STATS 413 Hw4

Shu Zhou

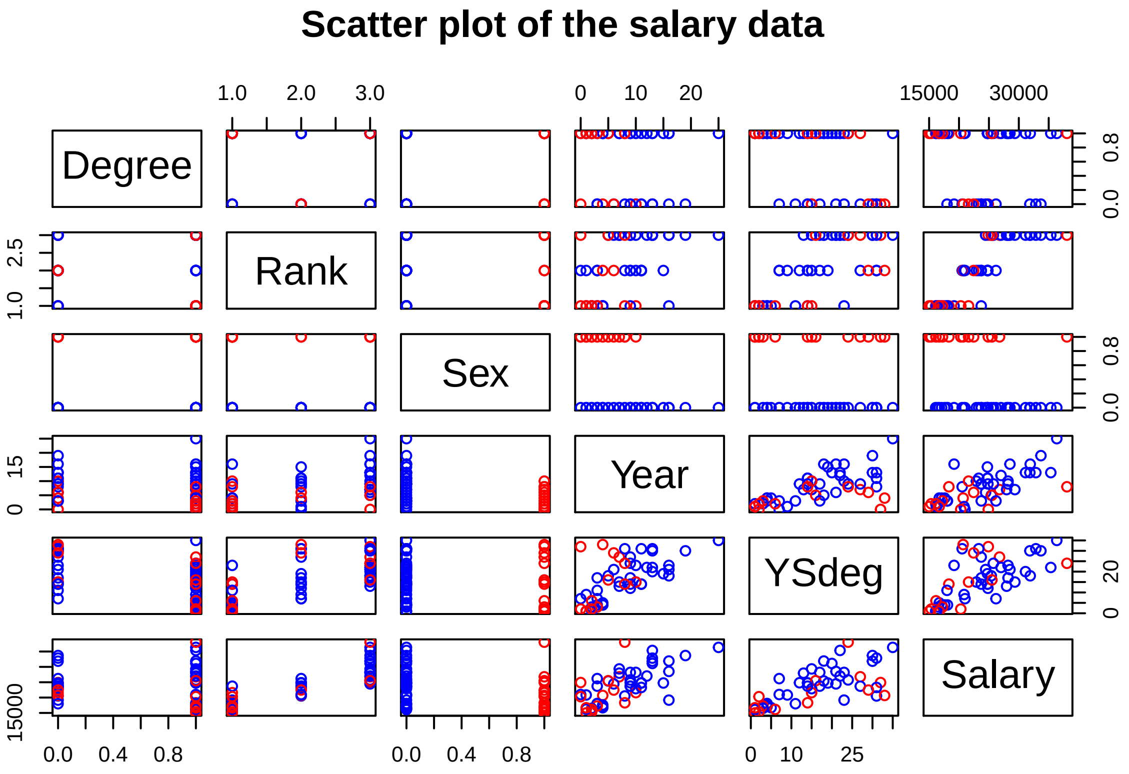
2020/11/4

This is the Assignment 4 of STATS 413 Author: Shu Zhou UMID: 19342932

Exercise 5.17

(5.17.1.)

```
data("salary")
cols <- character(nrow(salary))
cols[] <- "black"
cols[salary$Sex == 1] <- "red"
cols[salary$Sex == 0] <- "blue"
pairs(salary[,c(1:6)],col = cols, main = "Scatter plot of the salary data")
```



In this plot we use blue points to represent Males and red points to represent females. From this plot, we can see that

- Females generally have fewer years in rank.
- The variance of salary is much higher with a person who own a master's degree.
- Females generally have a lower salary than males.
- The mean function of YSdeg might have a different slope for males and females.

(5.17.2)

```
salary<-read.csv("salary.csv")
salary<-as.data.frame(salary)
summary(lm(salary~sex,data= salary))
```

```
##
## Call:
## lm(formula = salary ~ sex, data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8602.8 -4296.6  -100.8  3513.1 16687.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21357       1545  13.820   <2e-16 ***
## sexMale         3340       1808   1.847   0.0706 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5782 on 50 degrees of freedom
## Multiple R-squared:  0.0639, Adjusted R-squared:  0.04518
## F-statistic: 3.413 on 1 and 50 DF,  p-value: 0.0706
```

