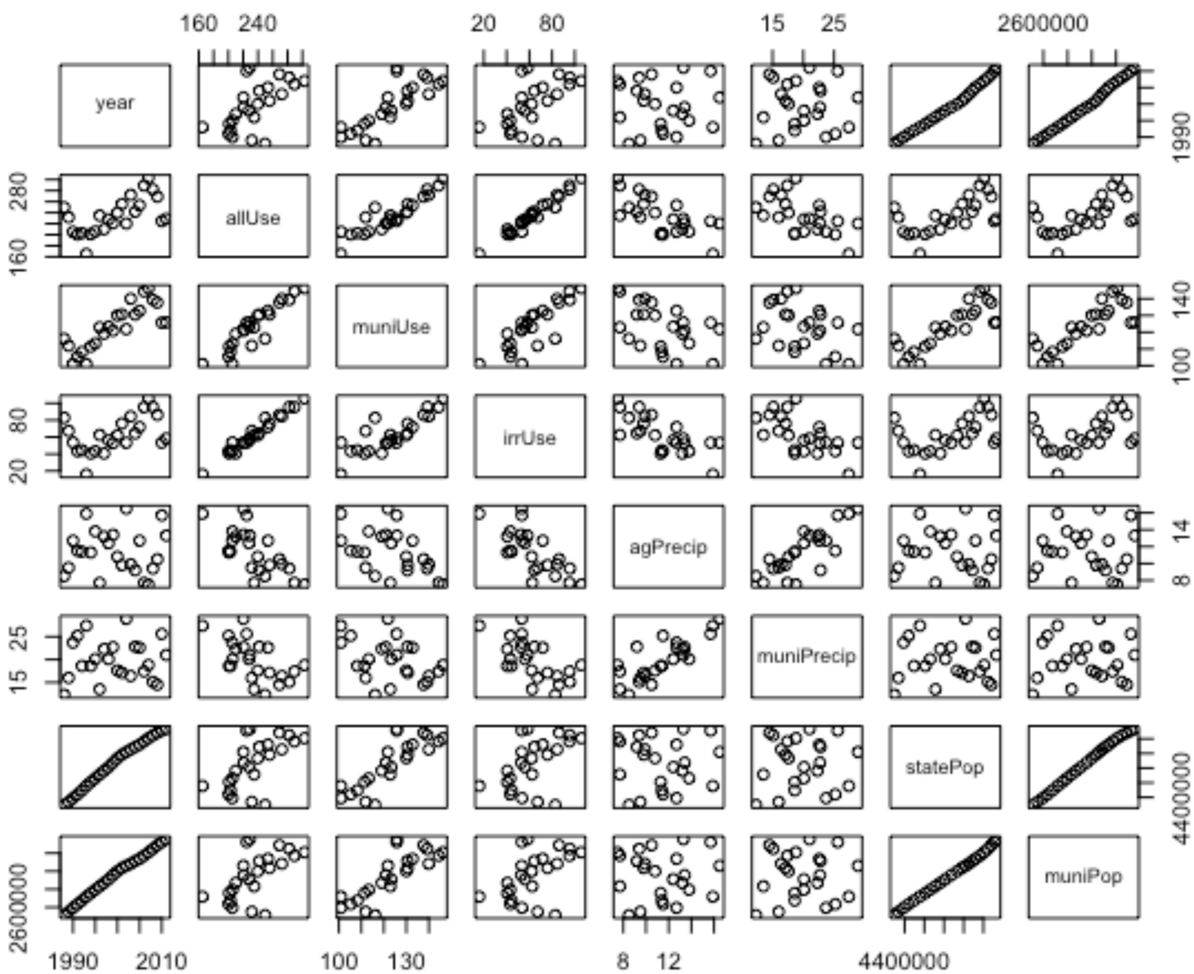
HW3

Name: Zixin Ouyang

1.

(a)

```
> library(alr4)
> pairs(MinnWater)
```



(b) year, statePop and muniPop have especially high sample correlations with each other.

(c)

```
> minmod1<-lm(muniUse~., data=MinnWater)
> summary(minmod1)

Call:
lm(formula = muniUse ~ ., data = MinnWater)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.38834 -0.44331  0.02071  0.47878  1.26227 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.163e+02  1.170e+03   0.527  0.60567    
year        -3.763e-01  6.034e-01  -0.624  0.54167    
allUse       6.909e-01  7.140e-02   9.677 4.33e-08 ***
irrUse      -6.535e-01  8.681e-02  -7.528 1.21e-06 ***
agPrecip     1.623e-01  1.562e-01   1.040  0.31398    
muniPrecip  -2.491e-01  7.980e-02  -3.122  0.00658 ** 
statePop     6.082e-05  2.237e-05   2.719  0.01517 *  
muniPop     -5.228e-05  3.876e-05  -1.349  0.19621    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8131 on 16 degrees of freedom
Multiple R-squared:  0.9973,	Adjusted R-squared:  0.9962 
F-statistic: 855.2 on 7 and 16 DF,  p-value: < 2.2e-16
```

(d)

```
> vif(minmod1)
      year     allUse     irrUse   agPrecip muniPrecip   statePop    muniPop 
 633.34563  190.15277  118.11767    5.72228    4.28763 1904.44626 3441.37710 
```

year, allUse, irrUse, StatePop and muniPop have a VIF indicating a problem of (approximate) collinearity.

