STAT 343 Final Exam

Siqi Wang

December 10th, 2015

Problem.1

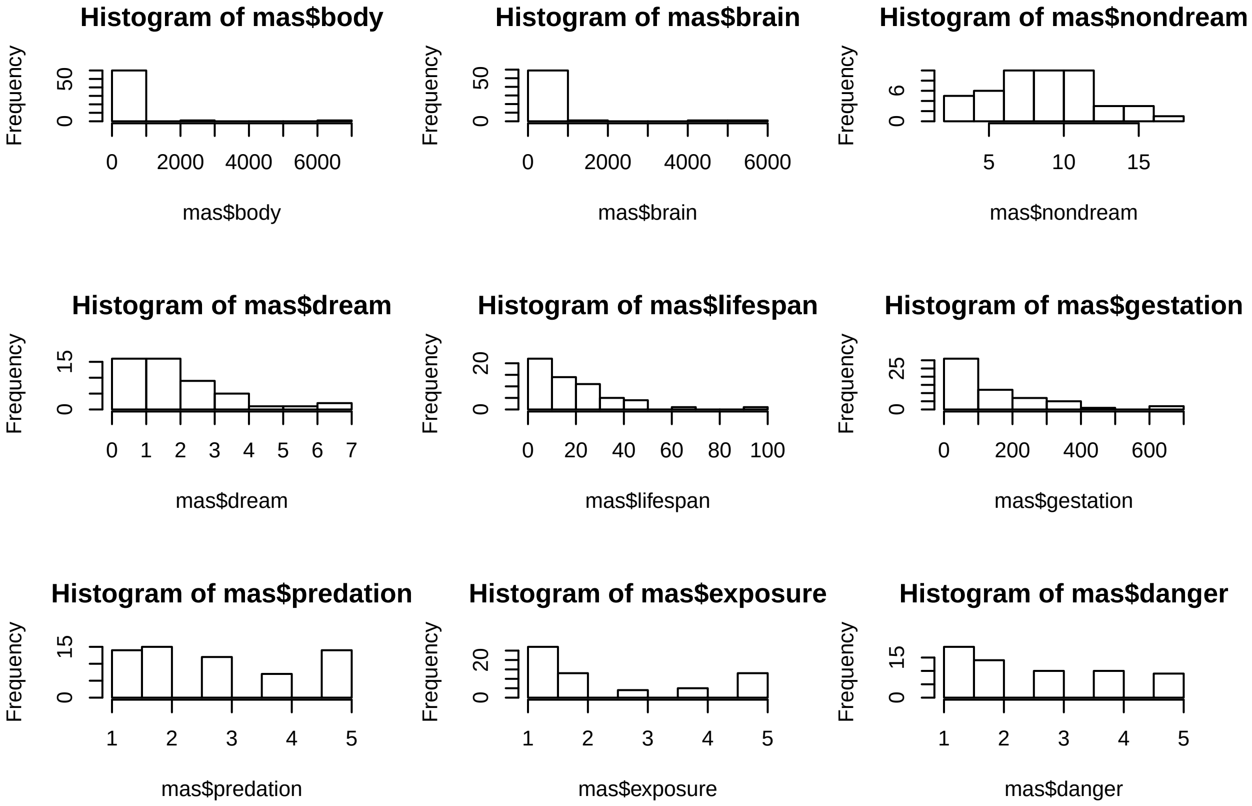
-Information about the dataset

The data set includes 62 observations and 10 variables. Each row has denoted by the species of animal. Here is a brief introduction of each variable:

1.body: body weight in kg
2.brain: brain weight in g
3.nondream: slow wave ("nondreaming") sleep (hrs/day)
4.dream: paradoxical ("dreaming") sleep (hrs/day)
5.sleep: total sleep (hrs/day) (sum of slow wave and paradoxical sleep)
6.lifespan: maximum life span (years)
7.gestation: gestation time (days)
8.predation: predation index (1-5) - This variable indicates the extent to which the species are preyed upon
* 1= minimum (least likely to be preyed upon)
* 5= maximum (most likely to be preyed upon)
9.exposure: sleep exposure index (1-5)
* 1 = least exposed (e.g. animal sleeps in a well-protected den)
* 5 = most exposed
10.danger: overall danger index (1-5) (based on the above two indices and other information)
* 1 = least danger (from other animals)
* 5 = most danger (from other animals)

-Univariate analysis

In this project, since we want to look at the effect of the factors on slow-wave sleep and paradoxical sleep separately, so we are interested in variables nondream and dream instead of sleep. Although danger is a index which involves the influence of both predation and exposure and it may create multicolinearity issue if we include all three of them in our model. Since danger also contains other factors, we will still use it in the full model. In our data, predation, exposure and danger are discrete. However the value of these variables represents the level of a certain condition (e.g the animal is the most vulnerable one if predation=5).Thus we should not treat them as categorical variables.



Now if we look at the histogram of each variable, we can find that there are some extreme values existed in body, and brain. The spread of nondream is approximately symmetric. The spread of dream, lifespan and gestation are skewed to the right.

-Dealing with missing data

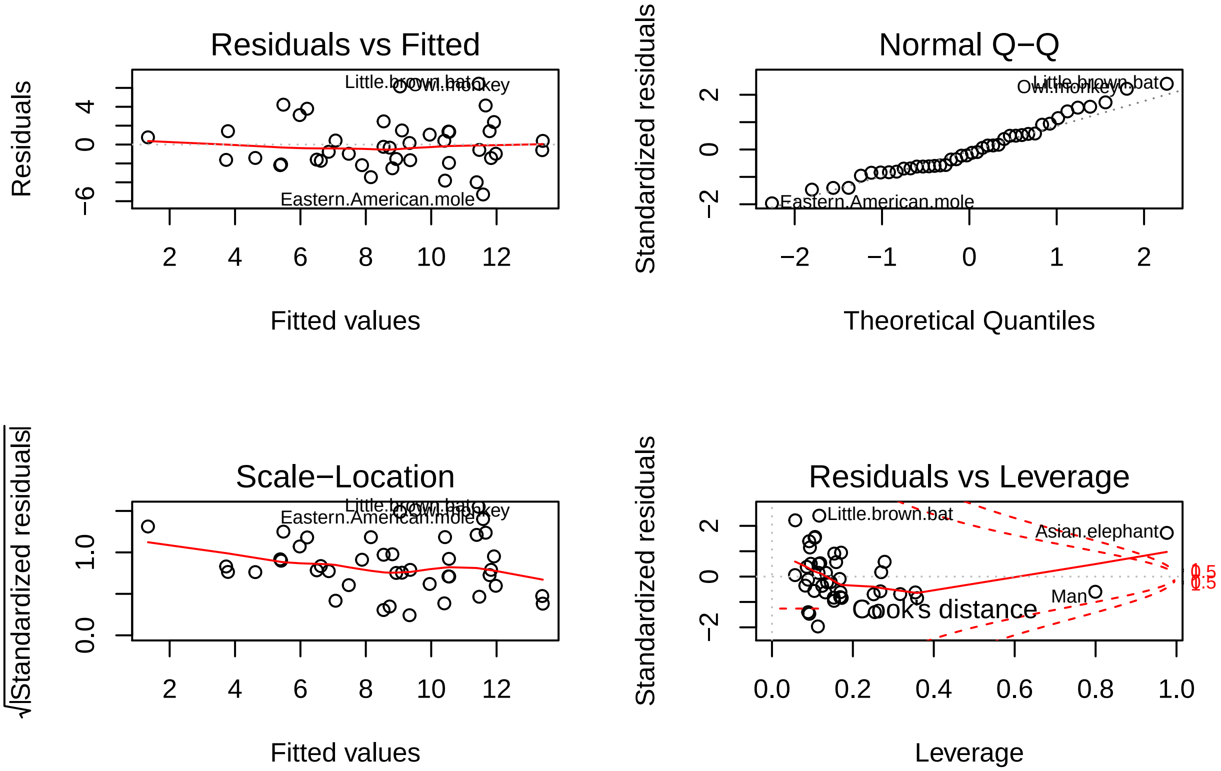
```
##       body              brain            nondream          dream      
##  Min.   :   0.005   Min.   :   0.14   Min.   : 2.100   Min.   :0.000  
##  1st Qu.:   0.600   1st Qu.:   4.25   1st Qu.: 6.250   1st Qu.:0.900  
##  Median :   3.342   Median :  17.25   Median : 8.350   Median :1.800  
##  Mean   : 198.790   Mean   : 283.13   Mean   : 8.673   Mean   :1.972  
##  3rd Qu.:  48.203   3rd Qu.: 166.00   3rd Qu.:11.000   3rd Qu.:2.550  
##  Max.   :6654.000   Max.   :5712.00   Max.   :17.900   Max.   :6.600  
##                                       NA's   :14       NA's   :12     
##      sleep          lifespan         gestation        predation    
##  Min.   : 2.60   Min.   :  2.000   Min.   : 12.00   Min.   :1.000  
##  1st Qu.: 8.05   1st Qu.:  6.625   1st Qu.: 35.75   1st Qu.:2.000  
##  Median :10.45   Median : 15.100   Median : 79.00   Median :3.000  
##  Mean   :10.53   Mean   : 19.878   Mean   :142.35   Mean   :2.871  
##  3rd Qu.:13.20   3rd Qu.: 27.750   3rd Qu.:207.50   3rd Qu.:4.000  
##  Max.   :19.90   Max.   :100.000   Max.   :645.00   Max.   :5.000  
##  NA's   :4       NA's   :4         NA's   :4                       
##     exposure         danger     
##  Min.   :1.000   Min.   :1.000  
##  1st Qu.:1.000   1st Qu.:1.000  
##  Median :2.000   Median :2.000  
##  Mean   :2.419   Mean   :2.613
##  3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :5.000   Max.   :5.000
##
```

We need to deal with missing data first. There are 14 missing value in nondream, 12 in dream, 4 in lifespan and 4 in gestation. Since nondream and dream would be the response variables in two models, we need to delete rows that contain missing value of them (when nondream is the response, delete those 14 rows that contain missing value of nondream; when dream is the response, delete those 12 rows that contain missing value of dream).
To deal with missing values in explanatory variables, we may think about deletion and imputation by regression (multiple imputation is too complicated; since the ranges of lifespan and gestation are very wide, mean imputation is not proper). Regression imputation is only available when there is only one missing data in the row. However, there is one row (Desert.hedgehog) that both the values of lifespan and gestation are missing. Thus we will only impute three values for gestation and two values for lifespan (although the lifespan of Arctic.ground.squirrel is missing, this row will be deleted since the response value is also missing). The limited number of imputation would not cause much difference from using deletion to our data. Besides, single impuation may involve bias to the model. Therefore I chose to use deletion. Now, we will have 42 data entries when the response variable is nondream and 40 data entries when the response is dream.

-Fit the full model

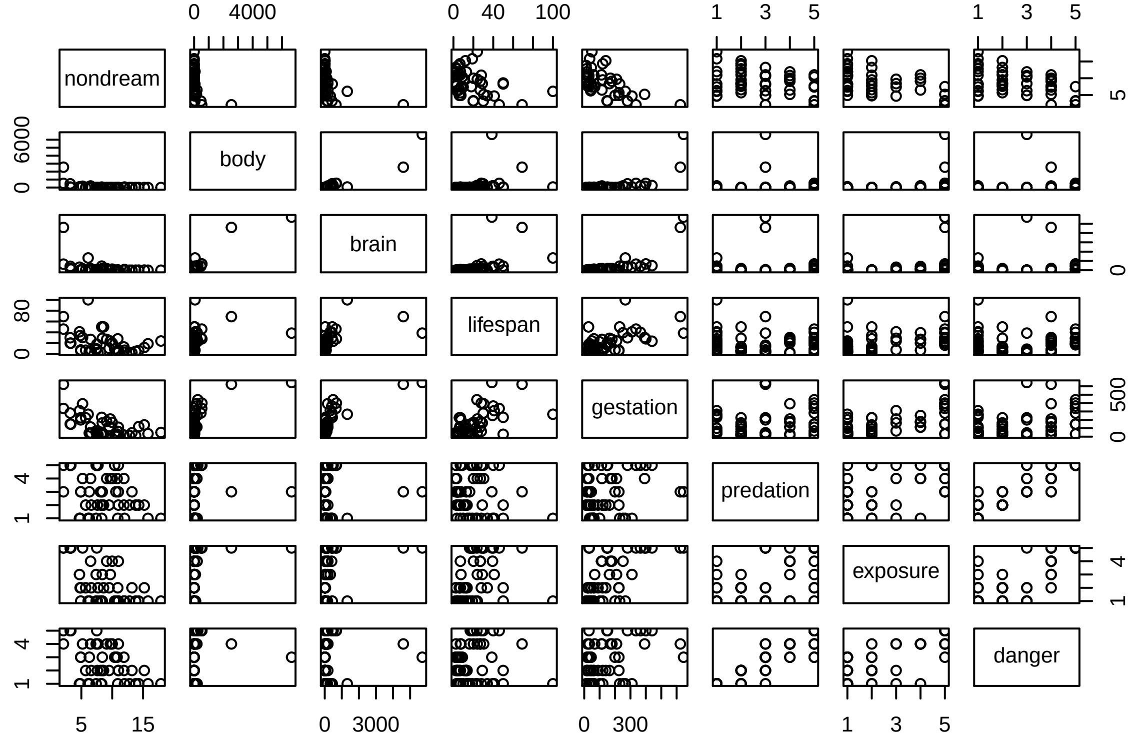
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

