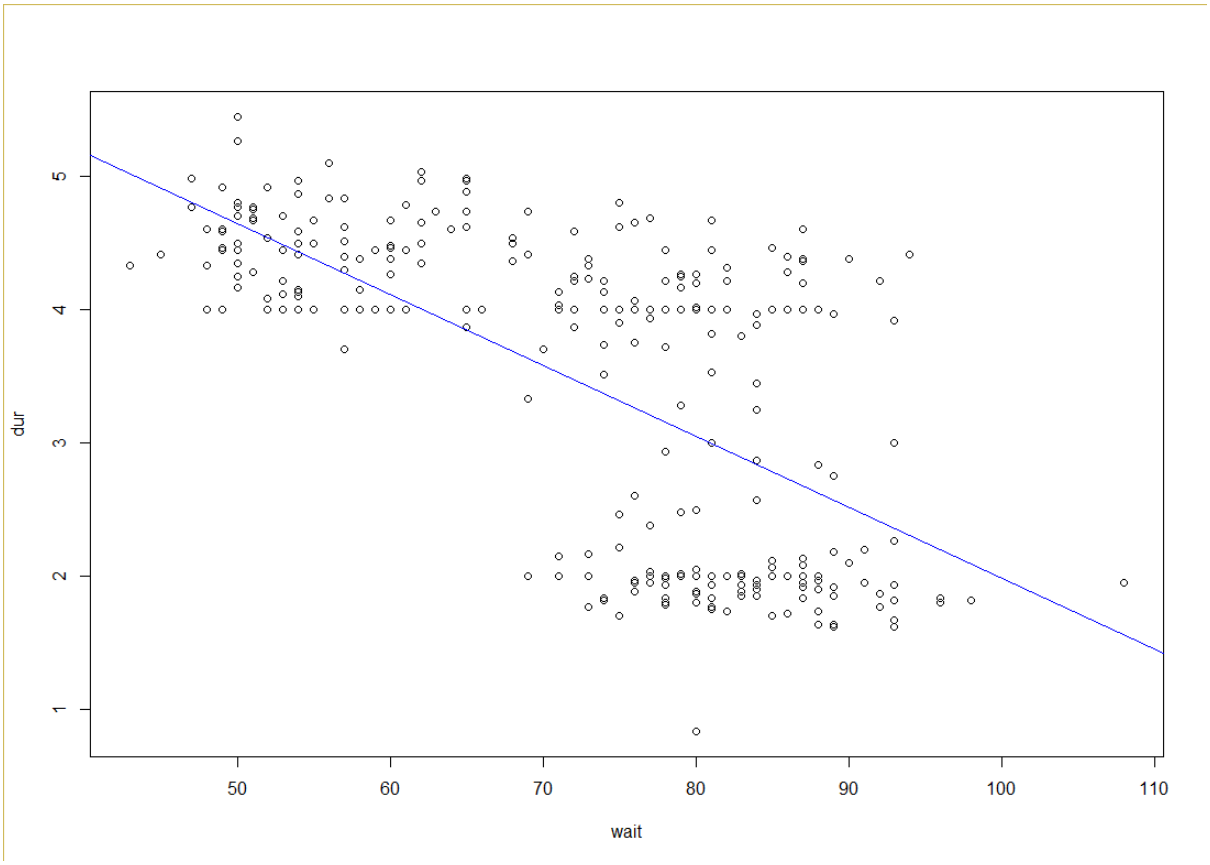
<회귀분석 과제> - 201511646 나여영

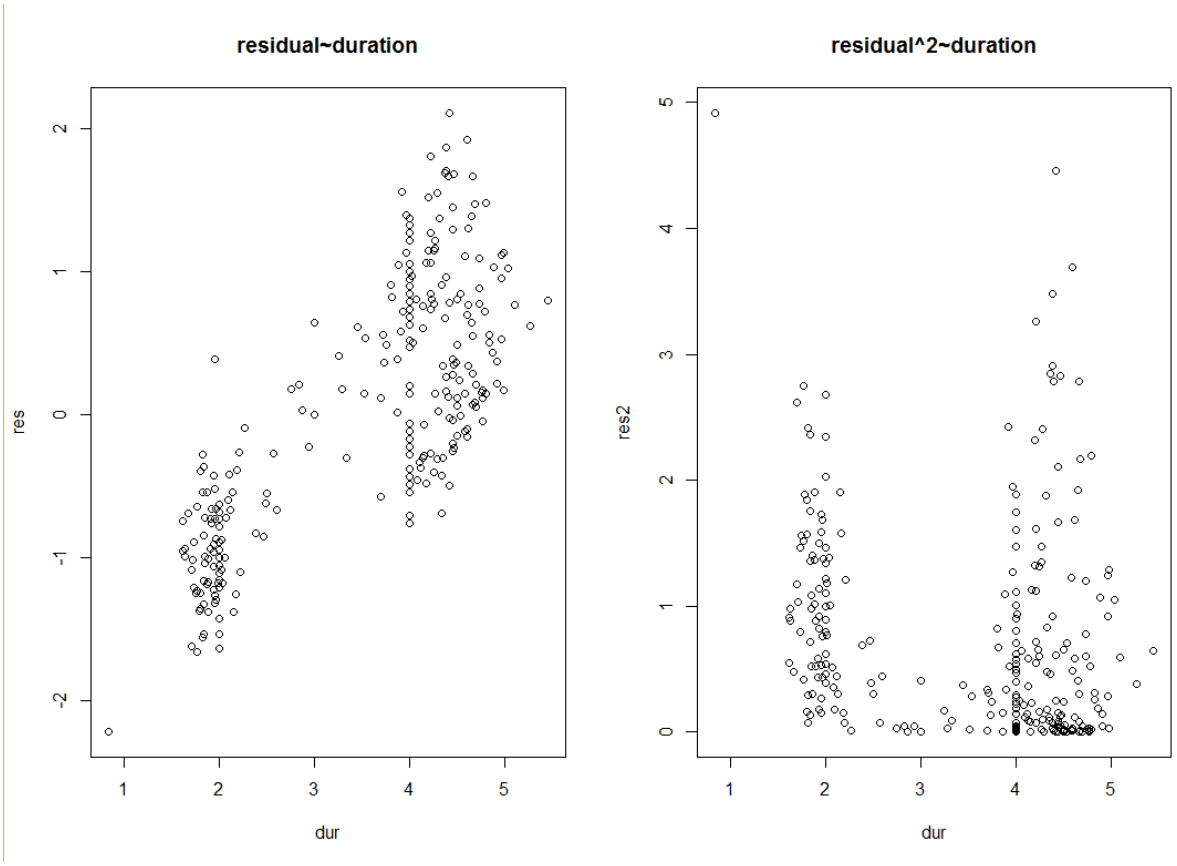
#Problem1.

A. • 자료의 산점도와 / 분출간격을 분출시간에 선형회귀모형을 최소제곱법으로 적합한 회귀선을 그리시오. 산점도와 회귀결과의 특이점을 설명하시오.



회귀식이 적절하지 않아 보인다. 모든 데이터를 커버하는?대표하는 식이라고 판단하기는 조금 어려워 보인다. 데이터가 한쪽으로 쏠려있는듯한 느낌을 받았고 추측하건데 Fitting 식 결과 선형식보단 이차식과 같은 곡선형태의 식을 fitting 하는 것이 더 적절해 보인다.

