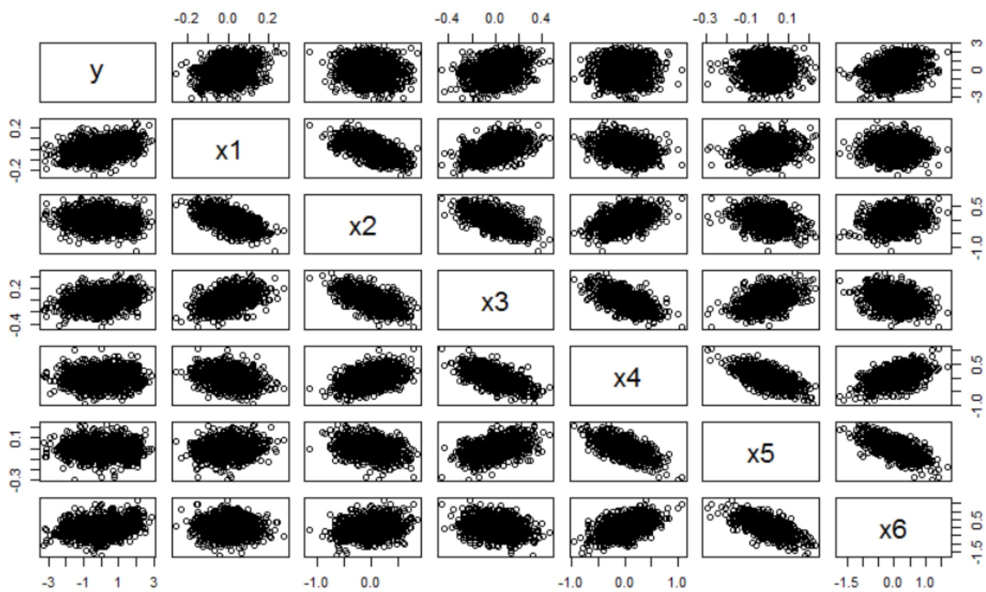
Homework #7

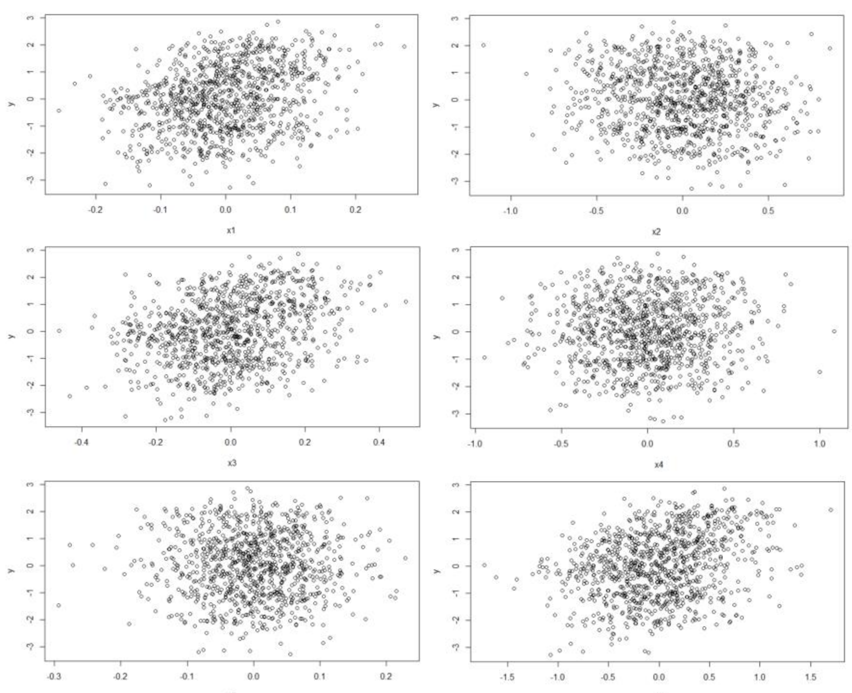
9.1.1 First draw a scatterplot matrix of all data and comment. Is there anything strange?

```
> library(alr4)
> data("Rpdata")
> pairs(y~x1+x2+x3+x4+x5+x6,Rpdata)
```



Answer: There is no strange thing found.