The significance level is 0.0706. Hence the sex factor is not statistically significant, we cannot reject the null hypothesis with 95% of confidence. The point estimate of the Sex effect is $3340 in favor of men. **(5.17.3)**_

```
model1<-lm(salary~.,data= salary)
summary(model1)
```

```
##
## Call:
## lm(formula = salary ~ ., data = salary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4168.1  -886.8  -275.6   694.0  9014.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21060.11    3044.29   6.918 1.51e-08 ***
## X              27.34      63.76   0.429    0.670
## degreePhD    1438.04    1034.55   1.390    0.172
## rankAsst    -5498.57    1251.93  -4.392 6.96e-05 ***
## rankProf     6127.86    1240.57   4.940 1.18e-05 ***
## sexMale     -1089.64     951.06  -1.146    0.258
## year          503.23     114.52   4.394 6.92e-05 ***
## ysdeg        -118.31      79.55  -1.487    0.144
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2420 on 44 degrees of freedom
## Multiple R-squared:  0.8556, Adjusted R-squared:  0.8327
## F-statistic: 37.26 on 7 and 44 DF,  p-value: < 2.2e-16
```

```
confint(model1)["sexMale", , drop=FALSE]
```

```
##             2.5 %  97.5 %
## sexMale -3006.367 827.092
```

We can see that the sex effect is much higher for females with higher salaries. Although we cannot reject our null hypothesis according to P-value, we cannot say that sex has no impact.

(5.17.4)_

```
model2<-lm(salary~.-rank,data= salary)
summary(model2)
```

```
##
## Call:
## lm(formula = salary ~ . - rank, data = salary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6725  -1950    -30   1871  11960
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28443.31    3487.89   8.155 1.75e-10 ***
## X            -245.29      64.64  -3.795  0.00043 ***
## degreePhD   -1728.43    1221.27  -1.415  0.16372
## sexMale      -331.09    1234.24  -0.268  0.78971
## year          154.35     136.05   1.135  0.26245
## ysdeg          95.56      95.84   0.997  0.32395
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3302 on 46 degrees of freedom
## Multiple R-squared:  0.7191, Adjusted R-squared:  0.6886
## F-statistic: 23.55 on 5 and 46 DF,  p-value: 1.164e-11
```

After the factor degree is excluded, we can see that the most variables become less significant. Hence we can argue that this sample is not unbiased and it is not a good choice to rely on this sample.