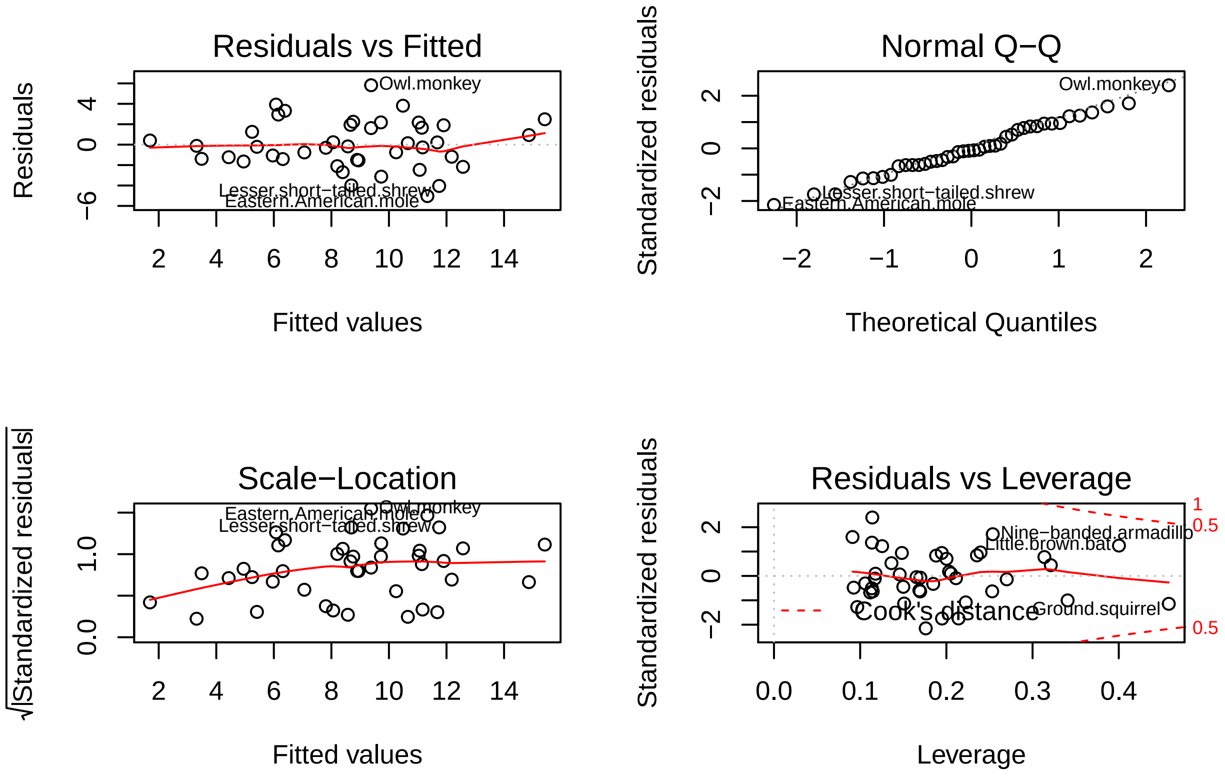


We will do regression with all the explanatory variables first. In this case, the model for slow-wave sleep is: $nondream = 13.33 + 0.003 body - 0.001 brain - 0.001 lifespan - 0.01 gestation + 1.41 predation + 0.22 exposure - 2.80 danger$. The residual standard error(SE) is 2.857 and the adjusted R^2 is 0.4458. Predictors gestation and danger are statistically significant on 5% level.

Now we look at the diagnostic plots. The residual plots shows that the spread of the points slightly tends to be wider as fitted values increases. This situation indicates that the constant variance assumption may be violated. Besides, the Q-Q plot shows that the graph is a bit skewed. The normality assumption may also be violated. Some transformation may be needed here.



We may look at the relationship between nondream and each variable from the scatter matrix plot. We can find that for body, brain, lifespan and gestation, there is a cluster of points on the plot and some points are spread out. We may need to transform these four predictors by log to solve this problem.

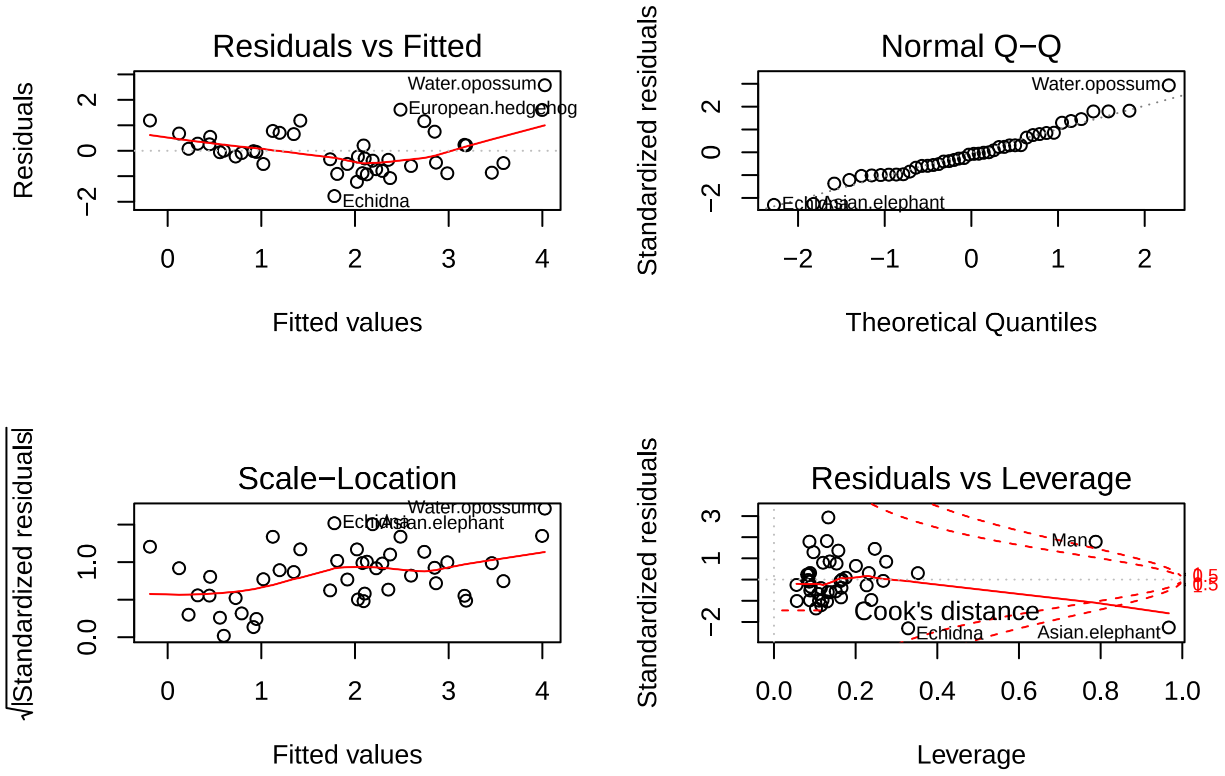


After the transformation, the residual plot has no obvious pattern and the Q-Q norm plot has a nearly straight line fit. The transformation seems valid. There are points that may need to care (such as owl.monkey). However, we will check the influential points and outliers later.

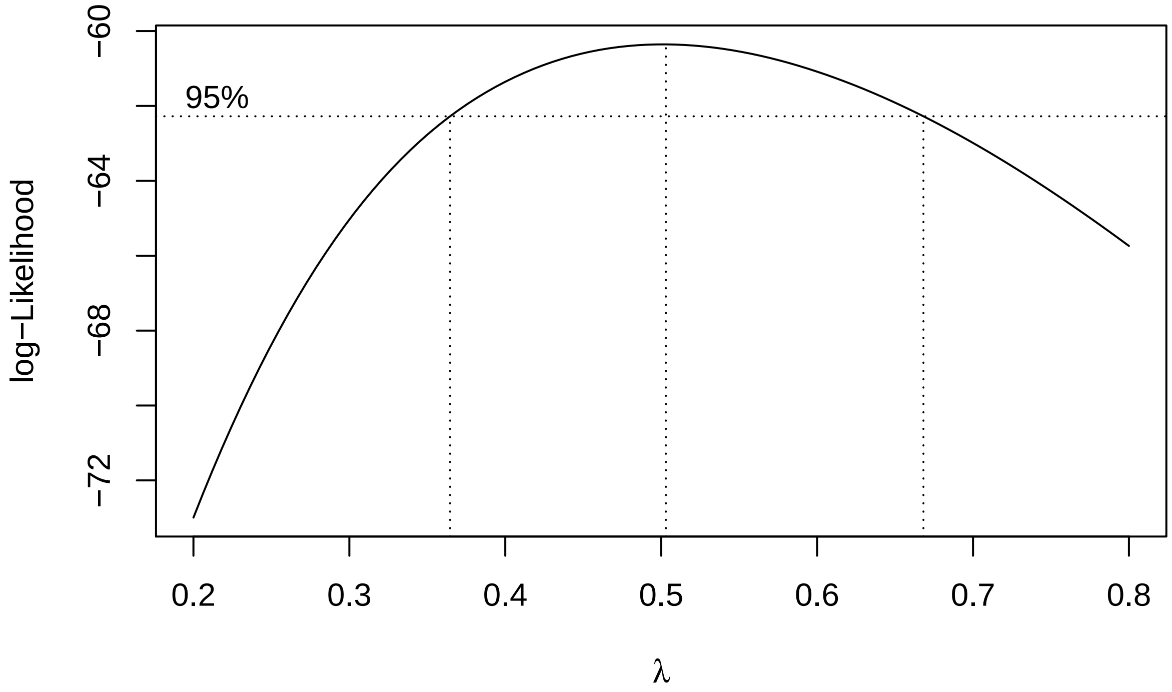
The full model of dream is $dream = 3.66 + 0.001 body + 0.0004 brain - 0.017 lifespan - 0.003 gestation + 0.926 predation + 0.35 exposure - 1.74 danger$. The residual standard error(SE) is 0.9428 and the adjusted R^2 is 0.5356. Predictors predation and danger are statistically significant on 5% level. If we look at the residual plot, we can see a non-linear pattern there. Besides, the Q-Q norm plot seems a bit skewed. The constant variance and normality assumptions may be violated. The result from diagnostic plots implies that transformation of the response variable may needed.

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```

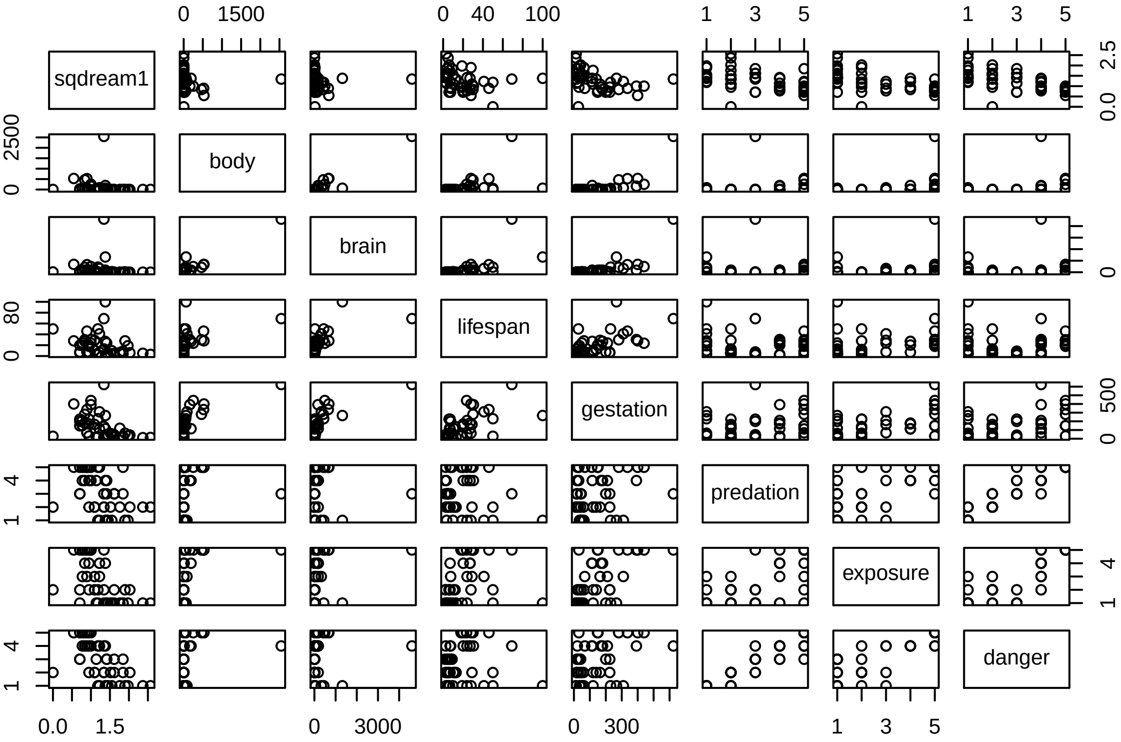
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



