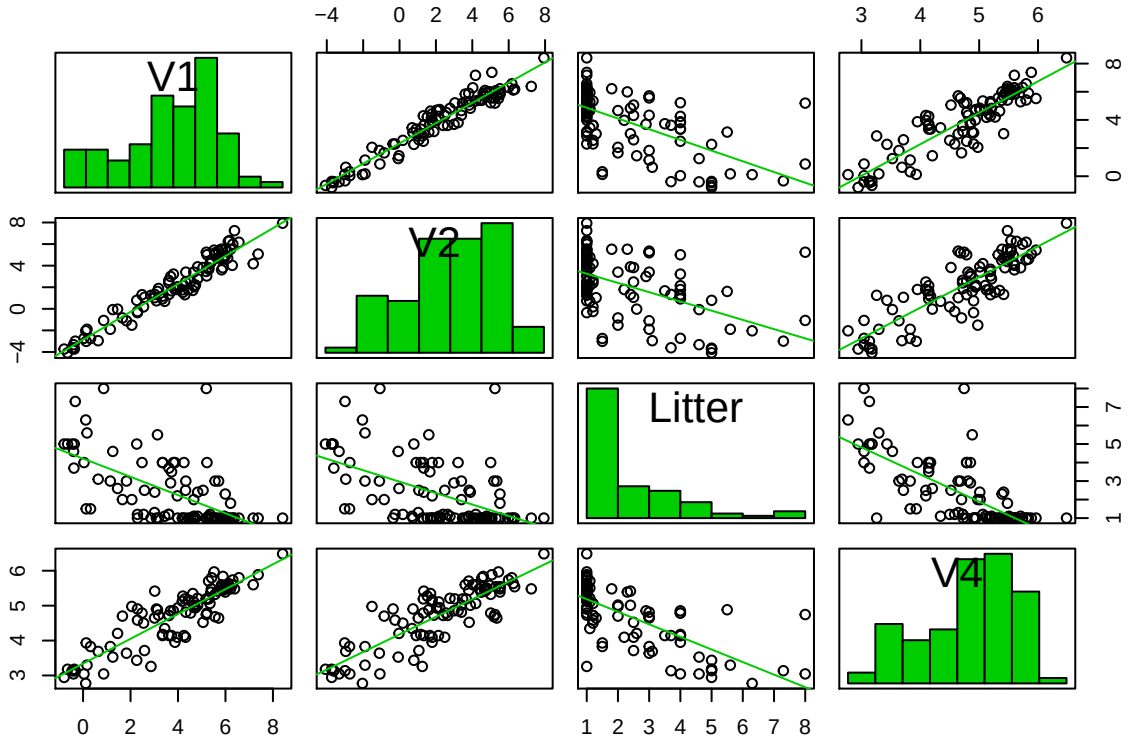
```
lm(log(Brain)~log(Body)+log(Litter)+log(Gestation))
```

```
##
## Call:
## lm(formula = log(Brain) ~ log(Body) + log(Litter) + log(Gestation))
##
## Coefficients:
##    (Intercept)       log(Body)     log(Litter)  log(Gestation)
##         0.8548          0.5751         -0.3101          0.4179
```

*c

```
myMatrix=cbind(log(Brain), log(Body), Litter, log(Gestation))
    if(require(car)){
    scatterplotMatrix(myMatrix,
```

```
      smooth=FALSE,
      diagonal="histogram")
      }
```



There's not much difference between these two, both log and normal scale does fine.