B. 구해진 잔차의 제곱과 분출시간의 산점도를 그리고 발견한 현상을 기술하시오.

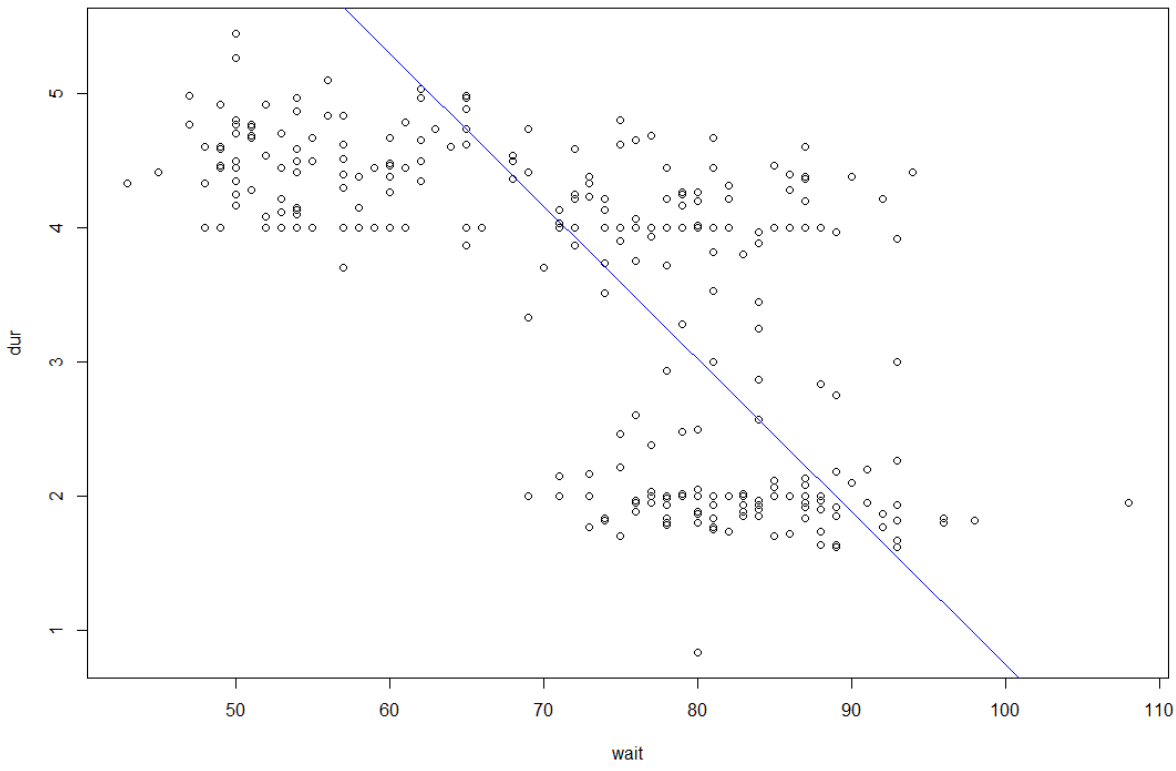


그냥 잔차plot을 보았을 때 잔차가 약간 퍼져서 약간 선형으로? 분산되어있다는 느낌을 준다. 잔차제곱을 plotting 한 결과를 보았을 때 x축과 가까이 골고루 쌓여있어야 정상인데 그렇지 않은걸 보아 적합하지 않은 식이 fitting 된 결과라고 볼 수 있다.

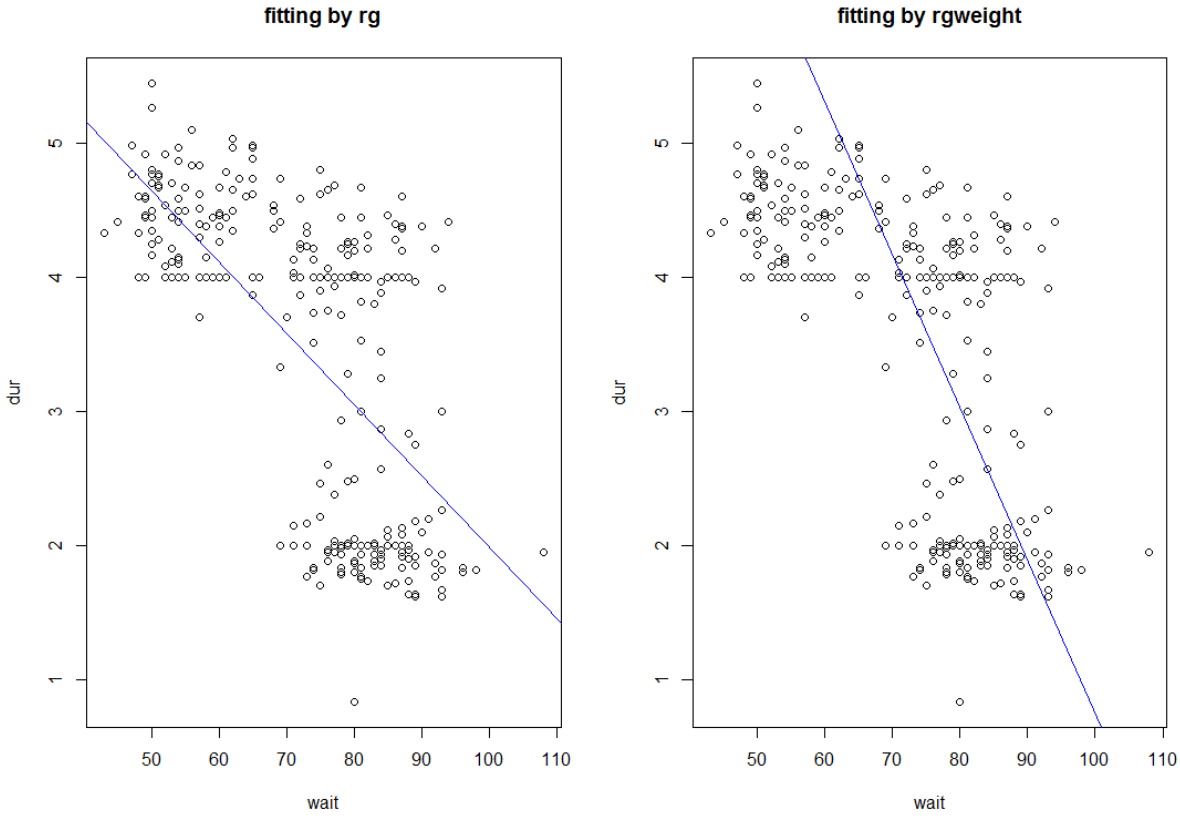
C. B에서 관찰한 내용에 따라 A의 적합의 잠재적 문제가 무엇인지 설명하고 이를 해결하기 위한 방법으로 가중최소제곱법을 적용해보시오.