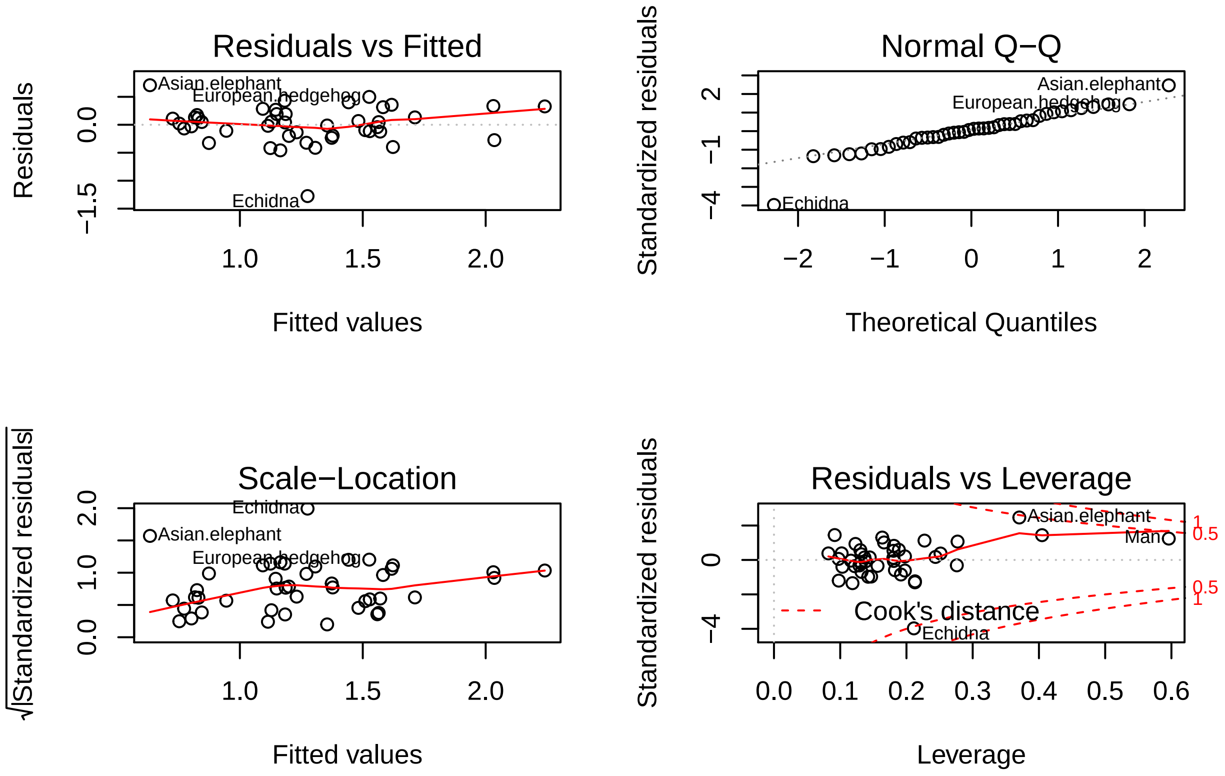
We use the Box-cox method to look for the information about the transformation. We can see from the plot that 1 is not in the 95% confidence interval of λ which means a transformation of dream is necessary and 0.5 is the value near the center of the confidence interval. Thus we will square root the response.



Now if we look at the relationship between each predictor and the transformed response from the matrix scatter plot, we will find that most of the data points of body and brain are gathered together. Thus we may still need to log transform body and brain.



From the diagnostic plots we can find that the curvature pattern in the residual plot is almost disappear and the Q-Q plot has a nearly straight line. There are some points (Aisan.elephant and Man) near the cook's distance curve that we may need to pay attention and the point Echidna may be an outlier candidate. We will investigate these points later. To sum up, the transformation is valid here.



-Model Selection

We will compare the results of the backward elimination, AIC, adjusted R^2 and Mallow's C_p methods to help us to select the correct model.

```
#backward elimination
fit1p=update(fit1p3,.~.-powerges1n)
fit1p=update(fit1p,.~.-powerbody1n)
fit1p=update(fit1p,.~.-exposure)
summary(fit1p)
```

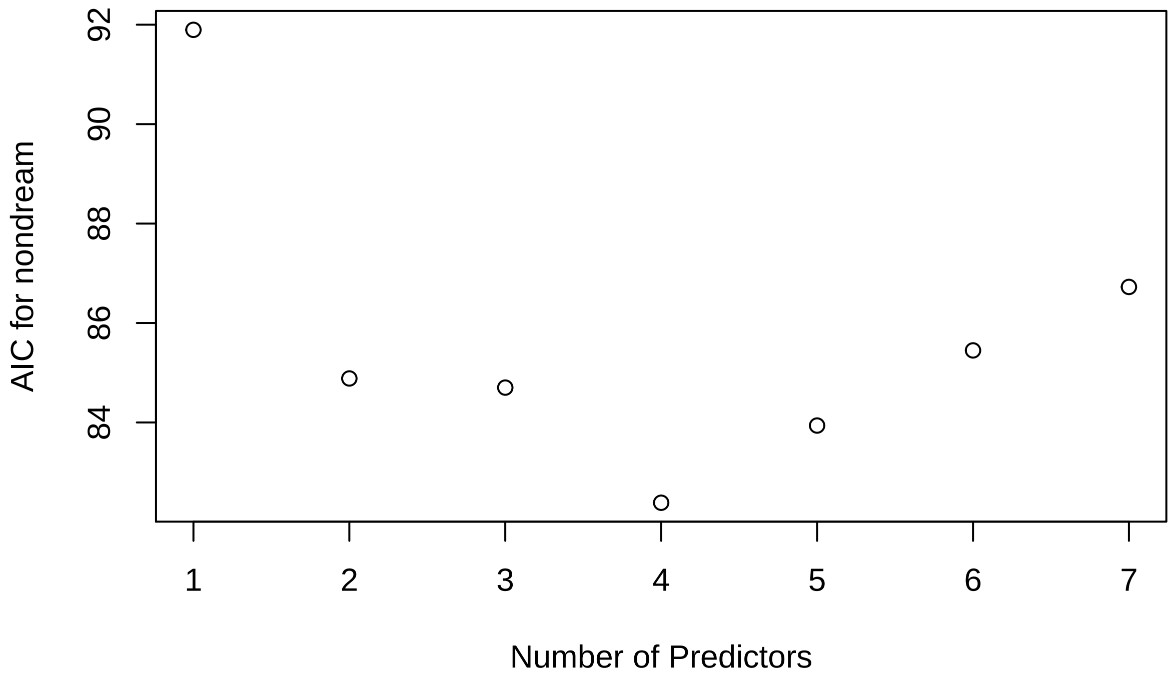
```
##
## Call:
## lm(formula = nondream ~ powerbrain1n + powerlife1n + predation +
##     danger, data = newdata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5813 -1.7883 -0.3657  1.7443  6.1858
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.0958     1.6671   6.056 5.28e-07 ***
## powerbrain1n  -1.1024     0.2716  -4.058 0.000245 ***
## powerlife1n    1.5961     0.6790   2.351 0.024169 *
## predation      1.8299     0.9145   2.001 0.052760 .
## danger        -2.8430     0.9741  -2.919 0.005950 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.522 on 37 degrees of freedom
## Multiple R-squared:  0.6102, Adjusted R-squared:  0.568
## F-statistic: 14.48 on 4 and 37 DF,  p-value: 3.316e-07
```

The backward elimination selects the nondream~log(brain)+log(lifespan)+predation+danger model. All predictors are significant.

```
aex1=regsubsets(nondream~powerbody1n+powerbrain1n+powerlife1n+powerges1n+predation+exposure+danger,data
rs1=summary(aex1)
rs1$which
```

```
##   (Intercept) powerbody1n powerbrain1n powerlife1n powerges1n predation
## 1        TRUE        TRUE        FALSE       FALSE      FALSE     FALSE
## 2        TRUE        TRUE        FALSE       FALSE      FALSE     FALSE
## 3        TRUE       FALSE         TRUE        TRUE      FALSE     FALSE
## 4        TRUE       FALSE         TRUE        TRUE      FALSE      TRUE
## 5        TRUE       FALSE         TRUE        TRUE      FALSE      TRUE
## 6        TRUE        TRUE         TRUE        TRUE      FALSE      TRUE
## 7        TRUE        TRUE         TRUE        TRUE       TRUE      TRUE
##   exposure danger
## 1    FALSE  FALSE
## 2    FALSE   TRUE
## 3    FALSE   TRUE
## 4    FALSE   TRUE
## 5     TRUE   TRUE
## 6     TRUE   TRUE
## 7     TRUE   TRUE
```

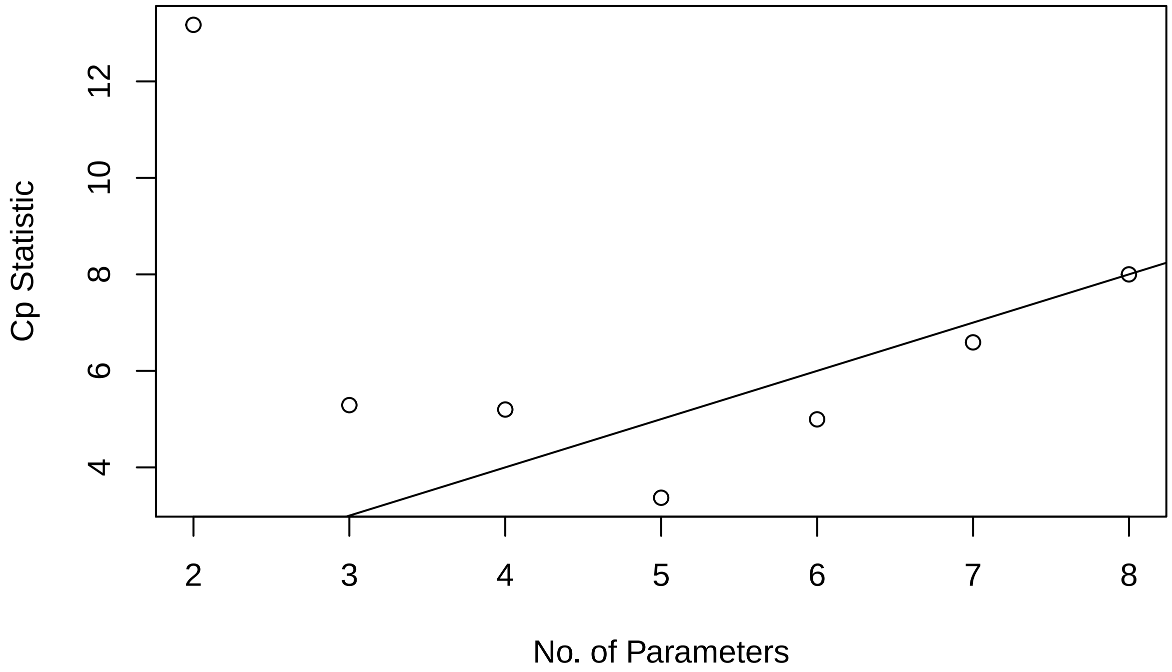
```
par(mfrow=c(1,2))
AIC1 = 42*log(rs1$rss/42) + (2:8)*2
par(mfrow=c(1,1))
plot(AIC1 ~ I(1:7), ylab="AIC for nondream", xlab="Number of Predictors")
```



```
which.max(rs1$adjr2) #model selection by adjr2, 4
```

```
## [1] 4
```

```
plot(2:8,rs1$cp,xlab="No. of Parameters",ylab="Cp Statistic") #Mallow's Cp, 7 parameters
abline(0,1)
```



The 4-predictor model has the smallest AIC. The adjusted R^2 and backward elimination have the same choice with AIC. And the 7-parameter model is chosen by Mallow's C_p method. We find that in the 7-parameter model, besides danger, all the other variables are not significant. The 4-predictor model is $nondream = 10.01 - 1.1log(brain) + 1.6log(life) + 1.83predation - 2.84danger$. All the predictors are significant. The residual standard error(SE) is 2.52 and the adjusted R^2 is 0.57. We may consider this model as the final model.
Now we will look at the model for paradoxical dream.

```
fit2p=update(fit2p1,.~.-gestation)
summary(fit2p)
```

```
##
## Call:
## lm(formula = sqdream1 ~ powerbody1d + powerbrain1d + lifespan +
##     predation + exposure + danger, data = newdata2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26957 -0.15135  0.03028  0.19260  0.69964
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.227934   0.219037  10.171 2.88e-12 ***
## powerbody1d   0.157340   0.061955   2.540  0.01543 *
## powerbrain1d -0.191085   0.084725  -2.255  0.03011 *
## lifespan     -0.002057   0.004371  -0.471  0.64074
## predation     0.286967   0.125500   2.287  0.02804 *
## exposure      0.047283   0.080920   0.584  0.56256
## danger       -0.540901   0.155719  -3.474  0.00133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3566 on 37 degrees of freedom
## Multiple R-squared:  0.5536, Adjusted R-squared:  0.4812
## F-statistic: 7.648 on 6 and 37 DF,  p-value: 2.201e-05
```

```
fit2p=update(fit2p,.~.-lifespan)
summary(fit2p)
```

```
##
## Call:
## lm(formula = sqdream1 ~ powerbody1d + powerbrain1d + predation +
##     exposure + danger, data = newdata2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32037 -0.15704  0.01993  0.19455  0.66449
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.21944    0.21604  10.273 1.61e-12 ***
## powerbody1d   0.16374    0.05982   2.737  0.00938 **
## powerbrain1d -0.21038    0.07337  -2.867  0.00672 **
## predation     0.29623    0.12267   2.415  0.02066 *
## exposure      0.04779    0.08008   0.597  0.55419
## danger       -0.54490    0.15389  -3.541  0.00107 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3529 on 38 degrees of freedom
## Multiple R-squared:  0.5509, Adjusted R-squared:  0.4919
## F-statistic: 9.324 on 5 and 38 DF,  p-value: 7.371e-06
```

```
fit2p=update(fit2p,.~.-exposure)
summary(fit2p)
```

