# hw12

*Dandong Tu*

*2017/12/5*

**1.**

(1)

(a)

$$\sum_{i=1}^{n} h_{ii} = tr(H) = tr(X(X^TX)^{-1}X^T)$$
$$= tr(X^TX(X^TX)^{-1})$$
$$= tr(I_{p'})$$
$$= p'$$

(b) Since $E(\hat{e}_i)=0$, so we have $\hat{e}' \cdot \underset{\sim}{1} = 0$ where $\sum(\hat{e}_i) = 0$

$\hat{e} = Y - \hat{Y} = Y - X(X^TX)^{-1}X^TY = (I-H)Y$ $\quad \hat{e}' \cdot \underset{\sim}{1} = Y'(I-H) \cdot \underset{\sim}{1}$

So, $(I-H)\underset{\sim}{1} = 0$ and $H \cdot \underset{\sim}{1} = \underset{\sim}{1}$

therefore, $\underset{\sim}{1} = \sum_{j=1}^{n} h_{ij}$, Since $H$ is symmetric, $h_{ij} = h_{ji}$

Thus, $\sum_{j=1}^{n} h_{ij} = \sum_{i=1}^{n} h_{ij} = 1$

(c)

$h_{ii} = \frac{1}{n} + (x_i^* - \bar{x})'(X'X)^{-1}(x_i^* - \bar{x}) = \frac{1}{n} + \|A(x_i^* - \bar{x})\|^2$, where $A = (X'X)^{-\frac{1}{2}}$

So, $\|A(x_i^* - \bar{x})\|^2 > 0$ ; $\frac{1}{n} + \|A(x_i^* - \bar{x})\|^2 \geq \frac{1}{n}$ and $h_{ii} > \frac{1}{n}$

Since $H$ is idempotent,

$$h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji} = \sum_{j=1}^{n} h^2_{ij} = \sum_{i \in J_i} h_{ij}^2 + h_{ij}^2 \qquad J_i = \{ i \in 1 \ldots n \mid x_i = x_j \}$$
$$= \#\{J_i\} h_{ii}^2 + \sum_{j \in J} h_{ij}^2 \qquad x_i = x_j \Rightarrow h_{ij} = h_{ji} = h_{ii}$$
$$= r h_{ii}^2 + \sum_{j \in J} h^2_{ij} \qquad r \text{ is the \# of rows}$$
$$\geq n h_{ij}^2$$