문제 : 등분산성을 만족하지 않는다고 본다. 만족하지 않는경우 multicollinearity 문제를 일으킬 수 있다.

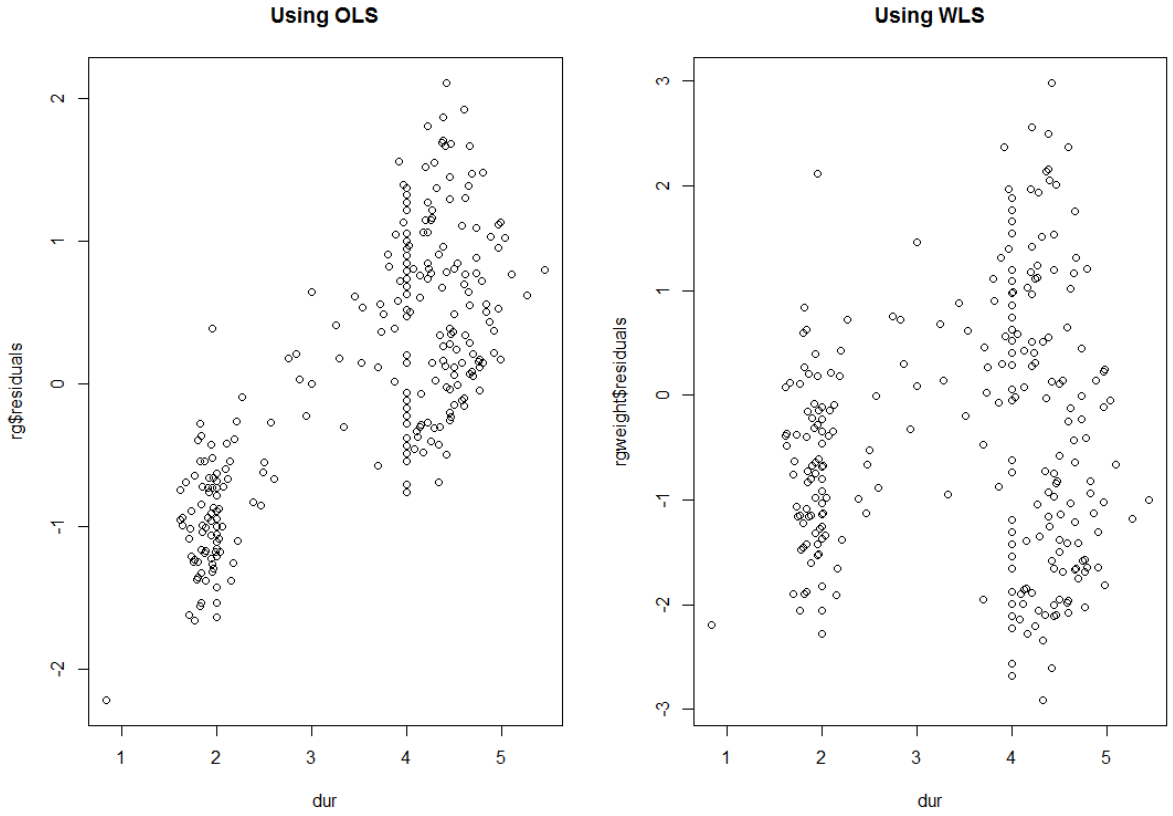
```
> rgweight

Call:
lm(formula = wait ~ dur, weights = (1/((dur - mean(dur))^2)))

Coefficients:
(Intercept)          dur
     131.35       -13.98
```