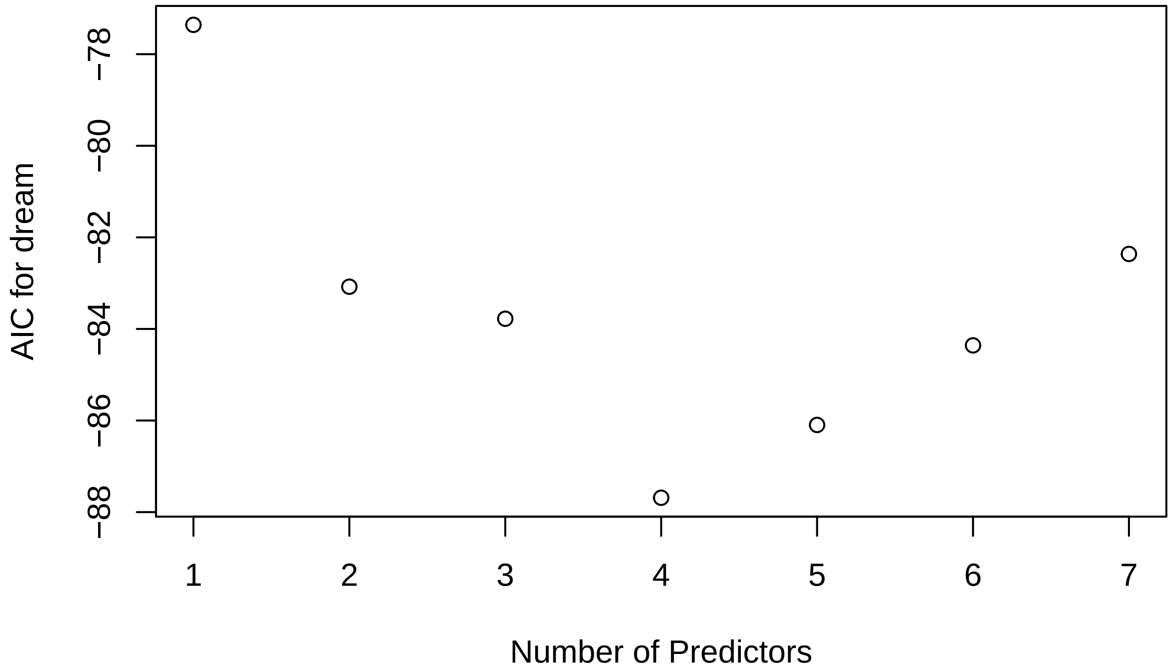
```
##
## Call:
## lm(formula = sqdream1 ~ powerbody1d + powerbrain1d + predation +
##     danger, data = newdata2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.32036 -0.16768  0.01894  0.18718  0.66610
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.22281    0.21418  10.378 8.85e-13 ***
## powerbody1d   0.16842    0.05882   2.863 0.006711 **
## powerbrain1d -0.20495    0.07220  -2.838 0.007161 **
## predation     0.28395    0.11993   2.368 0.022957 *
## danger       -0.49783    0.13104  -3.799 0.000497 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.35 on 39 degrees of freedom
## Multiple R-squared:  0.5467, Adjusted R-squared:  0.5002
## F-statistic: 11.76 on 4 and 39 DF,  p-value: 2.321e-06
```

The backward elimination selects the $dream^0$.5~log(body)+log(brain)+predation+danger model. All predictors are significant.

```
aex2=regsubsets(sqdream1~powerbody1d+powerbrain1d+lifespan+gestation+predation+exposure+danger,data=newd
rs2=summary(aex2)
rs2$which
```

```
##   (Intercept) powerbody1d powerbrain1d lifespan gestation predation
## 1        TRUE       FALSE        FALSE    FALSE     FALSE     FALSE
## 2        TRUE       FALSE        FALSE    FALSE     FALSE      TRUE
## 3        TRUE        TRUE         TRUE    FALSE     FALSE     FALSE
## 4        TRUE        TRUE         TRUE    FALSE     FALSE      TRUE
## 5        TRUE        TRUE         TRUE    FALSE     FALSE      TRUE
## 6        TRUE        TRUE         TRUE     TRUE     FALSE      TRUE
## 7        TRUE        TRUE         TRUE     TRUE      TRUE      TRUE
##   exposure danger
## 1    FALSE   TRUE
## 2    FALSE   TRUE
## 3    FALSE   TRUE
## 4    FALSE   TRUE
## 5     TRUE   TRUE
## 6     TRUE   TRUE
## 7     TRUE   TRUE
```

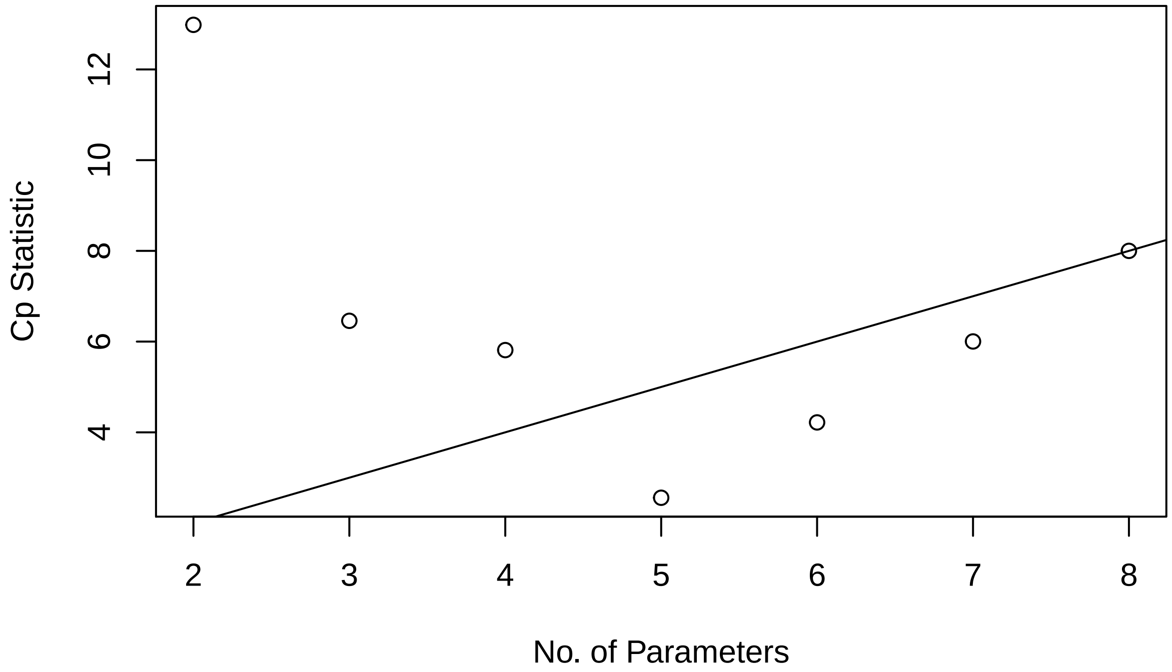
```
AIC2 = 44*log(rs2$rss/44) + (2:8)*2
par(mfrow=c(1,1))
plot(AIC2 ~ I(1:7), ylab="AIC for dream", xlab="Number of Predictors")
```



```
which.max(rs2$adjr2) #model selection by adjr2, 4
```

```
## [1] 4
```

```
plot(2:8,rs2$cp,xlab="No. of Parameters",ylab="Cp Statistic") #Mallow's Cp, 7 parameters
abline(0,1)
```



The 4-predictor model has the smallest AIC. The adjusted R^2 and backward elimination have the same choice with AIC. And the 7-parameter model is chosen by Mallow's C_p method. We find that in the 7-parameter model,lifespan and exposure are not significant parameters. The 4-predictor model is $dream^{0.5} = 2.22 + 0.17log(body) - 0.2log(brain) + 0.28predation - 0.49danger$. All the predictors are significant. The residual standard error(SE) is 0.36 and the adjusted R^2 is 0.48. We may consider this model as the final model.
###-Check Unusual Observation

```
#nondream model
stud=rstudent(fit1p)
stud[which.max(abs(stud))]
```

```
## Owl.monkey
##   2.707087
```

```
qt(.05/(42*2),34) #no influential outlier
```

```
## [1] -3.53774
```

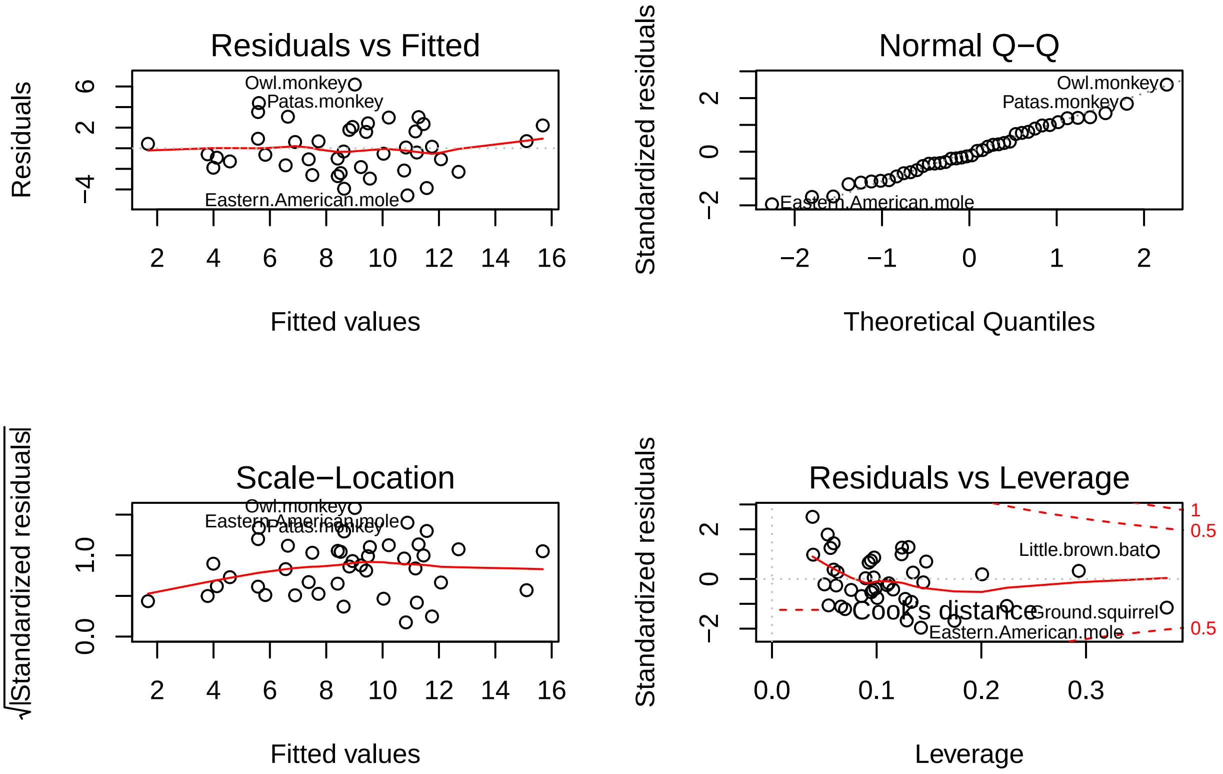
```
#dream model
stud=rstudent(fit2p)
stud[which.max(abs(stud))]
```

```
##   Echidna
## -4.814012
```

