STAT3032_Homework6

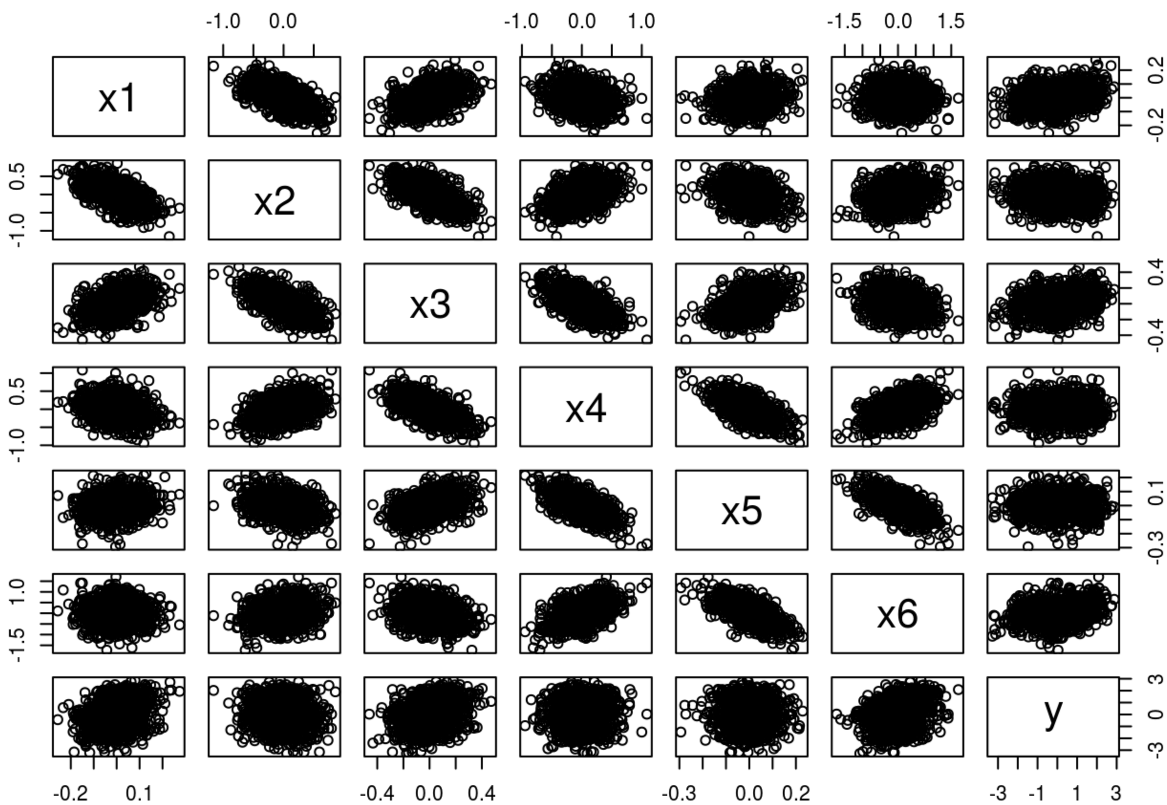
Guizhen Yu

November 29, 2017

Answer for 9.1

Answer for 9.1.1

```
pairs( ~ x1 + x2 + x3 + x4 + x5 + x6 + y, Rpdata)
```



There are some colinerarity between Xn and Xn+1 (where 1 < n < 6). Also, there is not a clear trend between regressors and Y. Otherwise, everyting is normal.