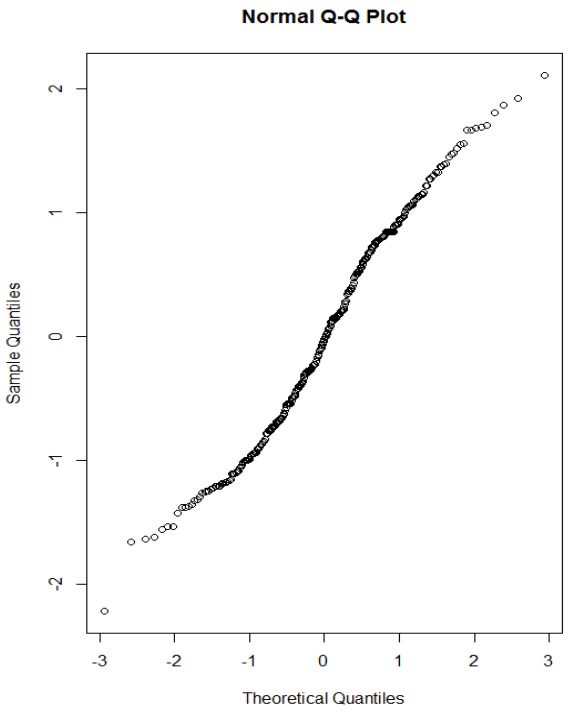
D. A와 C를 비교하여 설명하시오

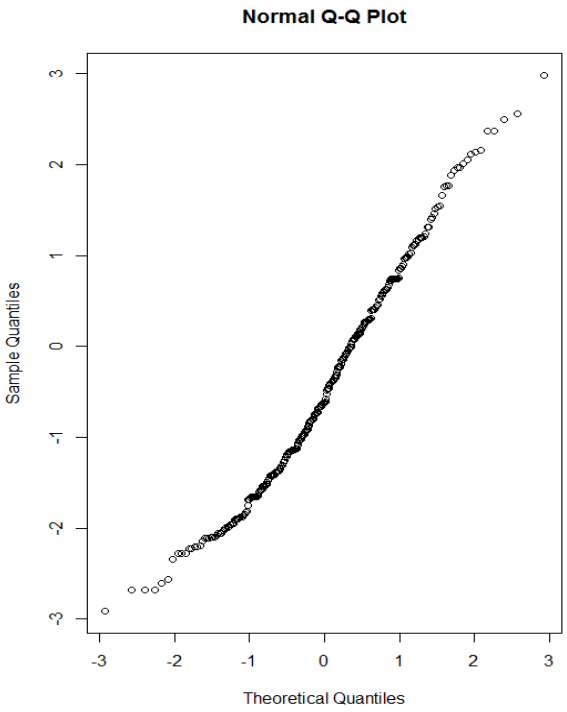


단순히 plotting 된 그림만 보아도 자료를 더 잘 설명하고 있는 것 같다는 느낌을 준다. 회귀선을 기준으로 데이터의 분포가 왼쪽보다 오른쪽이 조금 더 좌우가 균등하게 있다.



조금 더 면밀히 보기 위해 잔차를 비교해보면 가중치를 줌으로써 잔차가 0을기준으로 고르게 흩어져 있다. (물론 완벽하진 않지만 많이 해소되었다)





왼쪽을 보면 원래 왼쪽이 극단적으로 떨어져있던 것이 많이 해소 된 것을 볼 수 있다.

#Problem2

A. 부분F검정법에 기반한 후진제거방법을 이용하여 적합한 모형을 구하고 결과를 설명하시오.

```
> for(n in 1:ncol(data2)-1){
+   if(n==1) {
+     rgg=lm(y~., data2)
+   }else{
+     rgg=lm(y~.,new_data2)
+   }
+
+   re_rgg=drop1(rgg, test ="F")
+
+   f.val = as.numeric(na.omit(re_rgg$`F value`))
+   p.val = as.numeric(na.omit(re_rgg$`Pr(>F)`))
+
+   if(all(f.val > 2)){
+     result = rgg=lm(y~ . ,new_data2)
+     break
+   }
+
+   if(n==1){
+     new_data2=data2[,-(which(f.val==min(f.val))+1)]
+   }else{
+     new_data2=new_data2[,-(which(f.val==min(f.val))+1)]
+   }
+
+
+   }
> result

Call:
lm(formula = y ~ ., data = new_data2)

Coefficients:
(Intercept)           V2           V3           V5
     1.5634       2.5033       1.2208       0.7487

>
```