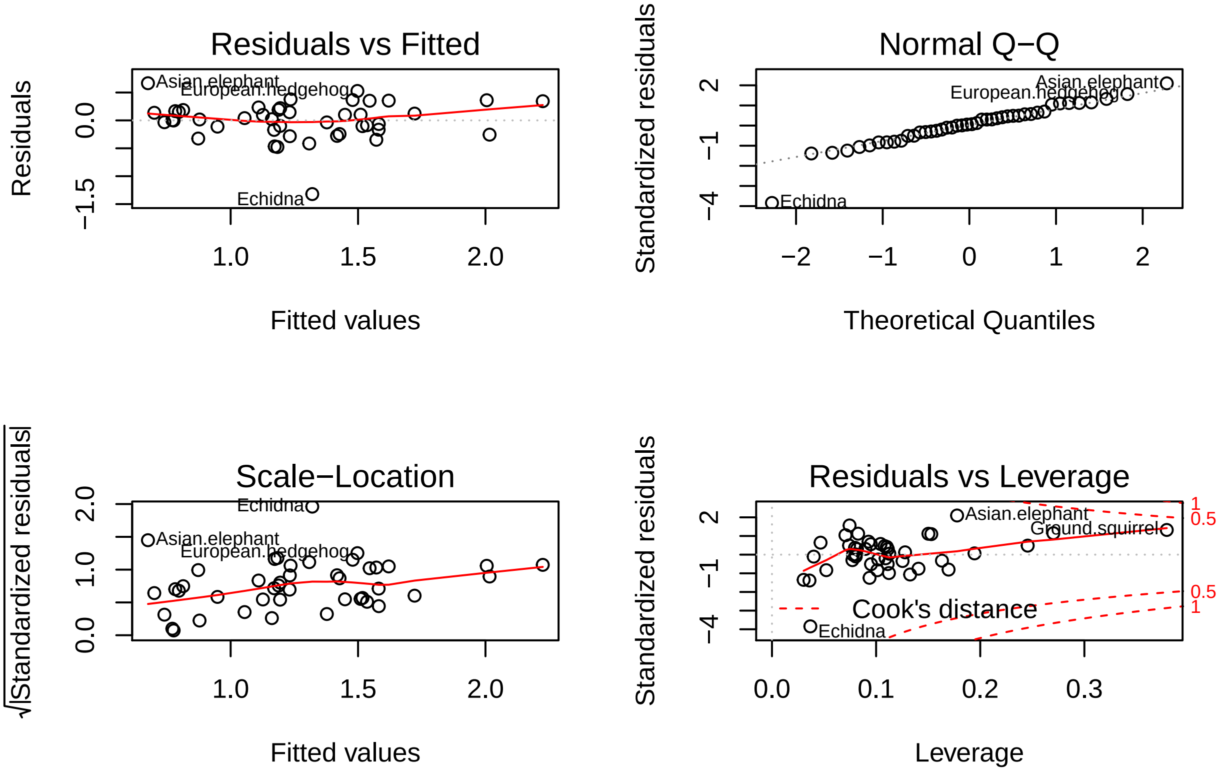
```r
qt(.05/(44*2),36) #one influential outlier, Echidna, line 11
```

```
## [1] -3.536637
```

```r
par(mfrow=c(2,2))
plot(fit1p)
```



```r
plot(fit2p)
```

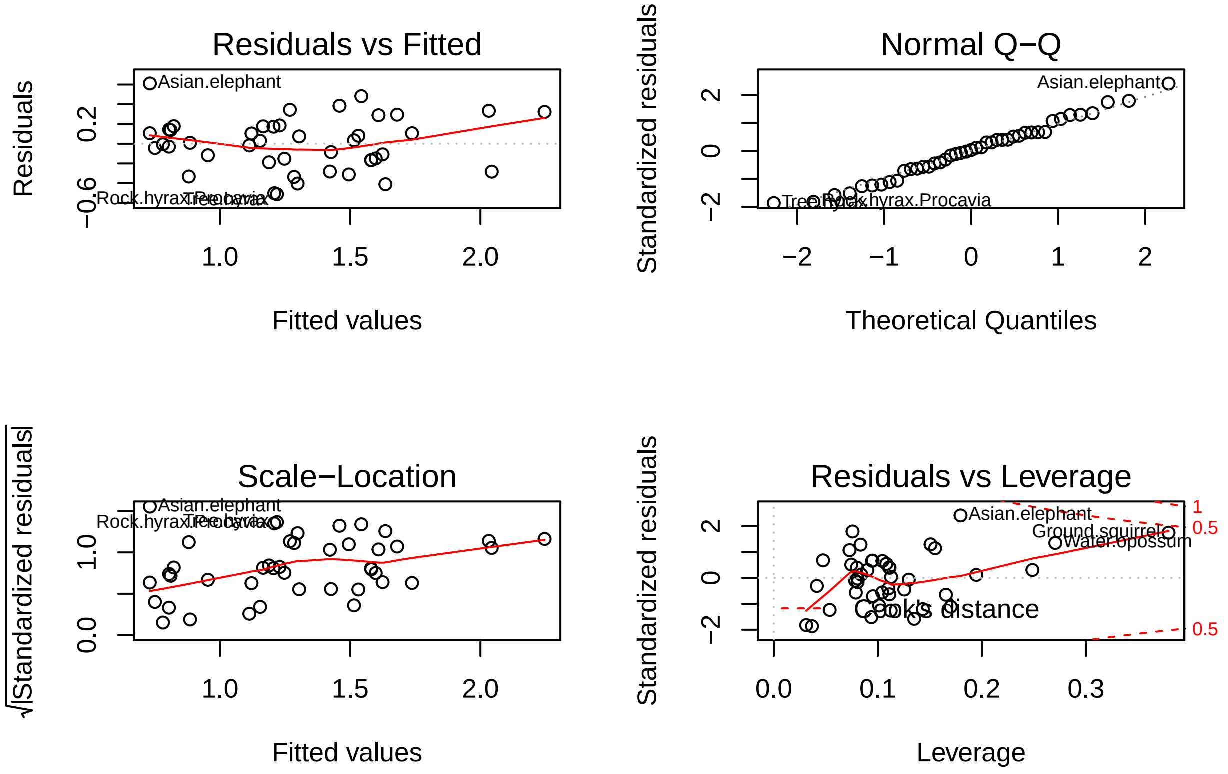


The outlier that we are interested in is the one that would influence the fit. Thus we will use the studentized residual to identify those influential outliers. We can find that in the nondream model, the maximum of studentized residual happens for Owl.donkey, which is 2.71. And it is less than the absolute value of Bonferroni critical value, 3.54. Thus it is not an influential outlier. For the dream model, the studentized residual of Echidna is -4.81. The absolute value of it is bigger than the absolute value of the Bonferroni critical value, 3.54. Thus Echidna is an influential outlier (I have checked that Echidna is the only influential outlier for the dream model. The second biggest studentized residual is less than 3.54).

If we look at the Residual VS Leverage plot in the diagnostic plots we can find that neither of the models has point that lie beyond the cook's distance curve. So there is no leverage points in two models.

```
##         body brain nondream dream sleep lifespan gestation predation
## Echidna    3    25      8.6     0   8.6       50        28         2
##         exposure danger
## Echidna        2      2
```

If we look at the data of Echidna, we will find that the time of its dream value is 0hr/day. This is the only row that has 0 as the dream value. Maybe this is an error. Even if it is not an error, this situation is not useful to be included in our case. Thus I decide to remove this outlier.



After removing the outlier, we can see from the diagnostic plots that the model does not change too much. The coefficient of each parameter has not big change.

-Final model and interpretation

Our model for slow-wave sleep is $nondream = 10.1 - 1.1 log(brain) + 1.6 log(life) + 1.83 predation - 2.84 danger$. The residual standard error(SE) is 2.52 and the adjusted R^2 is 0.57. We can see that every 1 percent increase of brain weight will decrease about 1 hour slow-wave sleep per day. Every 1 percent increase of life span will increase about 1.6 hours slow-wave sleep per day. If the animal is one unit more likely to be preyed-upon, the slow-wave sleep will increase by about 1.8 hours per day. If the environment of animal is one unit more danger, the slow-wave sleep will decrease by 2.84 hours per day. For slow-wave sleep, the size of brain is more influential. Besides, predator may have less slow-wave sleep compared to preys. The place that the animal sleeps may not affect the length of slow-wave sleep too much but other danger factors will decrease the hour slow-wave sleep.

Our model for paradoxical sleep is $dream^{0.5} = 2.29 + 0.16 log(body) - 0.2 log(brain) + 0.25 predation - 0.48 danger$. All the predictors are significant. The residual standard error(SE) is 0.28 and the adjusted R^2 is 0.63. Since we square-root transformed the response, the interpretation is hard. However we can see that the body size has a positive relationship with paradoxical sleep time but the brain size has a negative relationship with paradoxical sleep time. A predator would have more time for paradoxical sleep and other dangerous factors would decrease the paradoxical sleep time. Again the sleeping location does not matter too much to this response variable.

In general, animal's brain size is negatively related to both nondream time and dream time. Predators would have more sleep time in total than preys would. Sleep location do not affect the sleeping time so much. And dangerous factors would decrease both the nondream and dream sleeping times.

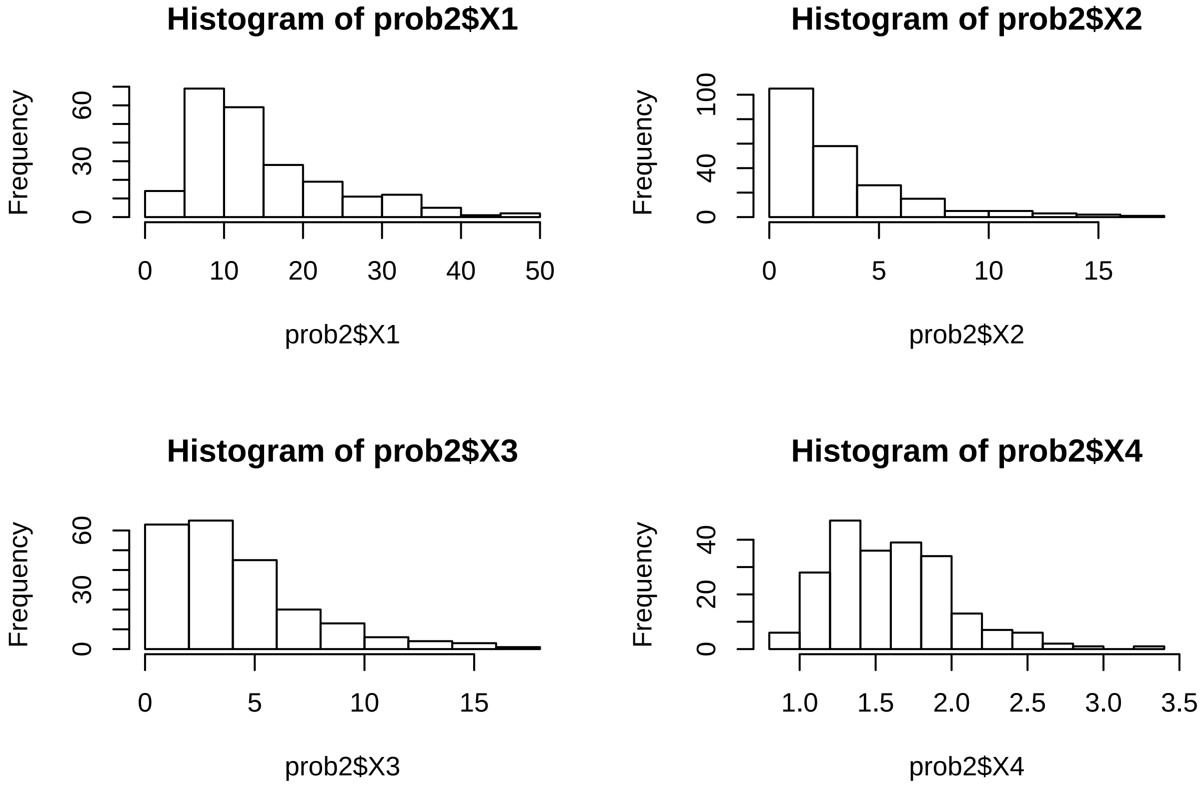
-Prediction and Assessment

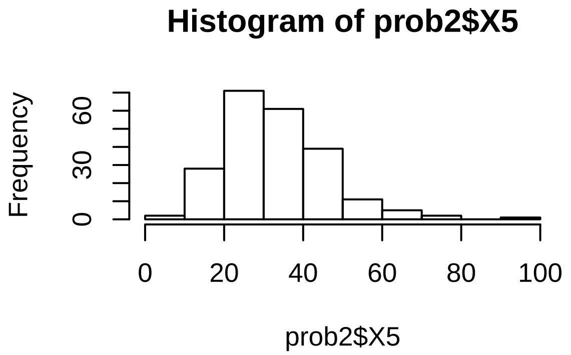
According to our model, the Kangaroo will have about 4.23hrs for slow-wave sleep and 0.91 hour for paradoxical sleep per day.

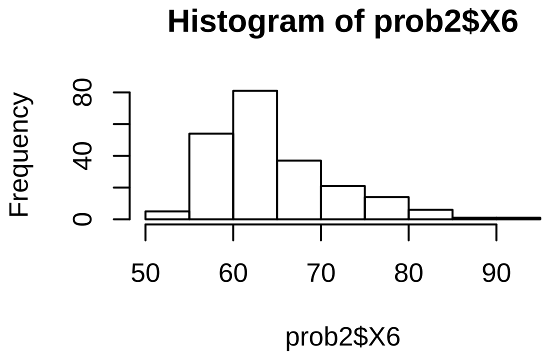
Our model is based on about 40 data points. The data size is not very large. We may need more data in order to let the model to be more reliable. Besizes, more data on the same species will also be helpful to build a reliable data.