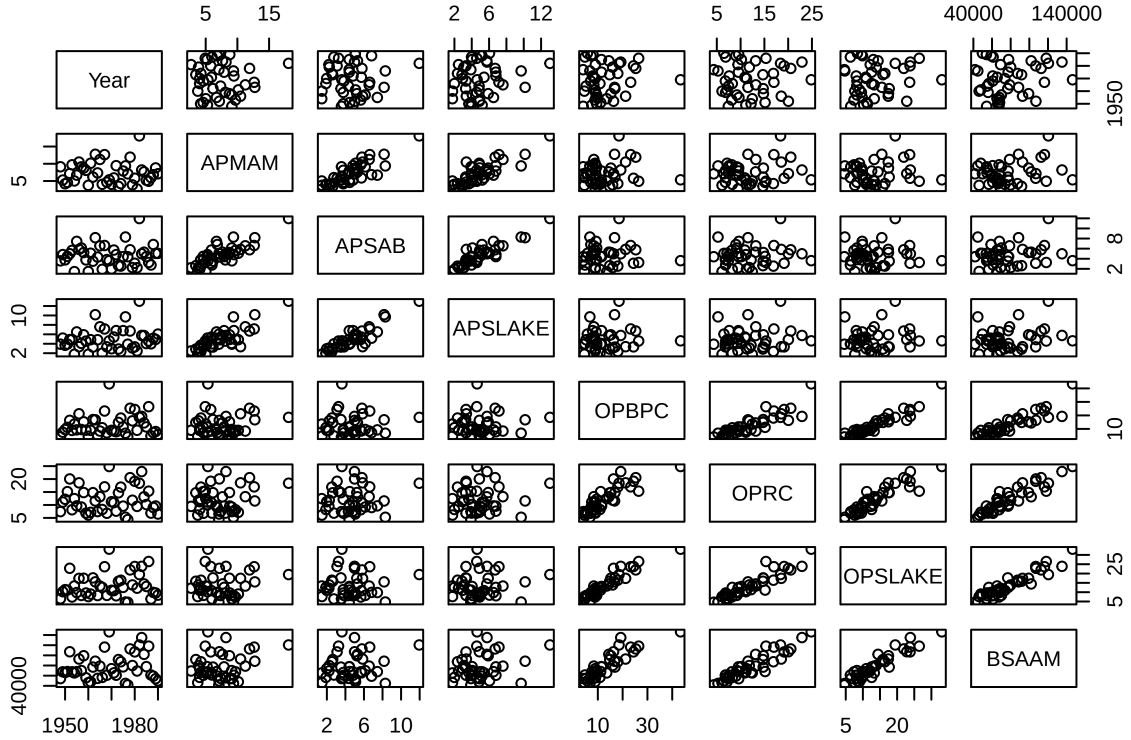
Exercise 8.3

(8.3.1)

```
water<-read.csv("water.csv")
cor(water)
```

```
##                     X          Year          APMAM      APSAB    APSLAKE
## X        1.0000000000  1.0000000000 -0.0007590557 0.05182523 0.17014669
## Year     1.0000000000  1.0000000000 -0.0007590557 0.05182523 0.17014669
## APMAM   -0.0007590557 -0.0007590557  1.0000000000 0.82768637 0.81607595
## APSAB    0.0518252272  0.0518252272  0.8276863704 1.00000000 0.90030474
## APSLAKE  0.1701466883  0.1701466883  0.8160759519 0.90030474 1.00000000
## OPBPC    0.1185994341  0.1185994341  0.1223856707 0.03954211 0.09344773
## OPRC     0.0224682441  0.0224682441  0.1544154918 0.10563959 0.10638359
## OPSLAKE  0.1380333978  0.1380333978  0.1075421167 0.02961175 0.10058669
## BSAAM    0.1699631973  0.1699631973  0.2385695382 0.18329499 0.24934094
##              OPBPC       OPRC    OPSLAKE     BSAAM
## X       0.11859943 0.02246824 0.13803340 0.1699632
## Year    0.11859943 0.02246824 0.13803340 0.1699632
## APMAM   0.12238567 0.15441549 0.10754212 0.2385695
## APSAB   0.03954211 0.10563959 0.02961175 0.1832950
## APSLAKE 0.09344773 0.10638359 0.10058669 0.2493409
## OPBPC   1.00000000 0.86470733 0.94334741 0.8857478
## OPRC    0.86470733 1.00000000 0.91914467 0.9196270
## OPSLAKE 0.94334741 0.91914467 1.00000000 0.9384360
## BSAAM   0.88574778 0.91962700 0.93843604 1.0000000
```

```
pairs(water[ , 2:9])
```



- The correlations between "OPBPC", "OPRC","OPSLAKE" and "BSAAM" are very high.
- The correlations between year and other variables are low.
- The correlations between "APMAM", "APSAB" and "APSLAKE" are high but not as high as the "O" variables.