Thus $\frac{1}{n} \leq h_{in} \leq \frac{1}{r}$ for $i = 1, \ldots n$

# 2

```r
library(alr4)
```

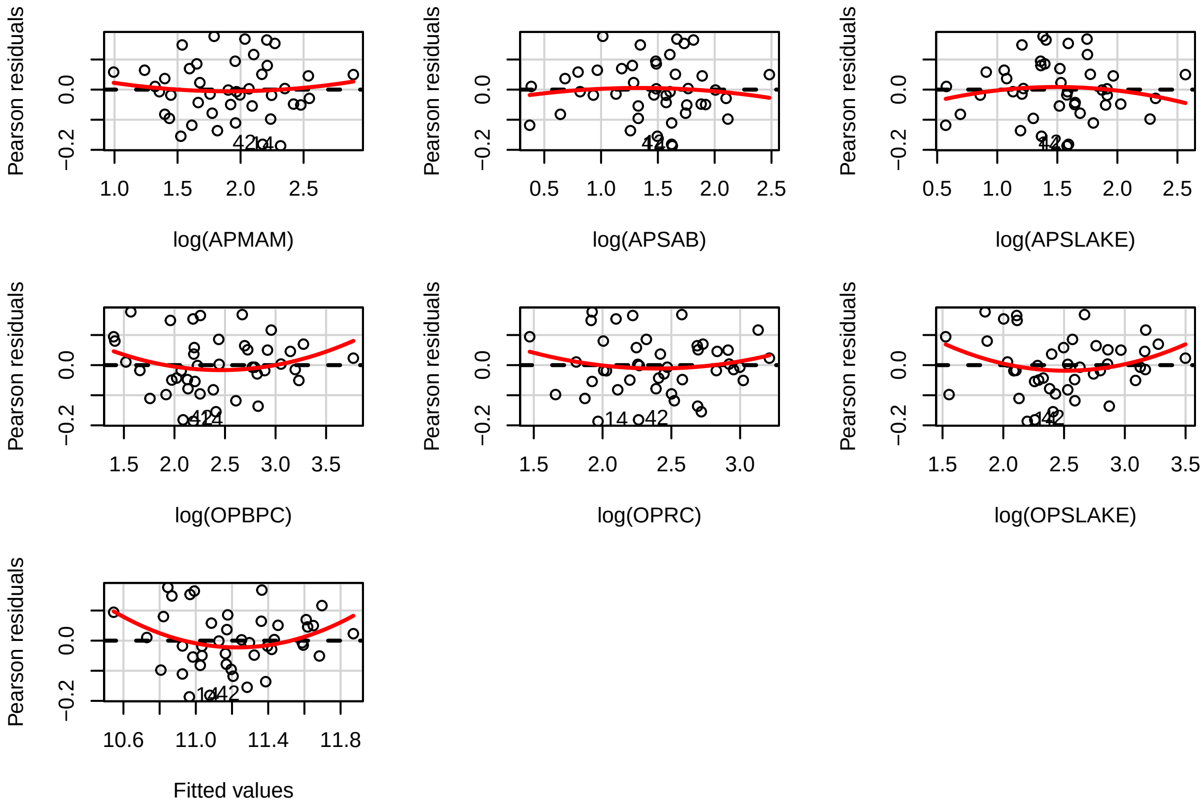
```
## Loading required package: car

## Loading required package: effects

##
## Attaching package: 'effects'

## The following object is masked from 'package:car':
##
##     Prestige
```

```r
m1 = lm(formula = log(BSAAM)~log(APMAM)+log(APSAB)+log(APSLAKE)+log(OPBPC)+log(OPRC)+log(OPSLAKE),data=
rp1=residualPlots(m1,id.n=2)
```




The residual plots shows that it seems to be a null plot.

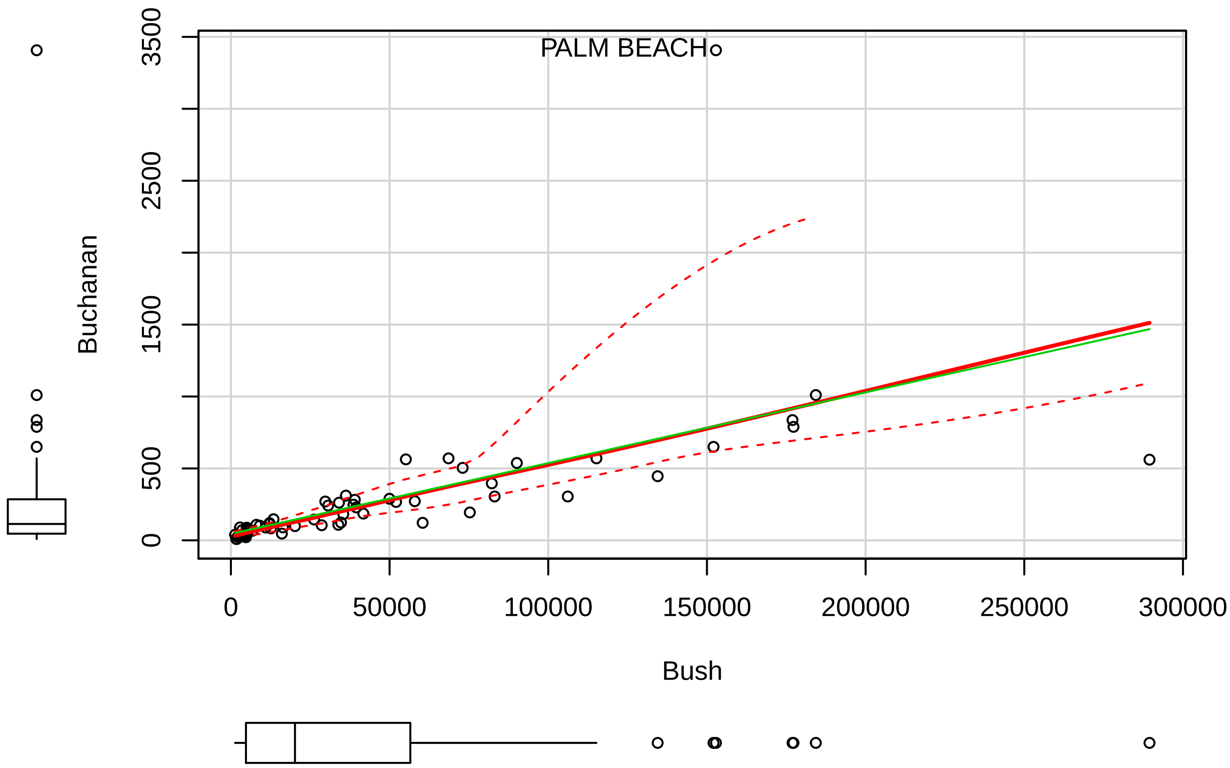
```r
rp1
```

```
##              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
```