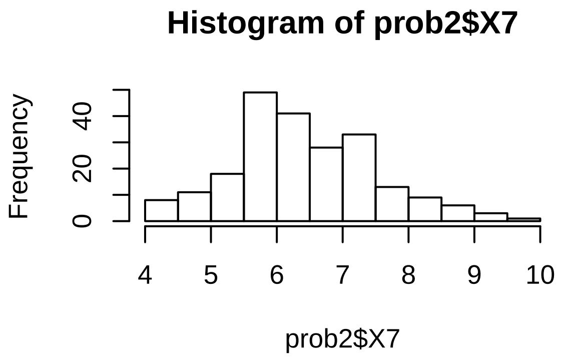
Problem.2

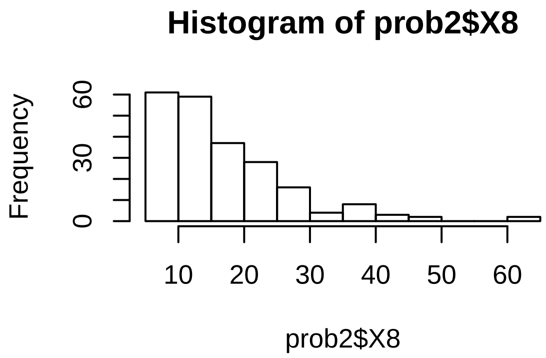
Univariate Analysis

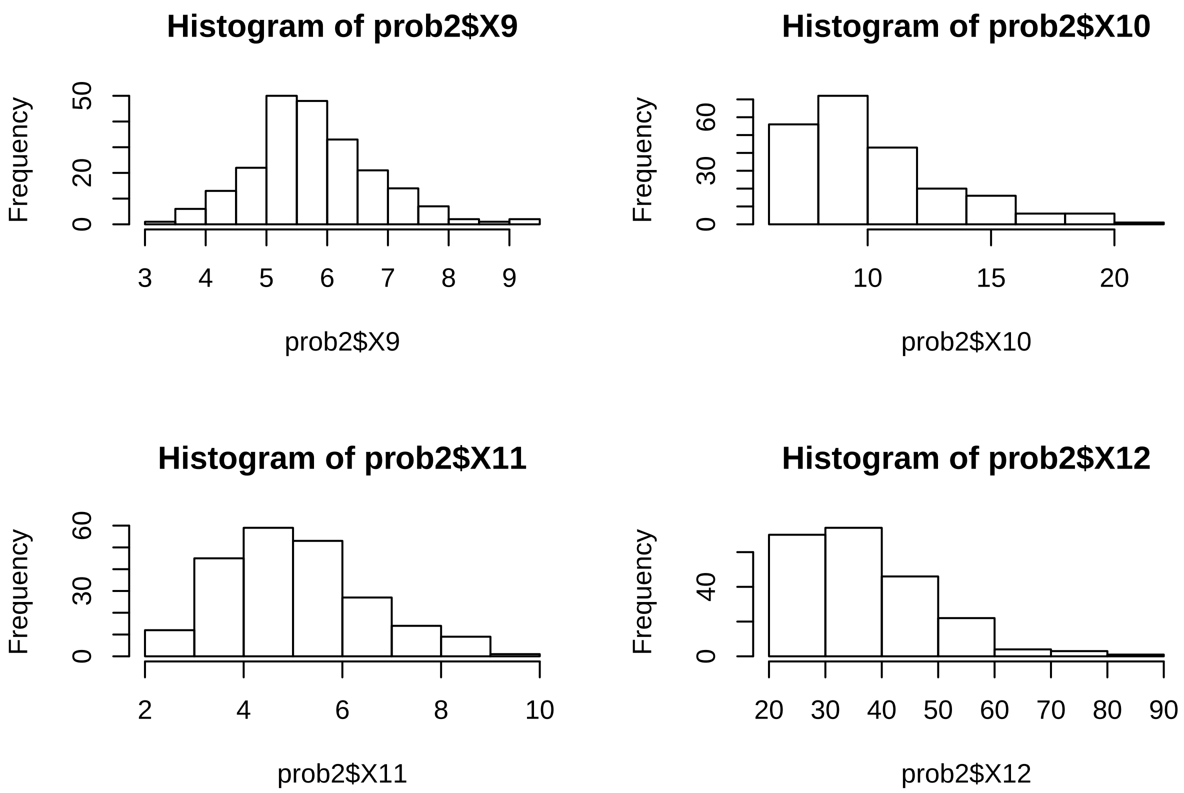


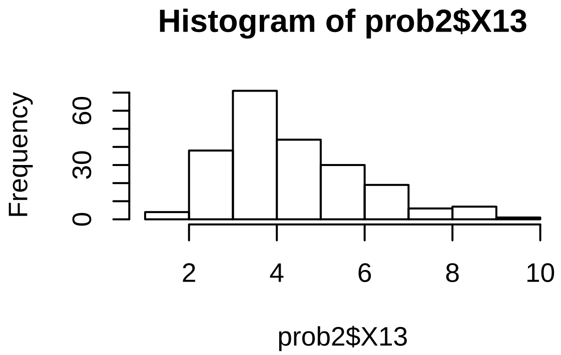

Histogram of prob2$X5
Frequency
0
30
60
0
20
40
60
80
100
prob2$X5

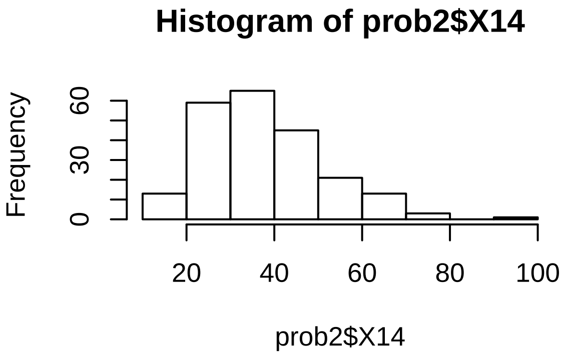

Histogram of prob2$X6
Frequency
0
40
80
50
60
70
80
90
prob2$X6

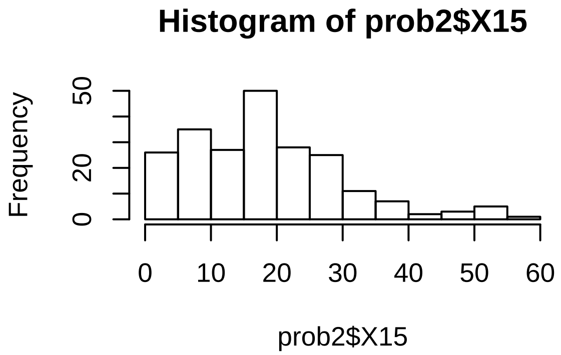

Histogram of prob2$X7
Frequency
0
20
40
4
5
6
7
8
9
10
prob2$X7

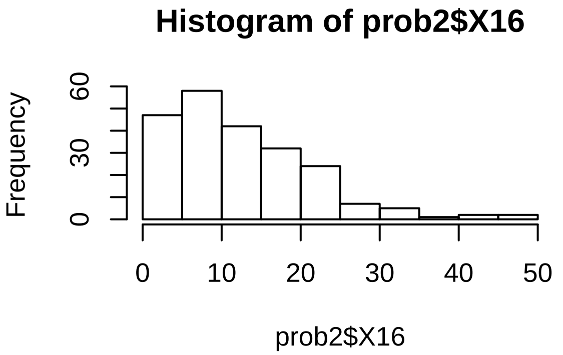

Histogram of prob2$X8
Frequency
0
30
60
10
20
30
40
50
60
prob2$X8

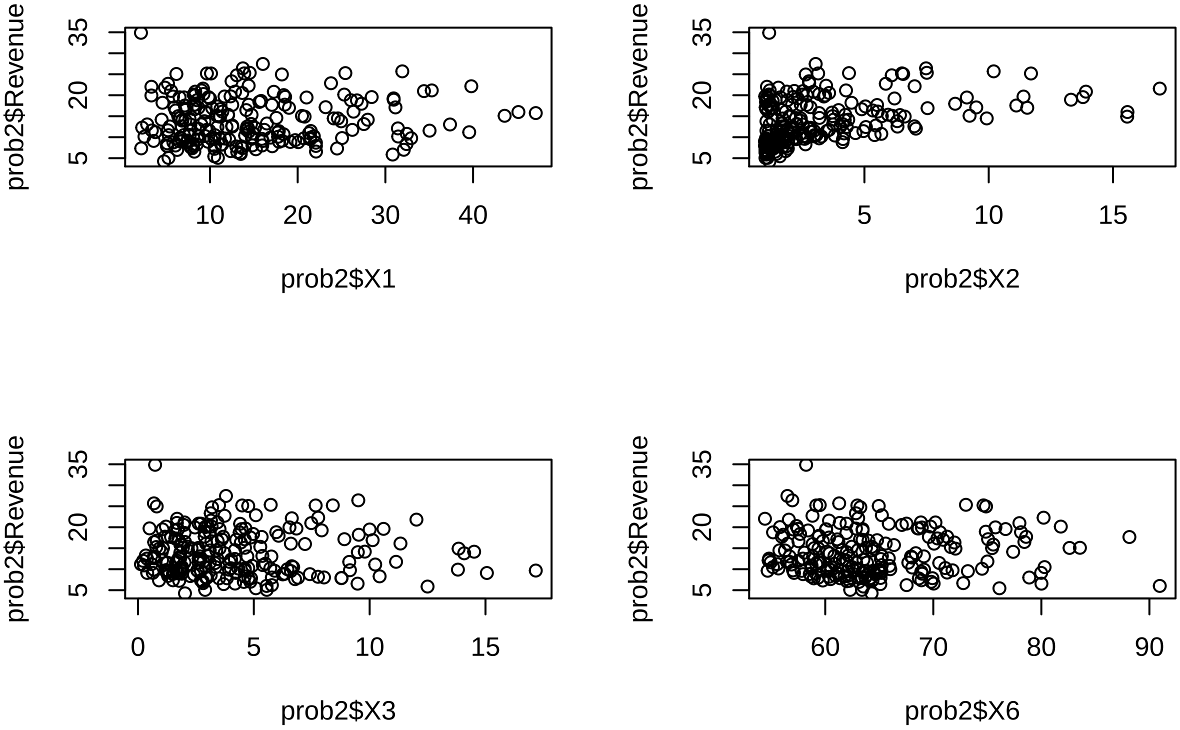

Histogram of prob2$X9
Frequency
0
20
50
3
4
5
6
7
8
9
prob2$X9
Histogram of prob2$X10
Frequency
0
30
60
10
15
20
prob2$X10
Histogram of prob2$X11
Frequency
0
30
60
2
4
6
8
10
prob2$X11
Histogram of prob2$X12
Frequency
0
40
20
30
40
50
60
70
80
90
prob2$X12