Answer for 9.1.2

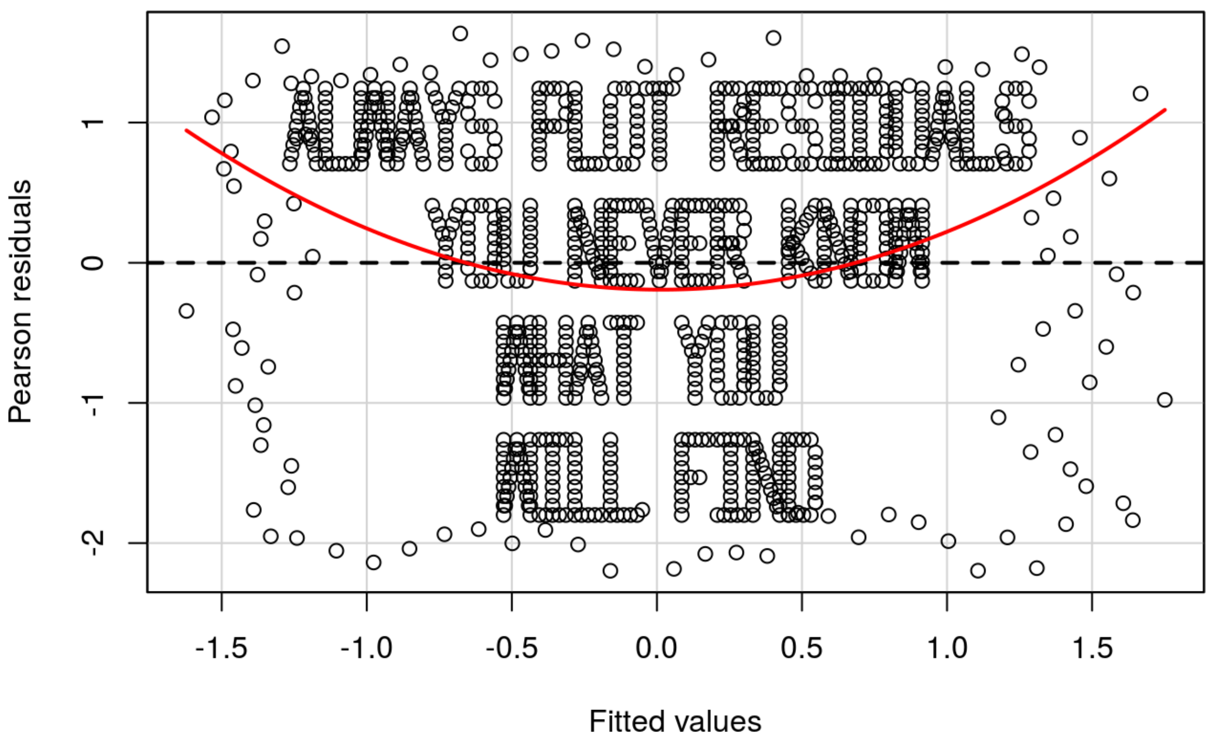
```
m912 <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
summary(m912)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = Rpdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1977 -0.7631  0.1729  0.8851  1.6359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02481    0.03188   0.778    0.437
## x1           4.14061    0.50954   8.126 1.32e-15 ***
## x2           1.01233    0.15522   6.522 1.11e-10 ***
## x3           3.99614    0.32663  12.234  < 2e-16 ***
## x4           0.96045    0.16657   5.766 1.09e-08 ***
## x5           3.75122    0.64726   5.796 9.17e-09 ***
## x6           0.95390    0.08561  11.142  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.003 on 983 degrees of freedom
## Multiple R-squared:  0.3112, Adjusted R-squared:  0.307
## F-statistic: 74.03 on 6 and 983 DF,  p-value: < 2.2e-16
```

Looking at the summary of model and I cannot find anything strange, all regressors are significant, However, R-squared is a little lower. This may not be a good model.