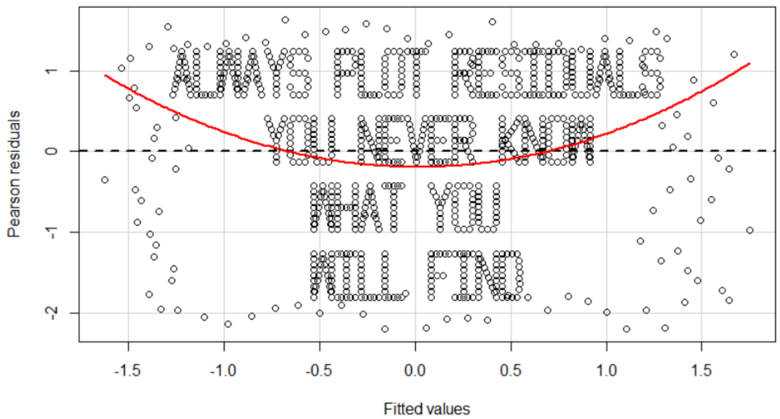
9.1.2



Answer: There is no strange thing found in each OLS fitted regression.

9.1.3

```
> residualPlot(lm(y ~ ., Rpdata))
```

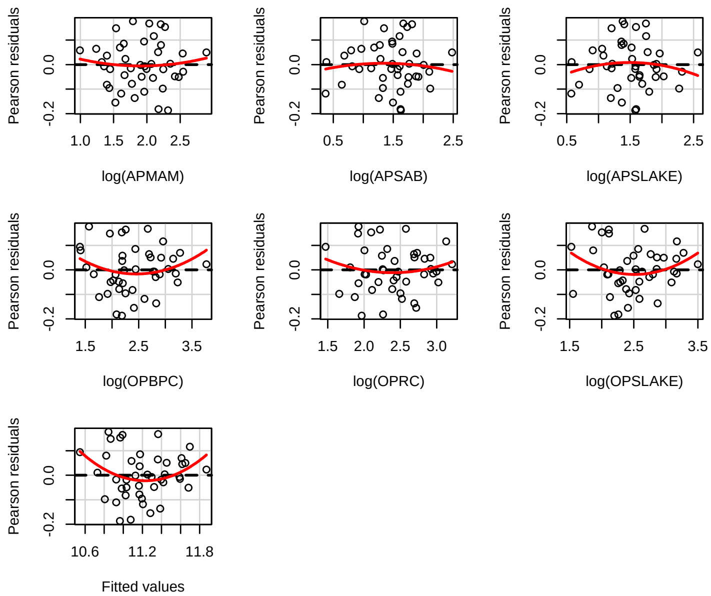


Answer: The residual plots makes a sentence "Always Plot Residuals. You never know what you will find."

9.8

```
> data(water)
> data = water
> fit = lm(log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) +
log(OPBPC) + log(OPRC) + log(OPSLAKE), data=data)
> residualPlots(fit)

             Test stat Pr(>|t|)
log(APMAM)       0.450    0.656
log(APSAB)      -0.465    0.645
log(APSLAKE)    -0.852    0.400
log(OPBPC)       1.385    0.175
log(OPRC)        0.839    0.407
log(OPSLAKE)     1.630    0.112
Tukey test       1.839    0.066
```

Answer: For the mean function, here are the residual plots.



Answer: As you can see, there are p-value data in the above data of residuals (fit). The null hypothesis is that there is no curvature, and the alternative hypothesis is that there is curvature. As a result, every p-value is larger than 0.05, which means that we do not reject the null hypothesis; there is no curvature.

9.10

```
> ehat <- c(1.000, 1.732, 9, 10.295)
> lev <- c(0.9, .75, .25, .185)
> sig <- 4
> r <- ehat/(sig*sqrt(1-lev))
> D <- (1/5)*r^2*(lev/(1-lev))
> ti <- r*sqrt((54-5-1)/(54-5-r^2))
> data.frame(ehat,lev,r,D,ti)
    ehat   lev         r         D        ti
1  1.000 0.900 0.7905694 1.1250000 0.7874992
2  1.732 0.750 0.8660000 0.4499736 0.8637532
3  9.000 0.250 2.5980762 0.4500000 2.7692308
4 10.295 0.185 2.8509366 0.3689939 3.0895439
```

Answer: As we examined in the R-script, we earned values of D and Ti for each of the cases. With analyzing data D, I can conclude that the first one is influential compared to other three (around 0.05) due to the largest D value, 1.125. And with Ti value we can also conclude which ones are outliers or not, and then we can conclude the third and fourth ones are outliers due to much higher t-values than the first two ones.