None of the tests has small significance levels, provding no evidence against the mean function. We do not have enough evidence to reject the H0 that there is no curvature.

# 3

```
m2=lm(Buchanan~Bush,data=florida)
scatterplot(Buchanan~Bush,data= florida,id.n=1)
```




```
## PALM BEACH
##         50
```

The Scatterplot shows **PALM BEACH** is an outlier.

```
outlierTest(m2)
```

```
##            rstudent unadjusted p-value Bonferonni p
## PALM BEACH 24.08014         8.6246e-34   5.7785e-32
```

```
cd2=cooks.distance(m2)
cd2[50]
```

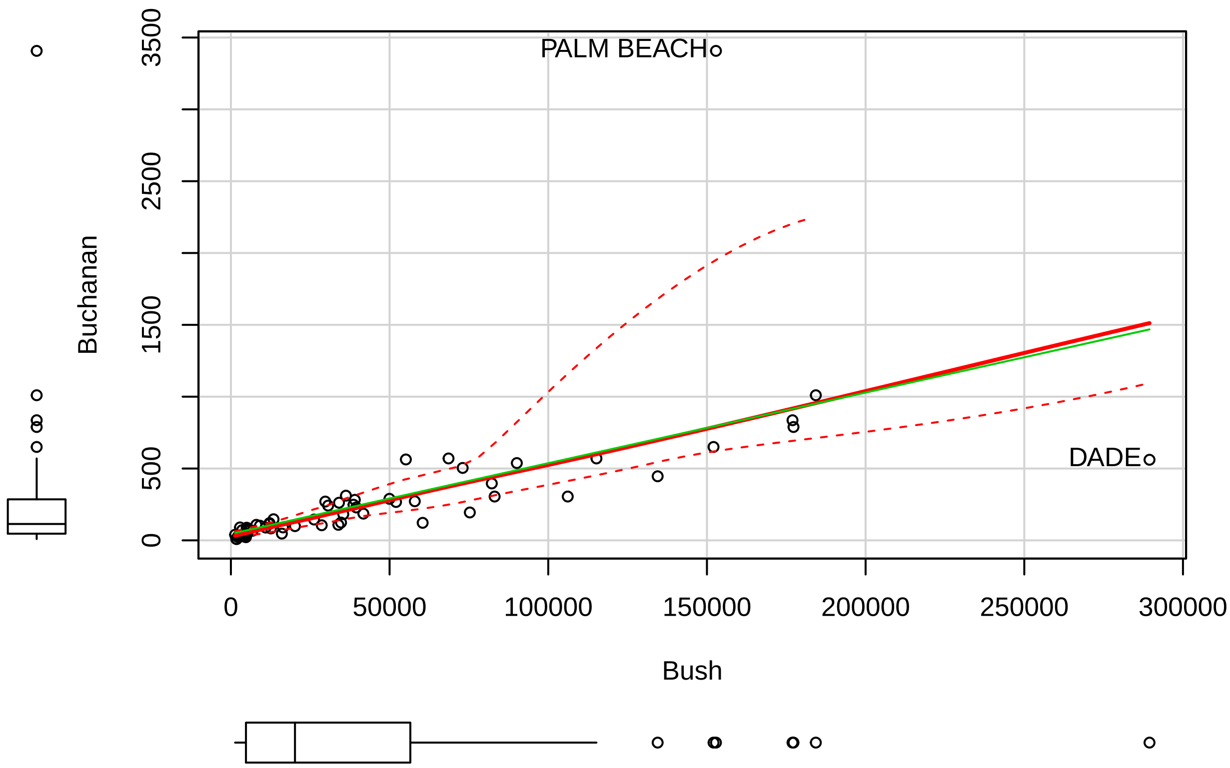
```
## PALM BEACH
##   2.231935
```

```
mean(cd2)
```

```
## [1] 0.06391973
```