Answer for 9.1.3

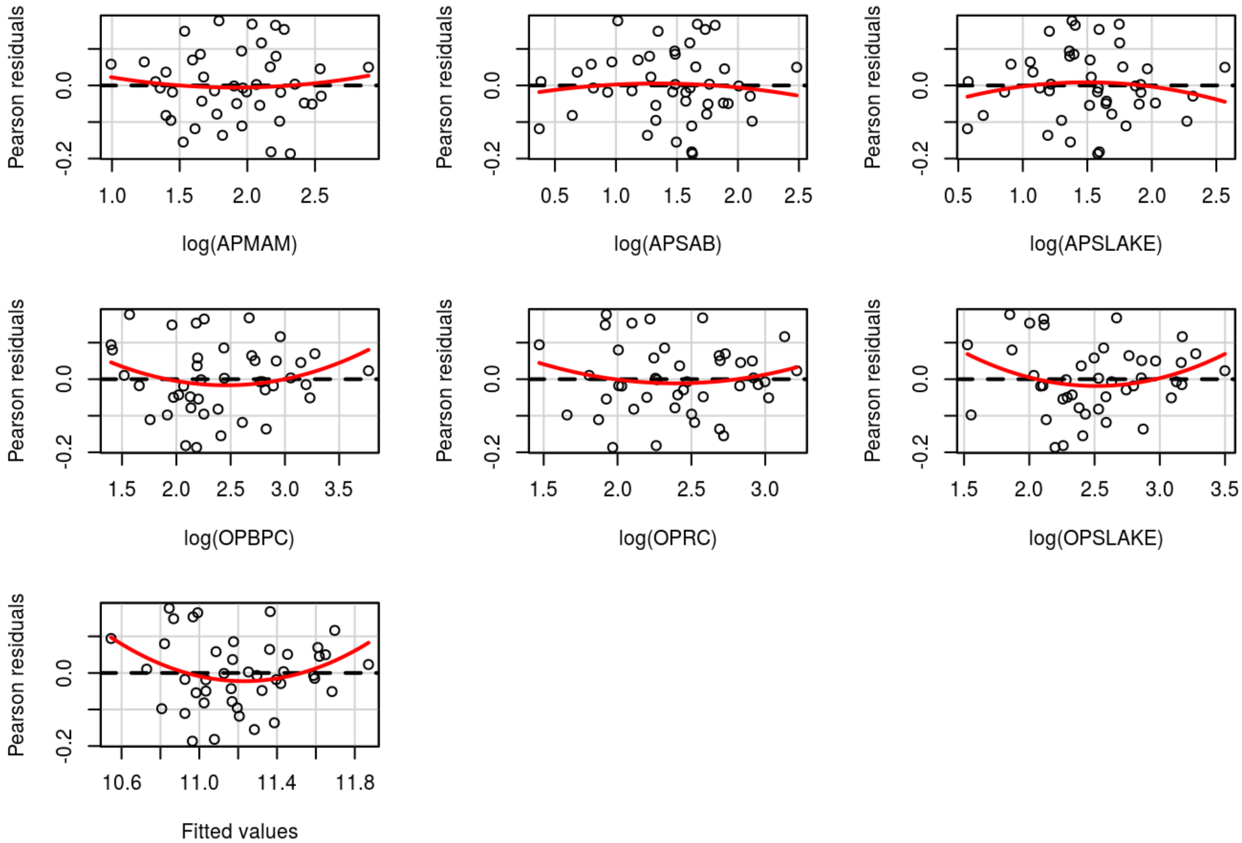
```
residualPlot(m912)
```



Yes, it is clearly that the residual plot have a strange shape. We expect residual plot to be random distributed on the graph. However, this residual plot form a sentence: “Always plot residuals, you never know what you will find”

Answer for 9.8

```
# model from 8.3.4
m834 <- lm(log(BSAAM) ~ log(APMAM) + log(APSAB) + log(APSLAKE) +
           log(OPBPC) + log(OPRC) + log(OPSLAKE), data = water)
residualPlots(m834)
```




```
##                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
```

These residual plots show that all of the regressors after transformation have a linear relationship with log(BSAAM) because from the plot, residuals are randomly distributed on all regressor's plots.

Test for curvature also show the same conclusion. For each regressor, we make a model with (long(regressor))^2, and test for significance of this squared term. All the P-values are a lot greater than 0.05, indicated that non of the squared terms are significant.