Y=1.5634+2.5033*V2+1.2208*V3+0.7487*V5+e 모형이 선택된 것을 확인할 수 있다.

```
> summary(result)

Call:
lm(formula = y ~ ., data = new_data2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81962 -0.48297  0.02251  0.47452  1.67474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.56342    0.05249   29.78   <2e-16 ***
V2           2.50334    0.02332  107.34   <2e-16 ***
V3           1.22079    0.01453   84.03   <2e-16 ***
V5           0.74874    0.02942   25.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7312 on 196 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9902
F-statistic:  6732 on 3 and 196 DF,  p-value: < 2.2e-16
```

Adjusted R-squared 를 통해 99%의 설명력을 가지는 것을 확인할 수 있고, 각 변수들이 매우 유의함을 알 수 있다.

(f.val>2일 때 멈추게 한 이유는, backward elimination 방법은 f값이 작은 값을 갖는 변수를 하나씩 제거해 나아가는 건데 아래의 함수결과를 보면 F값이 3개가 나머지에 비해 극단적으로 크고 나머지 변수들은 모두 2를 넘지 않기때문이다.)

```
> drop1(rggg, test ="F")
Single term deletions

Model:
y ~ V2 + V3 + V4 + V5 + V6 + V7 + V8 + V9 + V10 + V11
       Df Sum of Sq    RSS     AIC   F value Pr(>F)
<none>               102.8 -111.05
V2      1    5019.8 5122.6  668.62 9226.5242 <2e-16 ***
V3      1    3601.4 3704.3  603.79 6619.6050 <2e-16 ***
V4      1       1.0  103.8 -111.15    1.8041 0.1808
V5      1     284.8  387.6  152.33  523.4082 <2e-16 ***
V6      1       0.1  102.9 -112.90    0.1415 0.7072
V7      1       0.4  103.3 -112.20    0.8050 0.3707
V8      1       0.4  103.3 -112.23    0.7824 0.3775
V9      1       0.1  102.9 -112.91    0.1329 0.7158
V10     1       0.1  102.9 -112.85    0.1963 0.6582
V11     1       0.0  102.8 -113.01    0.0376 0.8465
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

B. 수정결정계수에 기반한 전진선택방법을 이용하여 적합한 모형을 구하고 결과를 설명하시오.

```
> for(n in 1:(ncol(data2)-1)){
+   k[n]=summary(lm(y~data2[,n+1], data2))$adj.r.squared
+
+   }
> k
 [1]  0.606517221  0.397454580  0.001420303  0.023285511  0.017715432  0.063643372 -0.002055118  0.063993004
 [9] -0.004063679 -0.003748222
> which(k==max(k))
[1] 1
```

변수 한 개가 포함된 모형에선 첫 번째 변수인 V2가 가장 높은 수정결정계수를 가지기 때문에 이 V2가 변수로 채택되고 lm(y~V2) 모형이 선택되었다.

V2 를 포함하여 두개의 변수로 이루어진 모형 중 가장 높은 수정결정계수를 가진 모형은

```
> names(data2[,2:11])
 [1] "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11"
> k=c()
> for(n in 2:(ncol(data2)-1)){
+     k[n]=summary(lm(y~V2+data2[,n+1], data2))$adj.r.squared
+
+   }
> k
 [1]        NA 0.9582000 0.6059283 0.6404913 0.6087106 0.6179381 0.6049638 0.6101262 0.6102084 0.6045683
> which(k==max(k, na.rm=TRUE))
[1] 2
```

2번째 즉 "V3"이 선택되었다. (V2+V3의 결과 95퍼센트의 설명력을 지닌다)

V2와 V3을 포함하여 세개의 변수로 이루어진 모형 중 가장 높은 수정결정계수를 가진 모형은