(e)

```
> minmod2<-lm(muniUse~allUse+irrUse+muniPrecip+statePop, data=MinnWater)
> summary(minmod2)

Call:
lm(formula = muniUse ~ allUse + irrUse + muniPrecip + statePop,
    data = MinnWater)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3681 -0.6835 -0.2439  0.7012  2.1665 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.256e+01  4.904e+00 -12.758 9.15e-11 ***
allUse       8.219e-01  5.251e-02  15.650 2.60e-12 ***
irrUse      -8.227e-01  7.270e-02 -11.317 6.92e-10 ***
muniPrecip  -1.332e-01  7.084e-02  -1.881   0.0754 .  
statePop     9.731e-06  1.244e-06   7.819 2.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.008 on 19 degrees of freedom
Multiple R-squared:  0.9951,	Adjusted R-squared:  0.9941 
F-statistic: 972.5 on 4 and 19 DF,  p-value: < 2.2e-16
```

The p-value of muniPrecip is 0.0754, larger than 0.05, so it’s not significant now. Other variables are still significant.

(f)

```
> vif(minmod2)
    allUse     irrUse muniPrecip   statePop 
 66.985801  53.944057   2.200120   3.838104 
```

The VIFs for these variables have decreased, but allUse and irrUse still have a VIF indicating a problem of (approximate) collinearity.

2.

(a)

```
> library(faraway)
> fit1<-lm(crawling~temperature, data=crawl)
```

```
> summary(fit1)

Call:
lm(formula = crawling ~ temperature, data = crawl)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0556 -0.5712  0.5221  0.8029  1.4334 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.67806    1.31753  27.080 1.09e-10 ***
temperature -0.07774    0.02510  -3.097   0.0113 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.319 on 10 degrees of freedom
Multiple R-squared:  0.4896,	Adjusted R-squared:  0.4386 
F-statistic: 9.592 on 1 and 10 DF,  p-value: 0.01131
```

(b)

```
> confint(fit1,"temperature")
                 2.5 %      97.5 %
temperature -0.1336661 -0.02181224
```

(c)

```
> fit2<-lm(crawling~temperature, weights=n,data=crawl)
> summary(fit2)

Call:
lm(formula = crawling ~ temperature, data = crawl, weights = n)

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-16.581  -4.325   2.001   4.026   9.146 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.70254    1.26003  28.335 6.97e-11 ***
temperature -0.07561    0.02454  -3.081   0.0116 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.399 on 10 degrees of freedom
Multiple R-squared:  0.487,	Adjusted R-squared:  0.4357 
F-statistic: 9.494 on 1 and 10 DF,  p-value: 0.01162
```

(d)

```
> crawl
          crawling   SD  n temperature
January      29.84 7.08 32          66
February     30.52 6.96 36          73
March        29.70 8.33 23          72
April        31.84 6.21 26          63
May          28.58 8.07 27          52
June         31.44 8.10 29          39
July         33.64 6.91 21          33
August       32.82 7.61 45          30
September    33.83 6.93 38          33
October      33.35 7.29 44          37
November     33.38 7.42 49          48
December     32.32 5.71 44          57
```

The first element of the weight matrix W is 32, and the first element of the matrix Σ is 1/32.

(e)

```
> confint(fit2, "temperature")
                 2.5 %      97.5 %
temperature -0.1302824 -0.02093222
```

(f)

```
> fit3<-lm(crawling~temperature, weights=n/(SD^2),data=crawl)
> summary(fit3)

Call:
lm(formula = crawling ~ temperature, data = crawl, weights = n/(SD^2))

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-2.1504 -0.6817  0.1688  0.4941  1.1009 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35.73262    1.21153   29.49 4.69e-11 ***
temperature -0.07332    0.02328   -3.15   0.0103 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9772 on 10 degrees of freedom
Multiple R-squared:  0.4981,	Adjusted R-squared:  0.4479 
F-statistic: 9.923 on 1 and 10 DF,  p-value: 0.01033
```

(g)

```
> 32/(7.08^2)
[1] 0.6383862
> 1/(32/(7.08^2))
[1] 1.56645
```

The first element of the weight matrix W is 0.6383862, and the first element of the matrix Σ is 1.56645.