Answer for 9.10

```
ei910 = c("Case 1" = 1,
          "Case 2" = 1.732,
          "Case 3" = 9,
          "Case 4" = 10.295)

hii910 = c("Case 1" = 0.9,
           "Case 2" = 0.75,
           "Case 3" = 0.25,
           "Case 4" = 0.185)
sigma910 = 4
#ri = ei/(sigma*sqrt(1-hii))
ri910 = ei910/(sigma910*sqrt(1-hii910))
print("ri is:")
```

```
## [1] "ri is:"
```

```
ri910
```

```
##    Case 1    Case 2    Case 3    Case 4
## 0.7905694 0.8660000 2.5980762 2.8509366
```

```
pprime910 = 5
#di = (1/p-prime)*(ri^2)*(hii/(1-hii))
di910 = (1/pprime910)*(ri910^2)*(hii910/(1-hii910))
print("di is :")
```

```
## [1] "di is :"
```

```
di910
```

```
##    Case 1    Case 2    Case 3    Case 4
## 1.1250000 0.4499736 0.4500000 0.3689939
```

```
n910 = 54
# ti = ri((n-p-1)/(n-p-ri^2))^(1/2)
ti910 = ri910*((n910-pprime910-1)/(n910-pprime910-ri910^2))^(1/2)
print("ti is : ")
```

```
## [1] "ti is : "
```

```
ti910
```

```
##    Case 1    Case 2    Case 3    Case 4
## 0.7874992 0.8637532 2.7692308 3.0895439
```

```
# test for outliers
2*pt(ti910, 49, lower.tail = F)
```

```
##      Case 1      Case 2      Case 3      Case 4
## 0.434782530 0.391932001 0.007912435 0.003299794
```

```
# a* is :
print(0.05/4)
```

```
## [1] 0.0125
```

Therefore Case 3 and Case 4 are outliers. Because the p values from Case 3 and Case 4 are both smaller than 0.0125 , which is statistically significance.

Answer for 9.11

```
# Create the needed variables for this data set
fuel2001$Dlic = 1000 * fuel2001$Drivers / fuel2001$Pop
fuel2001$fuel = 1000 * fuel2001$FuelC / fuel2001$Pop
fuel2001$Income = fuel2001$Income / 1000
m911 = lm(fuel ~ Tax + Dlic + Income + log(Miles), data = fuel2001)
summary(m911)
```

```
##
## Call:
## lm(formula = fuel ~ Tax + Dlic + Income + log(Miles), data = fuel2001)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.145  -33.039    5.895   31.989  183.499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.1928   194.9062   0.791 0.432938
## Tax          -4.2280     2.0301  -2.083 0.042873 *
## Dlic          0.4719     0.1285   3.672 0.000626 ***
## Income       -6.1353     2.1936  -2.797 0.007508 **
## log(Miles)   26.7552     9.3374   2.865 0.006259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.89 on 46 degrees of freedom
## Multiple R-squared:  0.5105, Adjusted R-squared:  0.4679
## F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

```
sighat = 64.891
df = 46
params = 5
n = df + params

# all given in the book
outliers = c("Alaska",
             "New York",
             "Hawaii",
             "Wyoming",
             "District of Columbia")
fuelEst = c("Alaska" = 514.279,
            "New York" = 374.164,
            "Hawaii" = 426.349,
            "Wyoming" = 842.792,
            "DC" = 317.492)        # estimated response
ehat = c("Alaska" = -163.145,
         "New York" = -137.599,
         "Hawaii" = -102.409,
         "Wyoming" = 183.499,
         "DC" = -49.452)          # residual: observed - predicted
hii = c("Alaska" = 0.256,
        "New York" = 0.162,
        "Hawaii" = 0.206,
        "Wyoming" = 0.084,
        "DC" = 0.415)
ri = ehat/(sighat * sqrt(1-hii))
ti = ri*sqrt((n - params - 1)/(n-params-ri^2))
(pvals = 2 * pt(abs(ti), df = n-params-1, lower = F))
```

```
##      Alaska    New York      Hawaii     Wyoming          DC
## 0.002571896 0.018798601 0.076237504 0.002208895 0.324435787
```

```
(adj.pvals = pmin(n*pvals,1))
```

```
##    Alaska  New York    Hawaii   Wyoming        DC
## 0.1311667 0.9587286 1.0000000 0.1126537 1.0000000
```

```
Di = (1 / params) * ri^2 * (hii / (1 - hii))    # Cooks distance
print("Di is: ")
```

```
## [1] "Di is: "
```

```
Di
```

```
##    Alaska  New York    Hawaii   Wyoming        DC
## 0.5846591 0.2074525 0.1627659 0.1601094 0.1408527
```

```
out1 = outlierTest(m911, cutoff = Inf, n.max = Inf)
states = c("AK", "NY", "HI", "WY", "DC")

print("Ti is: ")
```

```
## [1] "Ti is: "
```

```
out1$rstudent[states]
```

```
##         AK         NY         HI         WY         DC
## -3.1930222 -2.4382246 -1.8143653  3.2460899 -0.9962102
```

