Problem 1

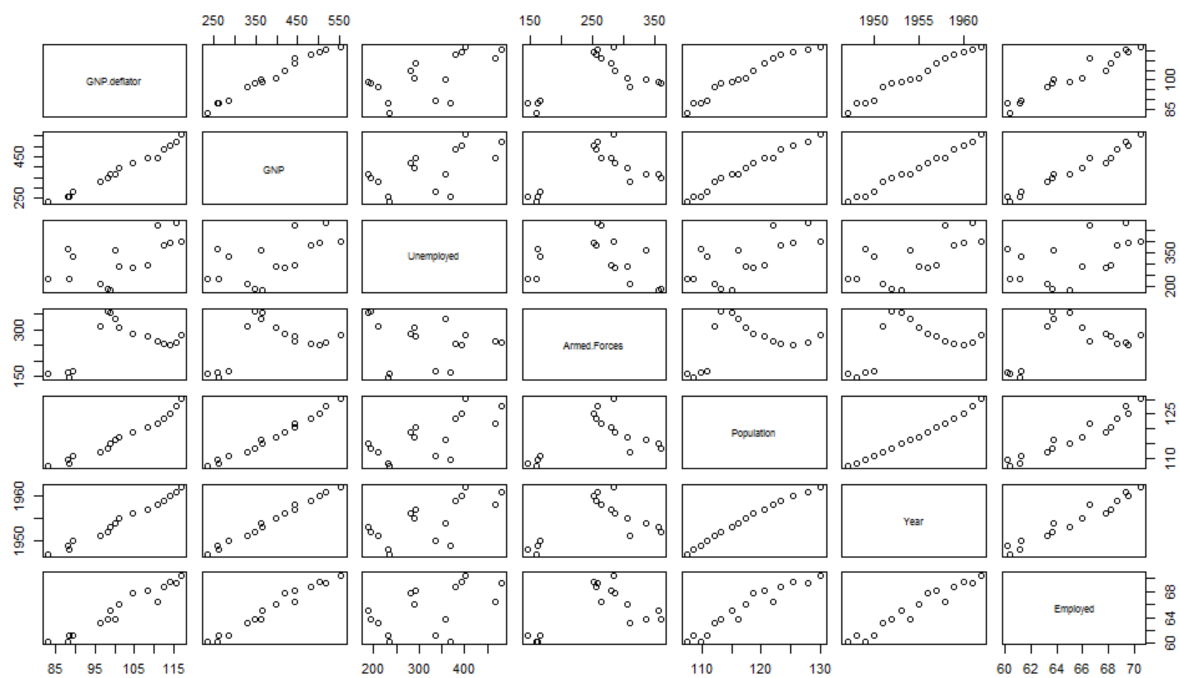
```
> library(faraway)
> data(longley)
> nrow(longley)
[1] 16
> longley[1:3,]
     GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
1947         83.0 234.289      235.6        159.0    107.608 1947   60.323
1948         88.5 259.426      232.5        145.6    108.632 1948   61.122
1949         88.2 258.054      368.2        161.6    109.773 1949   60.171
> attach(longley)
```

(a)

```
> options(digits=3)
> round(cor(longley),2)
             GNP.deflator  GNP Unemployed Armed.Forces Population Year Employed
GNP.deflator         1.00 0.99       0.62         0.46       0.98 0.99     0.97
GNP                  0.99 1.00       0.60         0.45       0.99 1.00     0.98
Unemployed           0.62 0.60       1.00        -0.18       0.69 0.67     0.50
Armed.Forces         0.46 0.45      -0.18         1.00       0.36 0.42     0.46
Population           0.98 0.99       0.69         0.36       1.00 0.99     0.96
Year                 0.99 1.00       0.67         0.42       0.99 1.00     0.97
Employed             0.97 0.98       0.50         0.46       0.96 0.97     1.00
```

(b)

```
> pairs(longley)
```



From the scatter plots and the correlation matrix from part (a), these pairs of variables seems to probably linearly related:

GNP.deflator & GNP (cor=0.99)	**GNP.deflator & Population (cor=0.98)**
GNP.deflator & Year (cor=0.99)	**GNP.deflator & Employed (cor=0.97)**
GNP & Population (cor=0.99)	**GNP & Year (cor=1.00)**
GNP & Employed (cor=0.98)	**Population & Year (cor=0.99)**
Population & Employed (cor=0.96)	**Year & Employed (cor=0.97)**

Especially, "GNP" and "Year" has a correlation 1.00 that indicates highly linear relationship.