(h)

```
> confint(fit3, "temperature")
                 2.5 %      97.5 %
temperature -0.1251886 -0.02145946
```

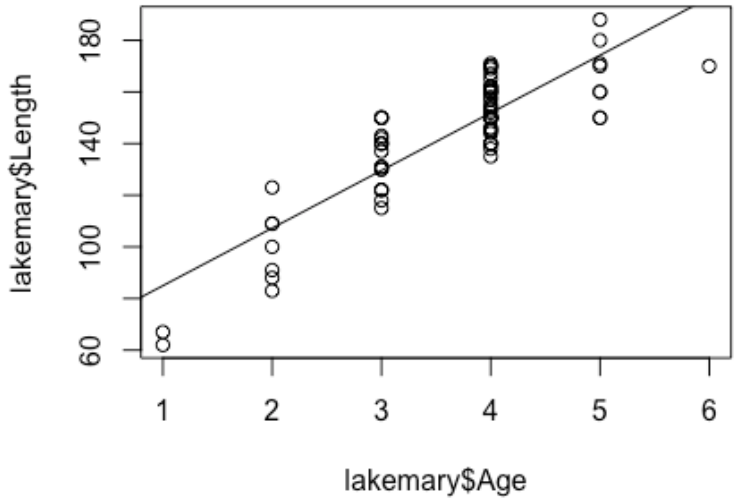
3.

(a)

```
> lakemod1<-lm(Length~Age, data=lakemary)
> summary(lakemod1)

Call:
lm(formula = Length ~ Age, data = lakemary)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.523  -7.586   0.258  10.102  20.414 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   62.649      5.755   10.89   <2e-16 ***
Age           22.312      1.537   14.51   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.51 on 76 degrees of freedom
Multiple R-squared:  0.7349,	Adjusted R-squared:  0.7314 
F-statistic: 210.7 on 1 and 76 DF,  p-value: < 2.2e-16
```

(b)

```
> plot(lakemary$Age, lakemary$Length)
> abline(lakemod1)
```



(c) There are repeated age values in the data set.

(d)

```
> lakemod2<-lm(Length~factor(Age), data=lakemary)
> anova(lakemod1,lakemod2)
Analysis of Variance Table

Model 1: Length ~ Age
Model 2: Length ~ factor(Age)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     76 11892.8
2     72  8812.7  4    3080.2 6.2912 0.0002125 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is less than 0.05, so we reject the null hypothesis and conclude that the simple linear regression model is lack of fit.

(e) According to d, an estimate of the pure error variance is 8812.7/72 = 122.4.

(f)

```
> lakemod3<-lm(Length ~ Age + I(Age^2), data=lakemary)
> summary(lakemod3)

Call:
lm(formula = Length ~ Age + I(Age^2), data = lakemary)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.846  -8.321  -1.137   6.698  22.098 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   13.622     11.016   1.237     0.22    
Age           54.049      6.489   8.330 2.81e-12 ***
I(Age^2)      -4.719      0.944  -4.999 3.67e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.91 on 75 degrees of freedom
Multiple R-squared:  0.8011,	Adjusted R-squared:  0.7958 
F-statistic: 151.1 on 2 and 75 DF,  p-value: < 2.2e-16
```

(g)

```
> anova(lakemod3,lakemod2)
Analysis of Variance Table

Model 1: Length ~ Age + I(Age^2)
Model 2: Length ~ factor(Age)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     75 8920.7                           
2     72 8812.7  3    108.01 0.2942 0.8295
```

The p-value is very large, so we cannot reject the null hypothesis and conclude that the quadratic regression model is adequate.

4.

(a)

$\hat{\beta}=(X^TX)^{-1}X^TY$, $Var(Y)=\sigma^2\Sigma$

$Var(\hat{\beta})=Var((X^TX)^{-1}X^TY)= (X^TX)^{-1}X^TYVar(Y)X(X^TX)^{-1}=\sigma^2(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}$

(b)

$\hat{\beta}_G=(X^T\Sigma^{-1}X)^{-1}X^T\,\Sigma^{-1}Y$, $Var(Y)=\sigma^2\Sigma$

$Var(\hat{\beta}_G)=Var((X^T\Sigma^{-1}X)^{-1}X^T\,\Sigma^{-1}Y)=(X^T\Sigma^{-1}X)^{-1}X^T\,\Sigma^{-1}Var(Y)\,\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}=\sigma^2(X^T\Sigma^{-1}X)^{-1}$

5.

(a)

```
> unweighted_res<-crawl$crawling-fitted(fit2)
> unweighted_res
   January   February      March      April        May       June       July 
-0.8724566  0.3367946 -0.5588127  0.9007215 -3.1909590 -1.3138540  0.4325021 
    August  September    October   November   December 
-0.6143199  0.6225021  0.4449313  1.3066118  0.9270776 
```

(b)

```
> diag(crawl$n)^(1/2)%*%unweighted_res
            [,1]
 [1,]  -4.935360
 [2,]   2.020768
 [3,]  -2.679972
 [4,]   4.592796
 [5,] -16.580709
 [6,]  -7.075321
 [7,]   1.981973
 [8,]  -4.120983
 [9,]   3.837360
[10,]   2.951341
[11,]   9.146282
[12,]   6.149537
```

(c)

```
> residuals(fit2)
   January   February      March      April        May       June       July 
-0.8724566  0.3367946 -0.5588127  0.9007215 -3.1909590 -1.3138540  0.4325021 
    August  September    October   November   December 
-0.6143199  0.6225021  0.4449313  1.3066118  0.9270776 
```

The values that R residuals function compute are the unweighted residuals.