9.11

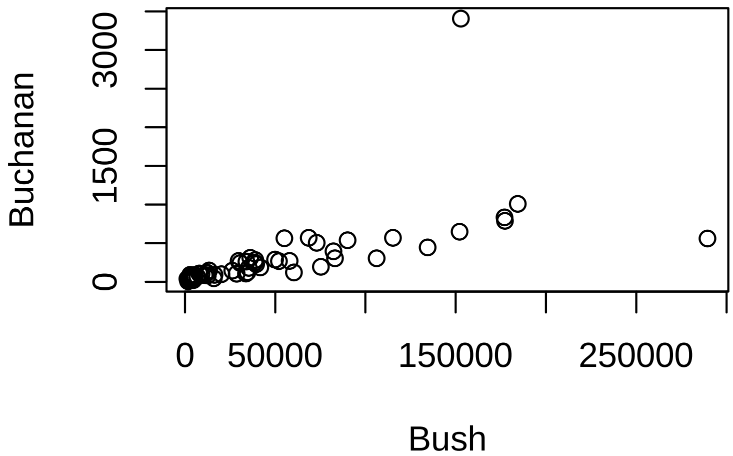
```
> library(MASS)
> data(fuel2001)
> fit = lm(FuelC ~ Tax + Drivers + Income + log(Miles),
data=fuel2001)
> summary(fit)

Call:
lm(formula = FuelC ~ Tax + Drivers + Income + log(Miles), data =
data)

Residuals:
     Min       1Q   Median       3Q      Max
-1676904  -126002   -21638   146118  1849371

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.274e+05  1.219e+06   0.186    0.853
Tax         -2.270e+04  1.436e+04  -1.581    0.121
Drivers      6.566e-01  2.198e-02  29.868   <2e-16 ***
Income      -1.820e+01  1.745e+01  -1.043    0.302
log(Miles)   7.579e+04  8.503e+04   0.891    0.377
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 453600 on 46 degrees of freedom
Multiple R-squared:  0.974,     Adjusted R-squared:  0.9717
F-statistic: 430.6 on 4 and 46 DF,  p-value: < 2.2e-16

> ehat <- c(-164.145, -137.599, -102.409, 183.499, -49.452)
> lev <- c(.256, .162, .206, .084, .415)
> s <- 64.891
> r <- ehat/(s*sqrt(1-lev))
> D <- (1/5)*r^2*(lev/(1-lev))
> t <- r*sqrt((46-1)/(46-r^2))
> data.frame(ehat,lev,r,D,t)
      ehat   lev          r         D          t
1 -164.145 0.256 -2.9326263 0.5918484 -3.2168359
2 -137.599 0.162 -2.3163746 0.2074525 -2.4376317
3 -102.409 0.206 -1.7711013 0.1627659 -1.8147106
4  183.499 0.084  2.9546191 0.1601094  3.2465847
5  -49.452 0.415 -0.9963719 0.1408527 -0.9962917

> pmin(51*2*pt(-abs(t),46),1)
[1] 0.1211242 0.9541017 1.0000000 0.1112947 1.0000000
```

Answer:
From the results, Wyoming would be the most significant outlier with the largest p-value of 3.2465847 for all five states outlier tests. Also, the state of Alaska has the largest influence in this regression because the largest D value of 0.5918484.

9.16

```
> library(alr4)
> data(florida)
> plot(Buchanan ~ Bush, data=florida)
> fit1 = lm(Buchanan ~ Bush, florida)
```

Here is the scatterplot of Buchanan versus Bush.

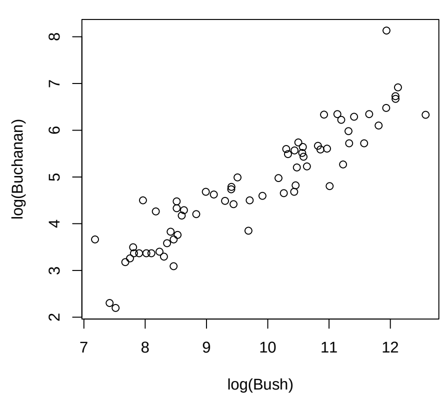



```
> outlierTest(fit1, cutoff=0.15)
```

	rstudent	unadjusted p-value	Bonferonni p
PALM BEACH	24.080144	8.6246e-34	5.7785e-32
DADE	-3.280922	1.6772e-03	1.1237e-01

Answer:
For the outlier test, we can conclude that PALM BEACH and DADE are both outliers because those p-values are less than 0.05 on the above outlier test.

```
> plot(log(Buchanan) ~ log(Bush), data=florida)
> fit2 = lm(log(Buchanan) ~ log(Bush), data=florida)
```



```
> outlierTest(fit2, cutoff=0.15)
```

	rstudent	unadjusted p-value	Bonferonni p
PALM BEACH	4.066282	0.00013325	0.0089278

Answer:
From this outlier test, we can know whether or not the data of PALM BEACH is outlier after transforming the variables. Using Bonferonnit p, the p-value of PAML BEACH is still less than 0.05. Therefore, we can predict that the PALM BEACH is still an outlier after transforming the variables.