(c)

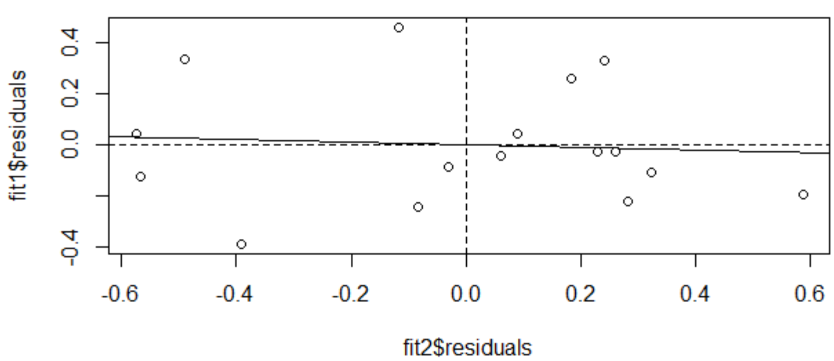
```
> options(digits=5)
> fit = lm(Employed~.,data = longley)
> vif(fit)
GNP.deflator          GNP   Unemployed Armed.Forces   Population         Year

    135.5324    1788.5135      33.6189       3.5889     399.1510     758.9806
```

Variance inflation factor that is greater than 5 is problematic. In this full model, only one VIF of "Armed.Forces" is smaller than 5, and all the others are extremely large, which suggests a huge multicollinearity issue.

(d)

```
> fit1 = lm(Employed~.-Population,data = longley)
> fit2 = lm(Population~.-Employed,data = longley)
> cor(fit2$residuals,fit1$residuals)
[1] -0.075137
> plot(fit2$residuals, fit1$residuals)
> abline(h=0,lty=2)
> abline(v=0,lty=2)
> abline(lm(fit1$residuals ~ fit2$residuals))
```



The partial correlation is very low, and the variable added plot almost shows no linear relationship, thus "Population" should not remain in the full model.

(e)

```
> summary(fit)

Call:
lm(formula = Employed ~ ., data = longley)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4101 -0.1577 -0.0282  0.1016  0.4554 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -3.48e+03   8.90e+02   -3.91  0.00356 ** 
GNP.deflator  1.51e-02   8.49e-02    0.18  0.86314    
GNP          -3.58e-02   3.35e-02   -1.07  0.31268    
Unemployed   -2.02e-02   4.88e-03   -4.14  0.00254 ** 
Armed.Forces -1.03e-02   2.14e-03   -4.82  0.00094 ***
Population   -5.11e-02   2.26e-01   -0.23  0.82621    
Year          1.83e+00   4.55e-01    4.02  0.00304 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.305 on 9 degrees of freedom
Multiple R-squared:  0.995,	Adjusted R-squared:  0.992 
F-statistic:  330 on 6 and 9 DF,  p-value: 4.98e-10
```

In the full model, the predictors “Unemployed”, “Armed.Forces” and “Year” are significant.

Fit a new model with these three predictors.

```
> fit_new = lm(Employed~Unemployed+Armed.Forces+Year)
> vif(fit_new)
  Unemployed Armed.Forces         Year 
      3.3179       2.2233       3.8909 
```

All the variance inflation factors are smaller than 5, thus none of them suggests multicollinearity.

(f)

```
> anova(fit_new,fit)
Analysis of Variance Table

Model 1: Employed ~ Unemployed + Armed.Forces + Year
Model 2: Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces + Populati
on + 
    Year
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     12 1.323                         
2      9 0.836  3     0.487 1.75   0.23
```

F test statistic = 1.75, and p-value = 0.23 > 0.1.

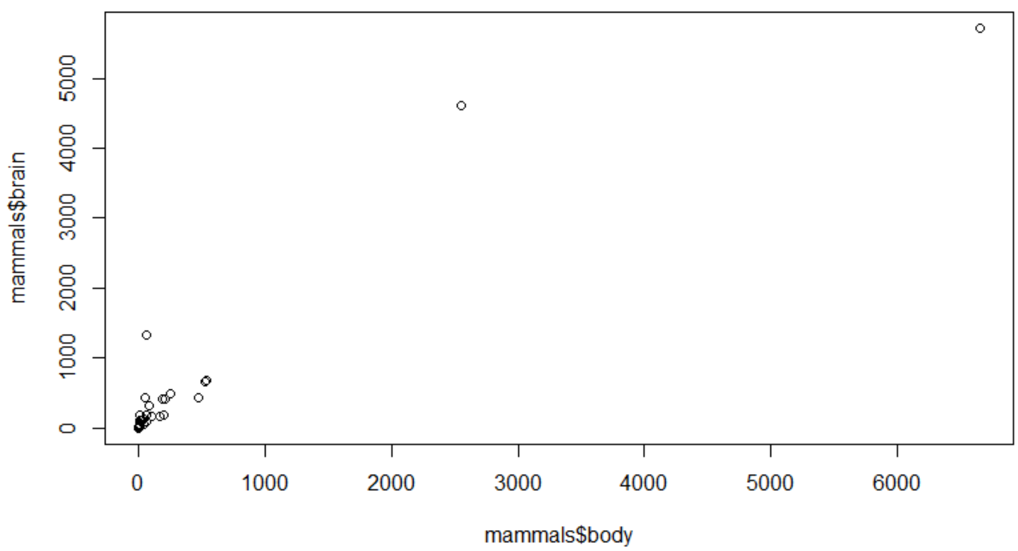
Therefore, Do not Reject Ho (Null Model) at α = 10% or smaller significance level. The new, smaller model is preferred, which is only explained by predictors “Unemployed”, “Armed.Forces” and “Year”.