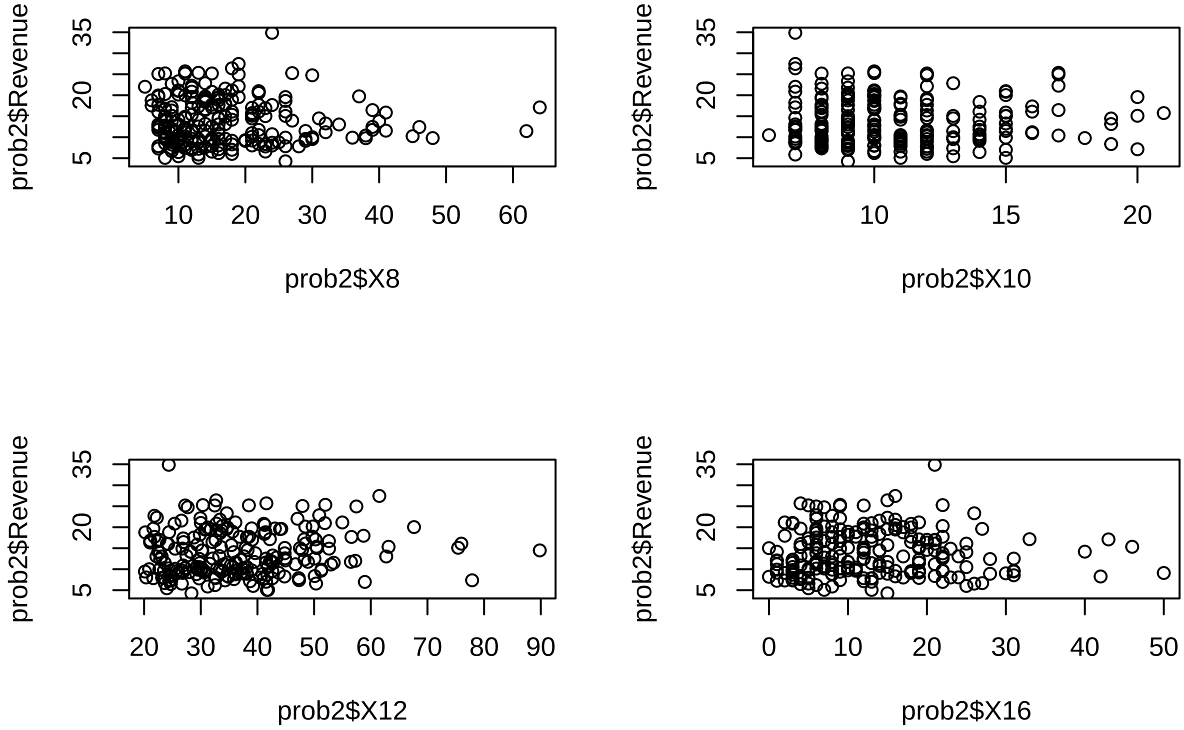

Histogram of prob2$X13
Frequency
0
30
60
2
4
6
8
10
prob2$X13


Histogram of prob2$X14
Frequency
0
30
60
20
40
60
80
100
prob2$X14


Histogram of prob2$X15
Frequency
0
20
50
0
10
20
30
40
50
60
prob2$X15


Histogram of prob2$X16
Frequency
0
30
60
0
10
20
30
40
50
prob2$X16

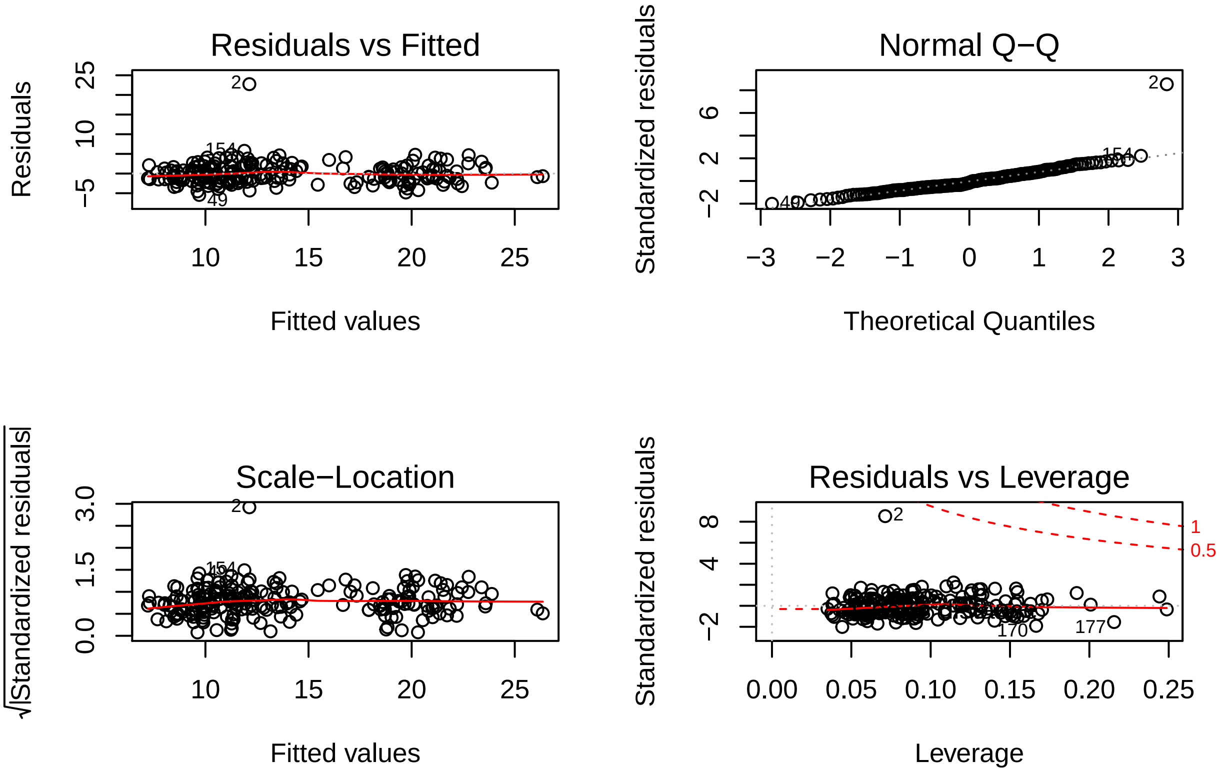

prob2$Revenue
35
20
5
10
20
30
40
prob2$X1
prob2$Revenue
35
20
5
5
10
15
prob2$X2
prob2$Revenue
35
20
5
0
5
10
15
prob2$X3
prob2$Revenue
35
20
5
60
70
80
90
prob2$X6



If we look at the histgram of each numerical variable, we can find that X1,X2, X3, X6, X8,X10, X12 and X16 are very skewed. Thus we decide to look at the relationship of each of these variables with the response. Finally we find the X2 and X8 seem not very linearly related with Revenue and we may need to transform X2 and X8 by log. However, let's look at the full model first.

-Full model

```
##
## Call:
## lm(formula = Revenue ~ . - X17 - X18 - X19 + x17f + x18f + x19f,
##     data = prob2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4391 -1.5315 -0.3816  1.3821 22.7393
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.928367   2.497385   4.776 3.45e-06 ***
## X1          -0.043494   0.027473  -1.583 0.114958
## X2           0.832263   0.066186  12.575  < 2e-16 ***
## X3          -0.062812   0.064287  -0.977 0.329722
## X4          -1.016428   0.633624  -1.604 0.110259
## X5           0.007399   0.019812   0.373 0.709213
## X6          -0.058212   0.035640  -1.633 0.103970
## X7           0.014898   0.433655   0.034 0.972629
## X8           0.038676   0.026318   1.470 0.143253
## X9          -0.502253   0.462697  -1.085 0.279011
## X10          0.113370   0.085706   1.323 0.187419
## X11          0.267399   0.180116   1.485 0.139226
## X12          0.008687   0.016929   0.513 0.608428
## X13          0.112787   0.147914   0.763 0.446649
## X14          0.056490   0.014275   3.957 0.000105 ***
## X15         -0.016259   0.017067  -0.953 0.341911
## X16          0.020584   0.022697   0.907 0.365552
## x17f1       -0.479563   0.877890  -0.546 0.585491
## x18f1       -0.170245   0.418941  -0.406 0.684905
## x19f1        9.359573   0.439703  21.286  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.767 on 200 degrees of freedom
## Multiple R-squared:  0.7664, Adjusted R-squared:  0.7442
## F-statistic: 34.53 on 19 and 200 DF,  p-value: < 2.2e-16
```

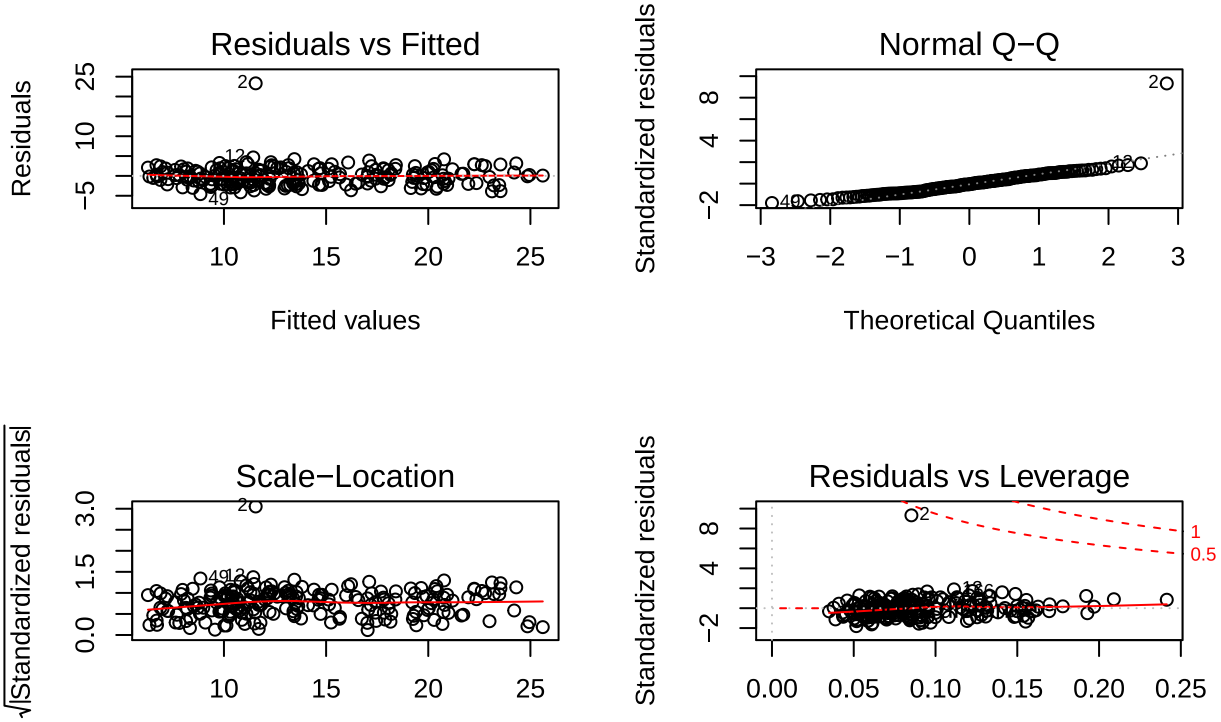


I am thinking about adding interaction terms into the full model. Since it seems (for example,) that whether the movie is released on holiday would influence the level of effect of advertisement spending on revenue. However, all the interaction terms are not statistically significant. Thus I decide not to consider about interaction terms in this problem.

The full model has residual standard error(SE) as 2.77 and the adjusted R^2 is 0.74. The result seem good. However, we need to delete those unsignificant predictors effectively. If we look at the diagnostics plot, we can find that the residual plots has two clusters. It may due to the effect of categorical variables. The Q-Q norm plot has a straight line which means the normality assumption is not violated. There is no point lie beyond the cook's distance curve on Residuals & Leverage plot. So no leverage point that need to be investigated here. However, row 2 seems to be an outlier candidate.

Transformation

We will log-transformed X2 and X8 since their relationship with Revenue seem not very linear



After tranformation, the residual standard error(SE) is 2.62 and the adjusted R^2 is 0.77. These results is a bit better than what we have in the full model. We can find that after log transformation, the clusters in the residual plot disppear. Everything seems better, the transformation is valid.

Model Selection

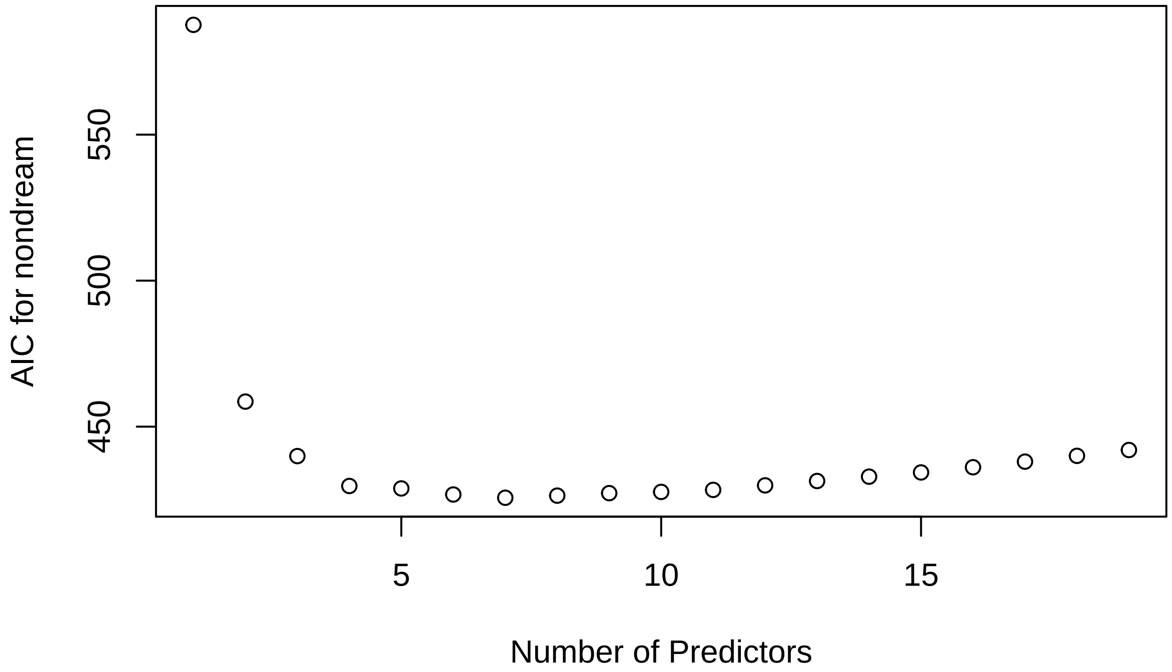
There are 19 potential variables here which is a bit too much to fit the data. Thus, I think about some methods that may help us shrink the dimension effectively. Here are the methods I tried:
1.Partial least square (pls): Looking at the predictors, it seems that some of these variables may be latent variables of others. Also the orthogonal design will avoid colinearity issue. However, the algorithm chooses 16 components as the best with cross validation and none of the predictors has zero coefficient. Pls does not sparse the dimension in this case. Besides, 16 components is still too many. Pls is not good to use here.
2.Lasso: Lasso has the sparsing effect. And it helps me to shrink the model to 9 predictors (X1, log(X2), X3, X4, X6, log(X8), X9, X14 and X19), which is good. However, lasso treat X17 X18 and X19 as numeric variables which is not quite right.