Based on the test, we obtained a very small p-value that is an indication that it is an outlier. From the cook test for the city **PALM BEACH**, compare with the mean of **cd2** we observed a high value, and we conclue that the city **PALM BEACH** has a very high chance it is an outlier.

```r
scatterplot(Buchanan~Bush,data= florida,id.n=2)
```




```
##        DADE PALM BEACH
##          13         50
```

It seems another country with an unusal value of the Buchanan vote, given its **Bush** value, is **DADE**

```r
cd2[13]
```

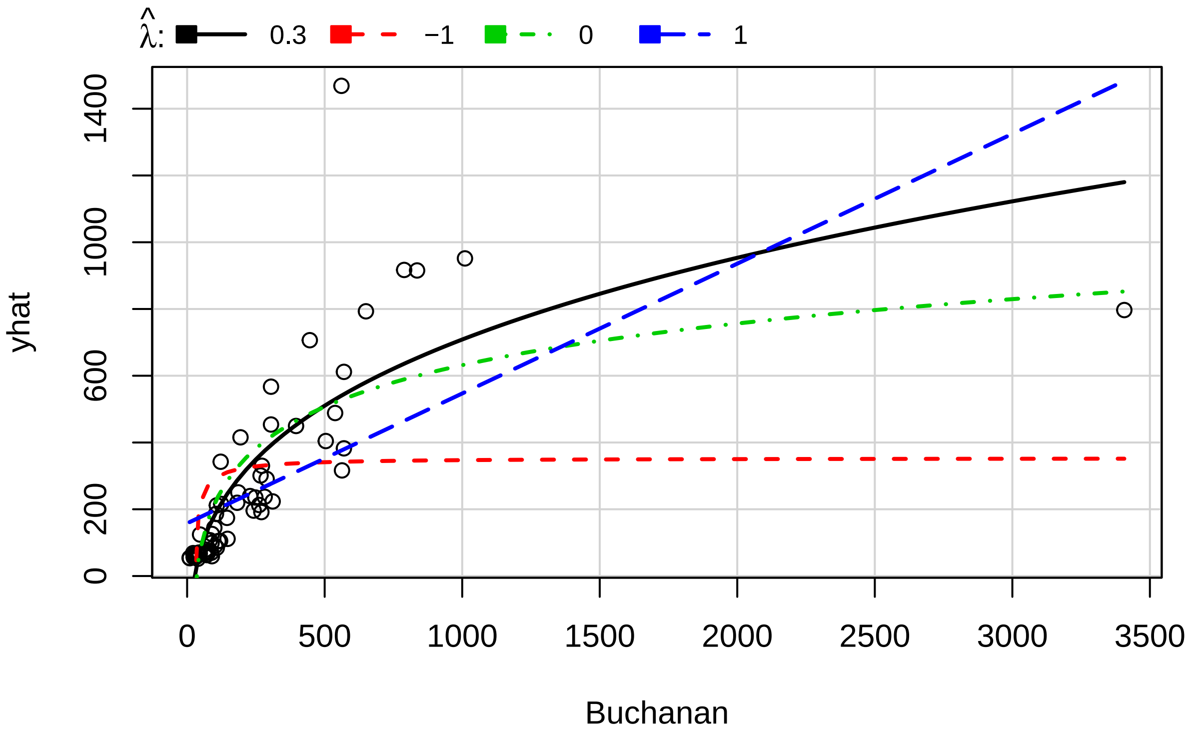
```
##     DADE
## 1.981366
```

```r
mean(cd2)
```

```
## [1] 0.06391973
```