```
print("Pi is: ")
```

```
## [1] "Pi is: "
```

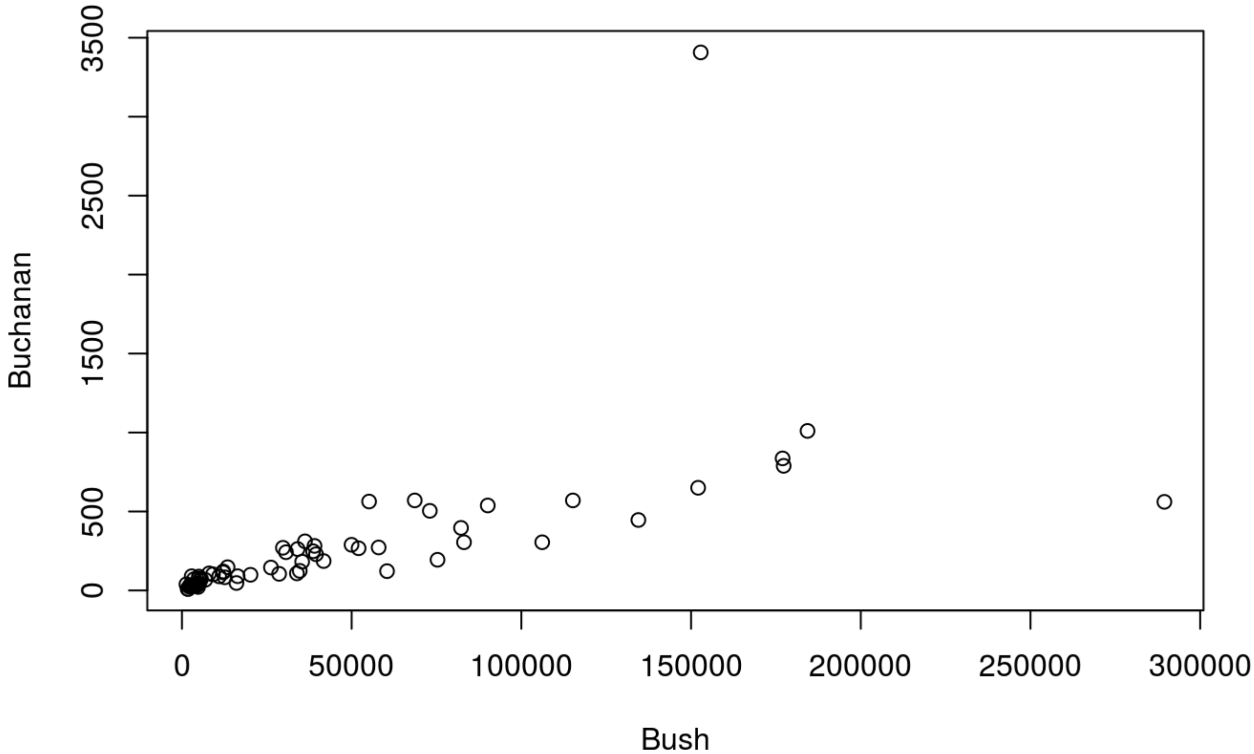
```
out1$p[states]           # this is the same as pvals above
```

```
##          AK          NY          HI          WY          DC
## 0.002570159 0.018771476 0.076291494 0.002212001 0.324474923
```

Therefore, Alaska is the most influential point because it have the hingest di.

Answer for 9.16

```
plot(Buchanan ~ Bush, data = florida)
```




```
m916 <- lm(Buchanan ~ Bush, data = florida)
print("Based on the graph, pinellas is also an unusual value.")
```

```
## [1] "Based on the graph, pinellas is also an unusual value."
```

```
out916 <- outlierTest(m916,cutoff = Inf, n.max = Inf)
out916$p["PINELLAS"]
```

```
##  PINELLAS
## 0.8627254
```

```
# Therefore, pinellas is not an outlier.

m916T <- lm(log(Buchanan) ~ log(Bush), data = florida)
out916T <- outlierTest(m916T)
out916T
```

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

```
print("From this test, palm beach is an outlier")
```

```
## [1] "From this test, palm beach is an outlier"
```