3.Criterion Selection: Finally I chose to use AIC, adjust R^2 and Mallow's C_p methods.

```
aex1=regsubsets(Revenue~X1+logx2+X3+X4+X5+X6+X7+logx8+X9+X10+X11+X12+X13+X14+X15+X16+x17f+x18f+x19f,nvma
rs1=summary(aex1)
rs1$which
```

```
##    (Intercept)    X1 logx2    X3    X4    X5    X6    X7 logx8    X9   X10
## 1         TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2         TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3         TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4         TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 5         TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 6         TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE
## 7         TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE
## 8         TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE
## 9         TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE
## 10        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE
## 11        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE FALSE
## 12        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
## 13        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
## 14        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## 15        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## 16        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## 17        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## 18        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE
## 19        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##      X11   X12   X13   X14   X15   X16 x17f1 x18f1 x19f1
## 1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## 2  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## 3  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 4  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 5  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 6  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 7  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 8  FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 9   TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 10  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 11  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 12  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE
## 13  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE
## 14  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE
## 15  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
## 16  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 17  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
## 18  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 19  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
AIC1 = 220*log(rs1$rss/220) + (2:20)*2
par(mfrow=c(1,1))
plot(AIC1 ~ I(1:19), ylab="AIC for nondream", xlab="Number of Predictors")
```



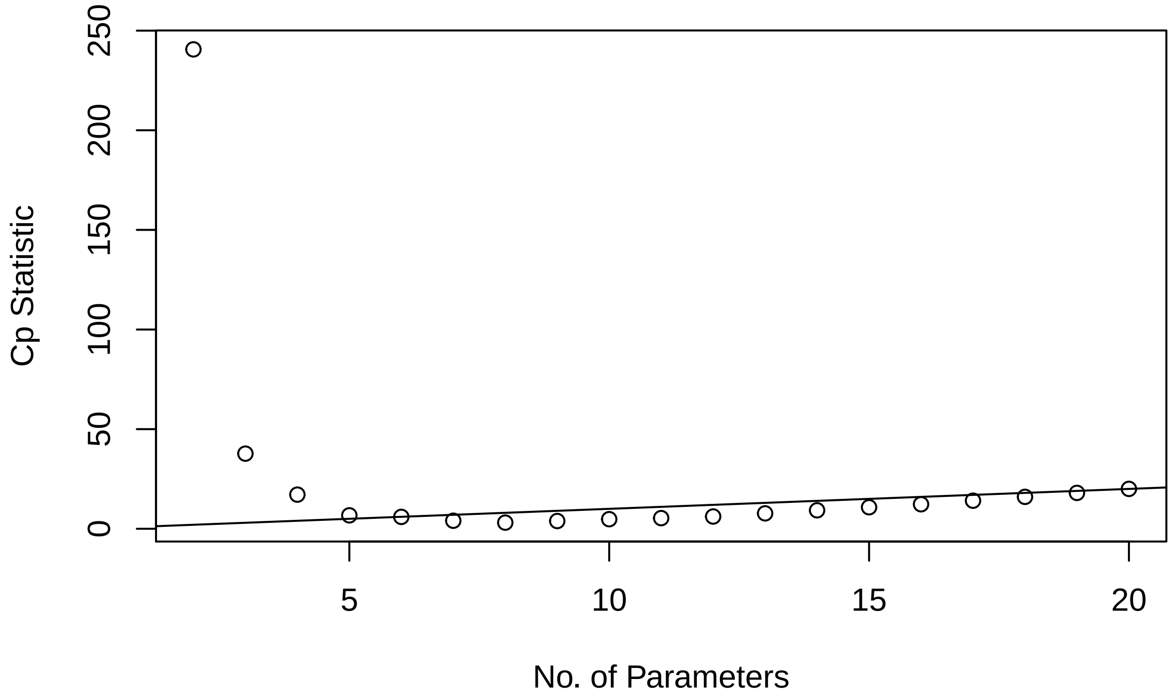
```
which.min(AIC1)
```

```
## [1] 7
```

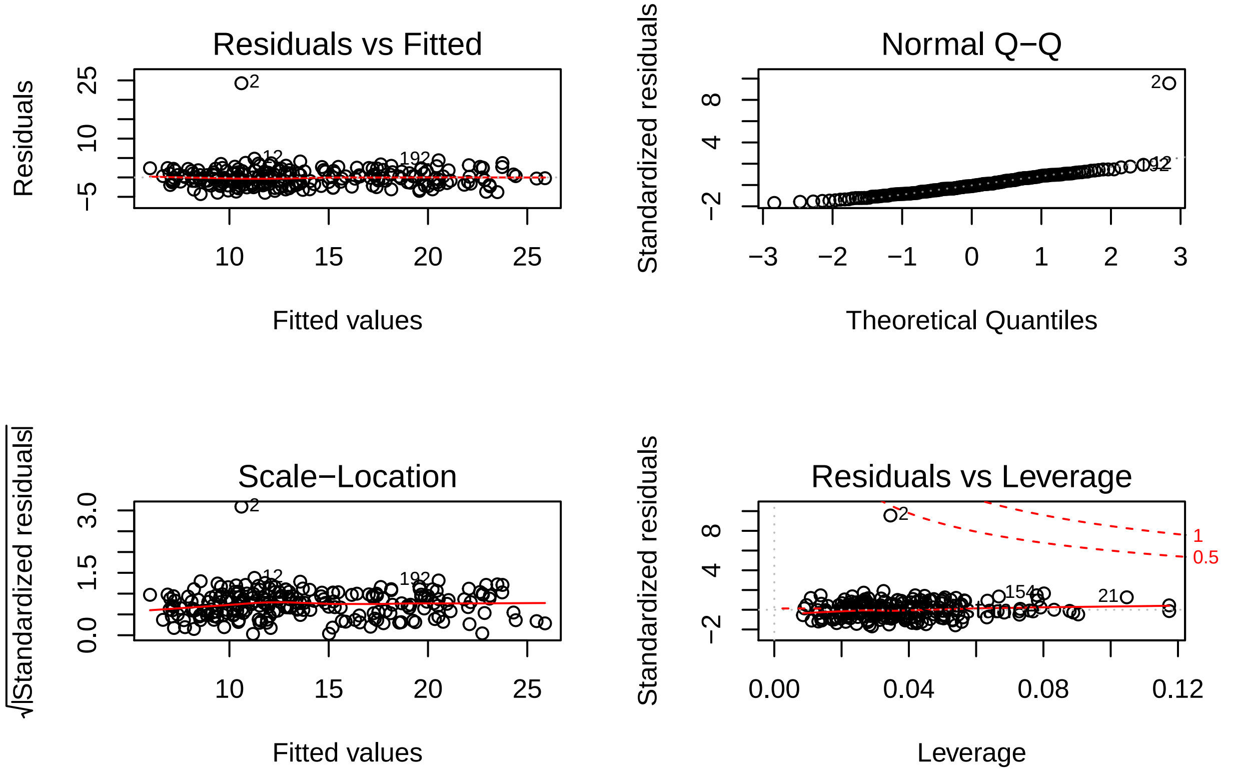
```
which.max(rs1$adjr2) #model selection by adjr2, 11
```

```
## [1] 11
```

```
plot(2:20,rs1$cp,xlab="No. of Parameters",ylab="Cp Statistic") #Mallow's Cp, 10 parameters
abline(0,1)
```



AIC chooses the 7-predictor model, adjust R^2 chosses the 11-predictor model and the C_p values of models with 8 parameters to 19 parameters are very close. The 11-predictor model has non-significant predictors (X3,X6,X11,X13). Also our client wants a simple model. Thus I choose the 7-predictor model to test the outlier. Let's look at the model first. The model is $Revenue = 9.22 - 0.04X1 + 3.72log(X2) - 1.44X4 - 0.05X6 + 0.83log(X8) + 0.07X14 + 9.15X19$. The residual standard error(SE) is 2.58 and the adjusted R^2 is 0.77. All predictors are significant on 10% level. The diagnostic plots seem no problem.



-Check Outliers

```
stud=rstudent(fitts1)
stud[which.max(abs(stud))]
```

```
##        2
## 12.63318
```

```
qt(.05/(220*2),211) # 2 is influential outlier
```

```
## [1] -3.751329
```

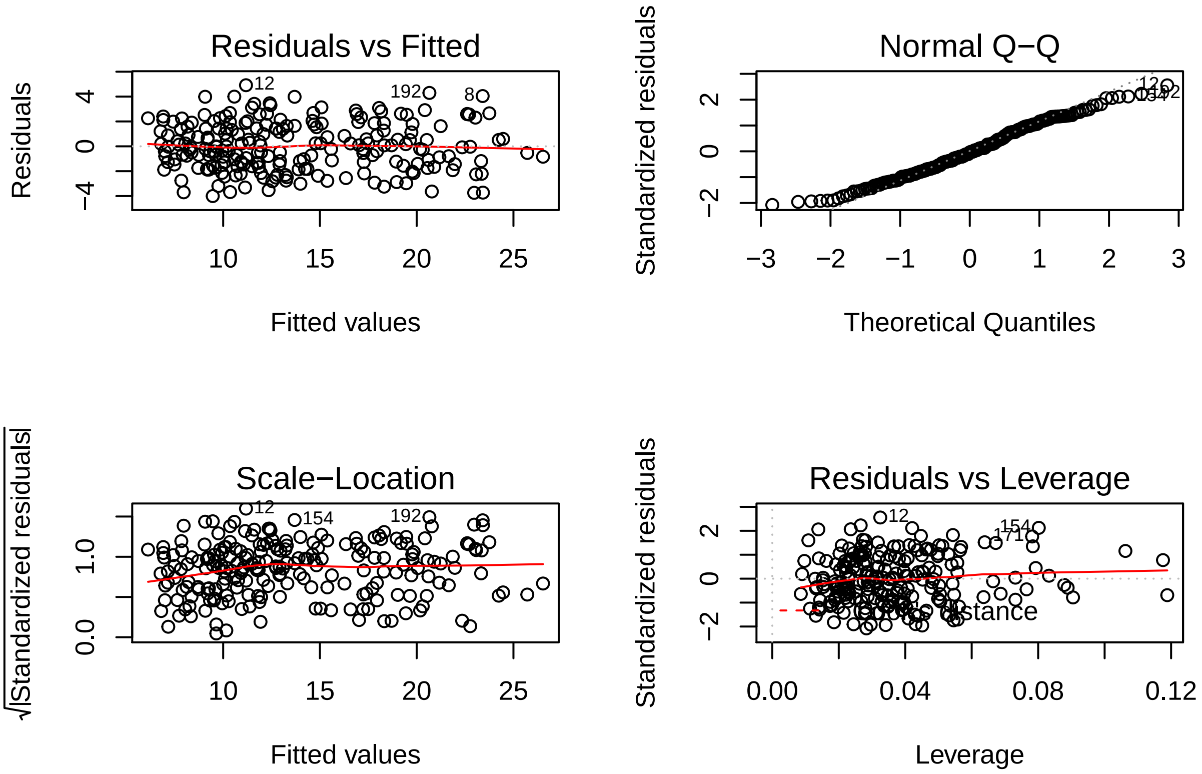
Again, our focus in on influential outliers. And we can find row 2 has studentized residual that is much bigger than the Bonferroni critical value. Thus row 2 is an influential outlier.

```
##     X1   X2   X3   X4    X5    X6   X7 X8   X9 X10  X11   X12  X13   X14
## 2 2.13 1.17 0.74 1.19 32.89 58.23 5.28 24 4.04   7 4.63 24.35 3.44 38.56
##   X15 X16 X17 X18 X19 Revenue
## 2   1  21   1   0   0   34.87
```

If we look at row 2, we can find that the spending of the movie is less and the scores on director, producer, actor and actress are not outstanding. However, the revenue is quite a lot. This may be a special success of low-cost movie. This case is not general. Also, since our data size is large, one outlier would not affect our model too much. We may delete our remain it in our model. I decided to delete it.

-Final Model

```
##
## Call:
## lm(formula = Revenue ~ X1 + x2new + X4 + X6 + x8new + X14 + x19new,
##     data = probnew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0008 -1.4579 -0.0805  1.4879  4.9111
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.457311   1.434329   5.896 1.46e-08 ***
## X1          -0.026065   0.016209  -1.608 0.109316
## x2new        3.855893   0.188339  20.473  < 2e-16 ***
## X4          -1.191367   0.346252  -3.441 0.000699 ***
## X6          -0.029745   0.019723  -1.508 0.133028
## x8new        0.430729   0.286948   1.501 0.134832
## X14          0.065109   0.009794   6.648 2.50e-10 ***
## x19new1      9.232368   0.306408  30.131  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.955 on 211 degrees of freedom
## Multiple R-squared:  0.8681, Adjusted R-squared:  0.8637
## F-statistic: 198.3 on 7 and 211 DF,  p-value: < 2.2e-16
```



Our final model is $Revenue = 8.45 - 0.03X1 + 3.85log(X2) - 1.19X4 - 0.03X6 + 0.43log(X8) + 0.06X14 + 9.23X19$. The residual standard error(SE) is 1.96 and the adjusted R^2 is 0.86. Our model says that advertisement spending, director experience, popularity of main actress has a positive relationsip with revenue. The production costs, length of movie and score of soundtrack are negatively related to the revenue. It is a bit strange since it says the movie that costs less would earn more.