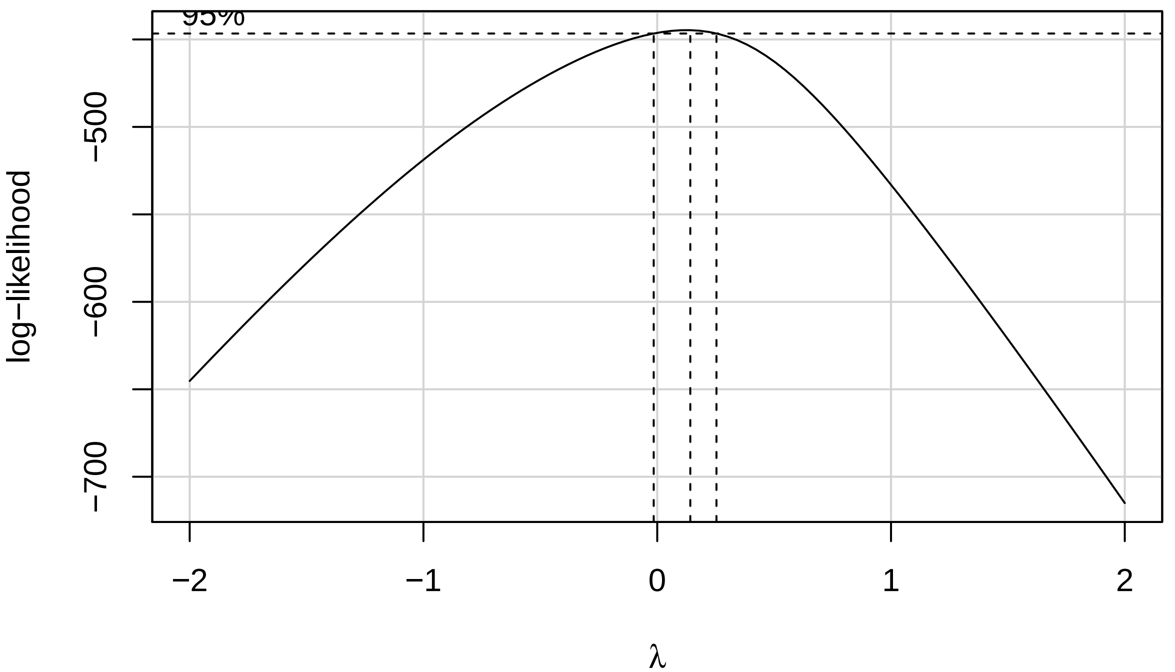
Same as the previous one, the distance value of the city **DADE** is large, we conclude that the city **DADE** has the high chance to be an outlier. It seems the butterfly ballot do have the issue of vote.

```r
inverseResponsePlot(m2)
```