Problem 2

```
> library(MASS)
> data(mammals)
> nrow(mammals)
[1] 62
> mammals[1:3,]
                 body brain
Arctic fox      3.385  44.5
Owl monkey      0.480  15.5
Mountain beaver 1.350   8.1
> attach(mammals)
```

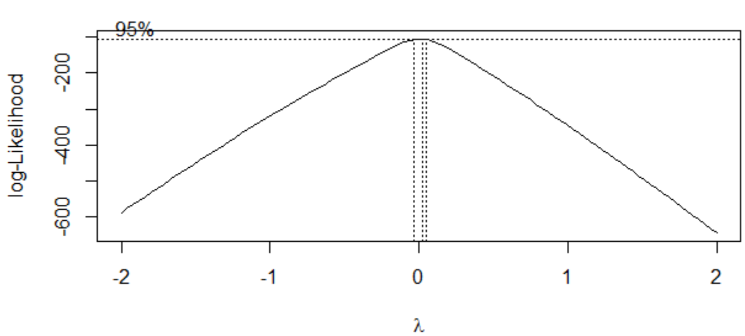
(a)

```
> plot(mammals$body,mammals$brain)
```



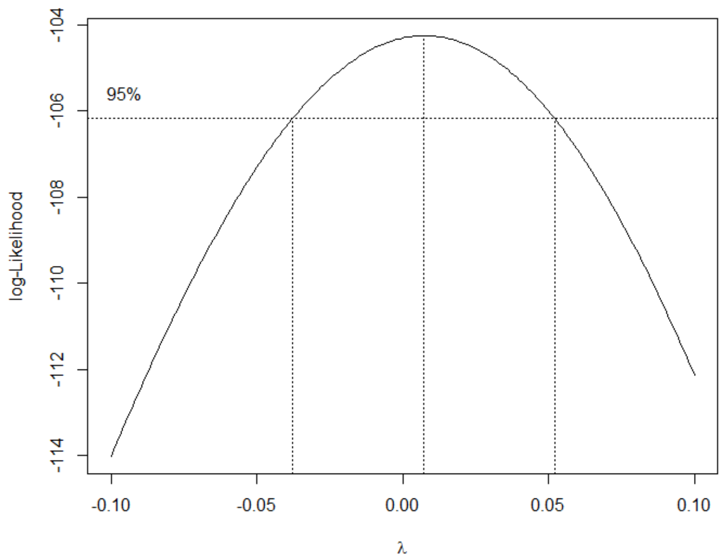
(b)

```
> fit = lm(brain~log(body),data = mammals)
> boxcox(fit,plotit = TRUE)
```



Then adjust the zoom.

```
> boxcox(fit,plotit = TRUE,lambda = seq(-0.1,0.1,by=0.001))
```

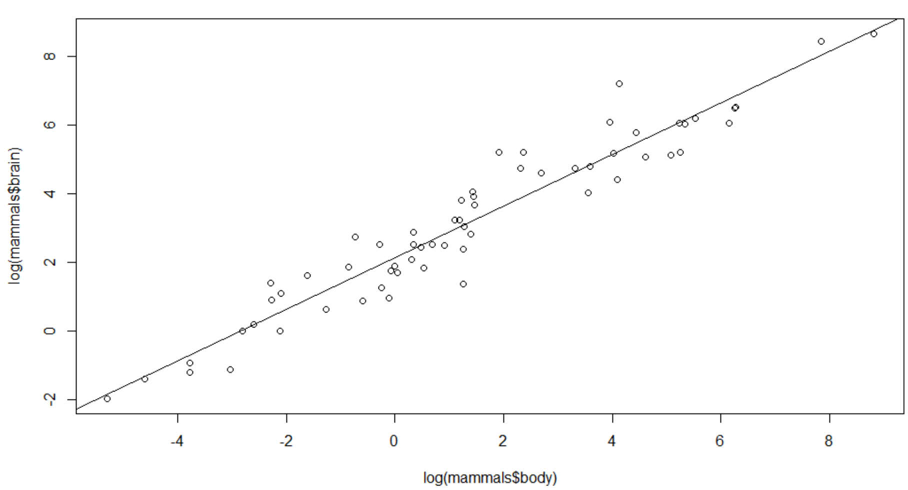


The Box-Cox result shows that λ = 0 is covered in the 95% confidence interval.

Therefore, log(brain weight) is the appropriate transformation of the response variable.

(c)

```
> plot(log(mammals$body),log(mammals$brain))
> fit_log = lm(log(brain)~log(body),data = mammals)
> abline(fit_log)
```



The plot of this new model suggests that this transformed linear model is appropriate.

```
> new=data.frame(body=254)
> predict.lm(fit_log,new,interval=c("prediction"),level=0.95)
     fit    lwr    upr
1 6.2971 4.8769 7.7174
> exp(predict.lm(fit_log,new,interval=c("prediction"),level=0.95))
     fit    lwr  upr
1 543.01 131.22 2247
```

The prediction of the average brain weight of a Siberian tiger is 543.01 g, of which the average body weight is 254 kg.

And the 95% prediction interval of its average brain weight is (131.22, 2247) g.