```
> names(data2[,2:11])
 [1] "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11"
> k=c()
> for(n in 3:(ncol(data2)-1)){
+     k[n]=summary(lm(y~V2+V3+data2[,n+1], data2))$adj.r.squared
+
+   }
> k
 [1]        NA        NA 0.9603705 0.9902415 0.9583724 0.9589043 0.9583719 0.9579915 0.9607106 0.9581711
> which(k==max(k, na.rm=TRUE))
[1] 4
> k[4]
[1] 0.9902415
```

4번째 즉 "V5"가 선택되었다.

V2+V3+V5를 적합시킨 결과 99퍼센트의 설명력을 지닌다.

변수선택을 그만하기 전에 확인차 변수 하나를 더 추가하여 명령문을 돌려보면

```
> names(data2[,2:11])
 [1] "V2"  "V3"  "V4"  "V5"  "V6"  "V7"  "V8"  "V9"  "V10" "V11"
> k=c()
> for(n in c(3,5,6,7,8,9,10)){
+     k[n]=summary(lm(y~V2+V3+V5+data2[,n+1], data2))$adj.r.squared
+
+   }
> k
 [1]        NA        NA 0.9902587        NA 0.9902018 0.9902276 0.9902269 0.9901977 0.9902031 0.9901917
> which(k==max(k, na.rm=TRUE))
[1] 3
> k[3]
[1] 0.9902587
```

3번째 즉 "V4"를 추가하게 되고 lm(y~V2+V3+V4+V5)는 V2+V3+V5를 적합 시킨 모형보다 대략 0.0001정도 나아진 효과를 보이므로 굳이 추가할 필요 없이 Y~V2+V3+V5 를 적합 시키면 된다.

식을 적합시킨 결과는 다음과 같다.

```
> lm(y~V2+V3+V5, data2)

Call:
lm(formula = y ~ V2 + V3 + V5, data = data2)

Coefficients:
(Intercept)           V2           V3           V5
     1.5634       2.5033       1.2208       0.7487
```

결과설명은 부분F검정을 통해 나온 식과 동일하기 때문에 생략한다.

C. AIC에 기반한 단계적 회귀적합 방법을 이용하여 적합한 모형을 구하고 결과를 설명하시오

```
Step:  AIC=-121.25
y ~ V2 + V3 + V5

       Df Sum of Sq    RSS     AIC
<none>               104.8 -121.25
- V5    1     346.4  451.2  168.72
- V3    1    3775.8 3880.6  599.09
- V2    1    6161.3 6266.1  694.92

Call:
lm(formula = y ~ V2 + V3 + V5, data = data2)

Coefficients:
(Intercept)           V2           V3           V5
     1.5634       2.5033       1.2208       0.7487
```

Y=1.5634+2.5033*V2+1.2208*V3+0.7487*V5+e 모형이 선택된 것을 확인할 수 있다.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.81962 -0.48297  0.02251  0.47452  1.67474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.56342    0.05249   29.78   <2e-16 ***
V2           2.50334    0.02332  107.34   <2e-16 ***
V3           1.22079    0.01453   84.03   <2e-16 ***
V5           0.74874    0.02942   25.45   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7312 on 196 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9902
F-statistic:  6732 on 3 and 196 DF,  p-value: < 2.2e-16
```

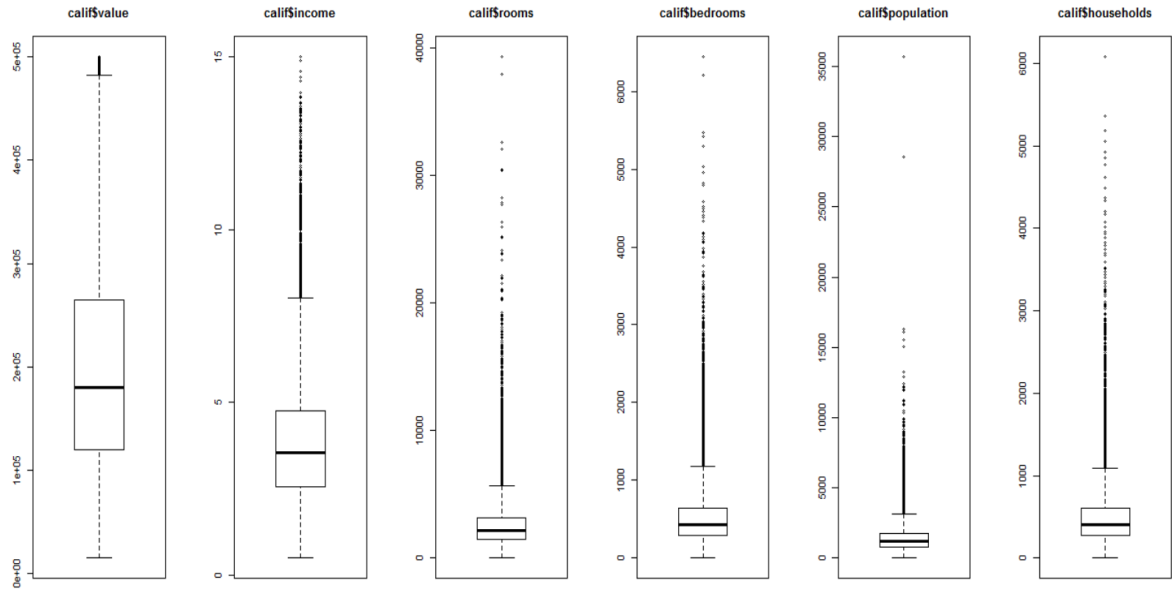
결과 설명 역시 위의 두 방법들과 동일하다.