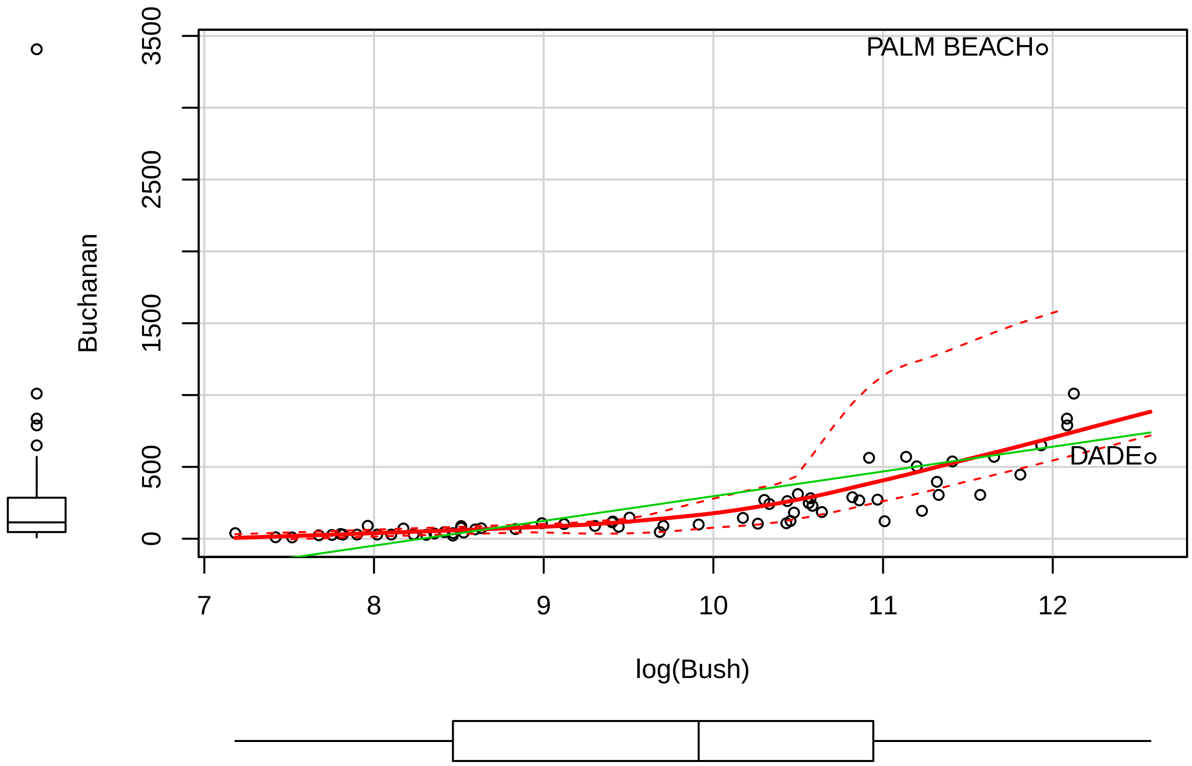

```
##       lambda     RSS
## 1  0.304032 1807862
## 2 -1.000000 4165308
## 3  0.000000 2099565
## 4  1.000000 3166621
```

```r
boxCox(m2)
```




From the graph we observe that the best lambda is about 0.3. And the boxCox shows that 0 is concluded under the 95% confidence interval. Therefore, we use **log** transformation to better fit a simple linear regression.

```r
m3=lm(Buchanan~log(Bush),data=florida)
scatterplot(Buchanan~log(Bush),data= florida,id.n=2)
```




```
##       DADE PALM BEACH
##         13         50
```

```r
outlierTest(m3)
```

```
##            rstudent unadjusted p-value Bonferonni p
## PALM BEACH 22.25891         7.7958e-32   5.2232e-30
```

```r
cd3=cooks.distance(m3)
cd3[50]
```

```
## PALM BEACH
##   1.448178
```

```r
mean(cd3)
```

```
## [1] 0.0242152
```

Based on the cook distance test, and compared with mean value, the city **PALM BEACH** is still have large distance value and we still conclude that it has a high chance to be an outlier.

```r
cd3[13]
```

```
##        DADE
## 0.009202589
```

```r
mean(cd3)
```

```
## [1] 0.0242152
```

Comparing the prevous steps to test the city **DADE**, We found that the DADE now has the distance below the mean, and we conclude that after the transformation, the city **DADE** seems no longer to be an outlier.