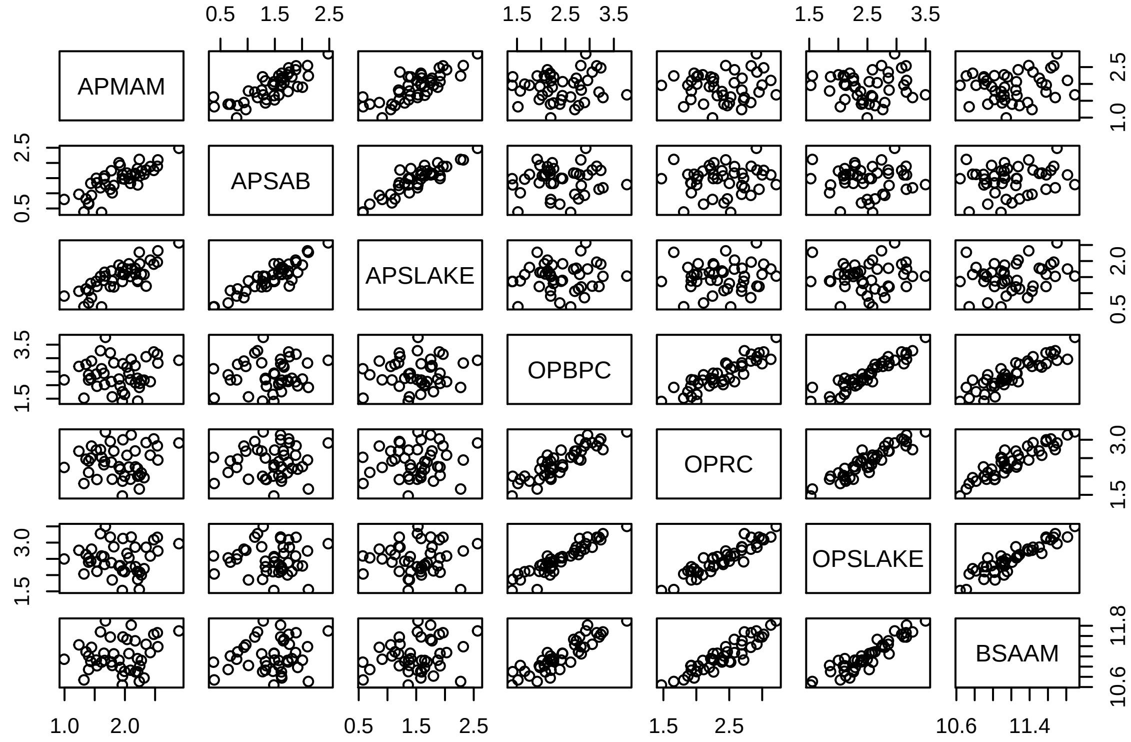
(8.3.2)

```
summary(ans <- powerTransform( as.matrix(water[ , 3:8]) ~ 1))
```

```
## bcPower Transformations to Multinormality
##         Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## APMAM      0.0982           0      -0.4625       0.6589
## APSAB      0.3450           0      -0.0533       0.7432
## APSLAKE    0.0818           0      -0.3466       0.5101
## OPBPC      0.0982           0      -0.2109       0.4073
## OPRC       0.2536           0      -0.2255       0.7328
## OPSLAKE    0.2534           0      -0.0921       0.5988
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                    LRT df     pval
## LR test, lambda = (0 0 0 0 0 0) 5.452999  6 0.48716
##
## Likelihood ratio test that no transformations are needed
##                                    LRT df        pval
## LR test, lambda = (1 1 1 1 1 1) 61.20312  6 2.5629e-11
```

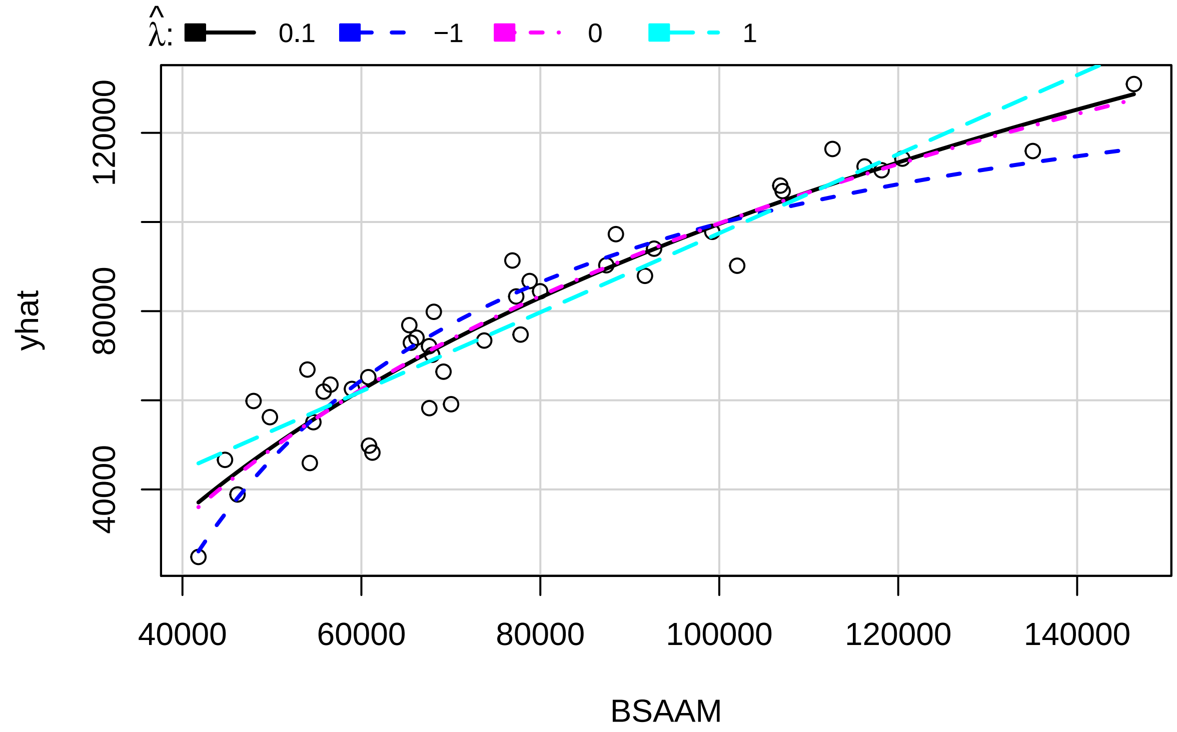
The transformation we found appear to achieve linearity. Since the p-value for the LR-test is 0.48716.

```r
pairs(log(water[ , 3:9]))
```



(8.3.3)

```r
model3 <- lm(BSAAM ~ log(APMAM) + log(APSAB) + log(APSLAKE) + log(OPBPC) + log(OPRC) + log(OPSLAKE), wa
invResPlot(model3)
```



```
##       lambda        RSS
## 1  0.1048461 2257433456
## 2 -1.0000000 3008670148
## 3  0.0000000 2264377190
## 4  1.0000000 2745251921
```

The fitted line for $\hat{\lambda}$ have the smallest RSS, hence it is the best fit, which indicates that the log transform is reasonable.

(8.3.4)

```r
model4 <- lm(formula = log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) +
log(OPBPC) + log(OPRC) + log(OPSLAKE), data = water)
summary(model4)
```

```
##
## Call:
## lm(formula = log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) +
##     log(OPBPC) + log(OPRC) + log(OPSLAKE), data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18671 -0.05264 -0.00693  0.06130  0.17698
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   9.46675    0.12354  76.626  < 2e-16 ***
## log(APMAM)   -0.02033    0.06596  -0.308  0.75975    
## log(APSAB)   -0.10303    0.08939  -1.153  0.25667    
## log(APSLAKE)  0.22060    0.08955   2.463  0.01868 *  
## log(OPBPC)    0.11135    0.08169   1.363  0.18134    
## log(OPRC)     0.36165    0.10926   3.310  0.00213 ** 
## log(OPSLAKE)  0.18613    0.13141   1.416  0.16524    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1017 on 36 degrees of freedom
## Multiple R-squared:  0.9098, Adjusted R-squared:  0.8948 
## F-statistic: 60.54 on 6 and 36 DF,  p-value: < 2.2e-16
```

The two negative estimates are log(APMAM) and log(APSAB). Both of them are not significant. The negative signs are caused by the correlations of other included regressors.

(8.3.5)

```
water$geometricmean_O<-rowSums(water[,6:8])/3
model5 <- lm(log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) +geometricmean_O , water)
anova(model5,model4)
```

```
## Analysis of Variance Table
## 
## Model 1: log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) + geometricmean_O
## Model 2: log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) + log(OPBPC) + 
##     log(OPRC) + log(OPSLAKE)
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)    
## 1     38 0.56725                                 
## 2     36 0.37243  2   0.19481 9.4155 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence, we can reject that the three "O" log predictors are not equal ,i.e. they are equal. Which shows that the geometric mean of the snow depth represents its valley as well as do the individual measurements.

```
water$geometricmean_A<-rowSums(water[,3:5])/3
model6 <- lm(log(BSAAM) ~ log(OPBPC) + log(OPRC) + log(OPSLAKE) +geometricmean_A , water)
anova(model6,model4)
```

```
## Analysis of Variance Table
## 
## Model 1: log(BSAAM) ~ log(OPBPC) + log(OPRC) + log(OPSLAKE) + geometricmean_A
## Model 2: log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) + log(OPBPC) + 
##     log(OPRC) + log(OPSLAKE)
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)  
## 1     38 0.42614                             
## 2     36 0.37243  2  0.053705 2.5956 0.0885 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hence, we cannot reject that the three "A" log predictors are not equal. Which shows that each "A" log predictor measurement is important.