Problem3.

A. 각 변수들의 특징을 요약하시오. 이중 특이한 관측치가 있으면 그 관측치를 보고하고 이유를 설명하시오

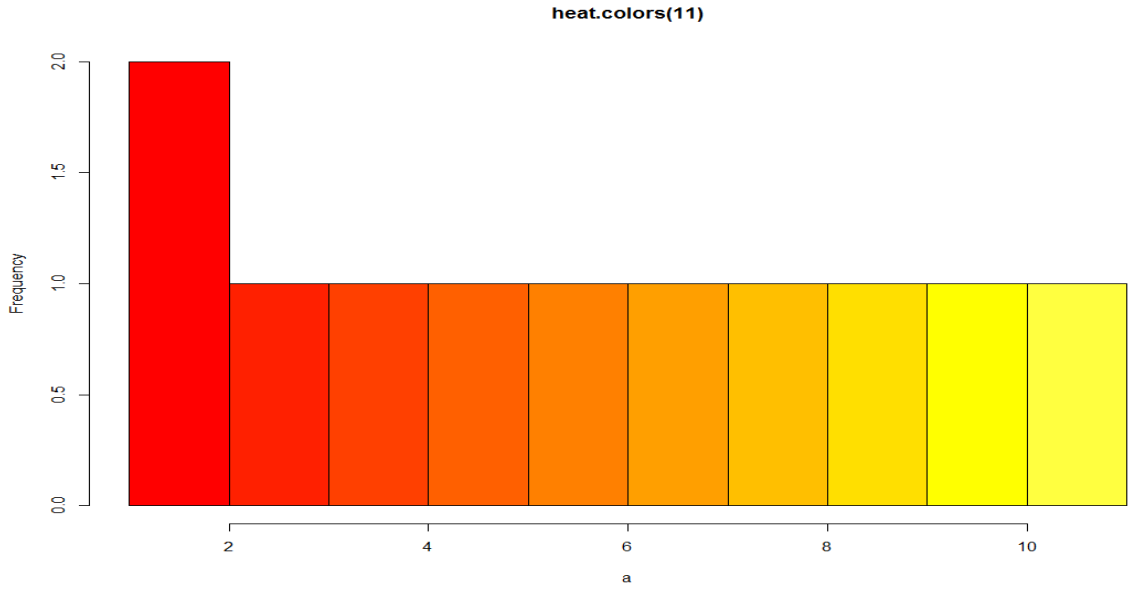
```
> summary(calif)
     value            income            age             rooms          bedrooms        population      households
 Min.   : 14999   Min.   : 0.4999   Min.   : 1.00   Min.   :    2   Min.   :   1.0   Min.   :    3   Min.   :   1.0
 1st Qu.:119600   1st Qu.: 2.5634   1st Qu.:18.00   1st Qu.: 1448   1st Qu.: 295.0   1st Qu.:  787   1st Qu.: 280.0
 Median :179700   Median : 3.5348   Median :29.00   Median : 2127   Median : 435.0   Median : 1166   Median : 409.0
 Mean   :206856   Mean   : 3.8707   Mean   :28.64   Mean   : 2636   Mean   : 537.9   Mean   : 1425   Mean   : 499.5
 3rd Qu.:264725   3rd Qu.: 4.7432   3rd Qu.:37.00   3rd Qu.: 3148   3rd Qu.: 647.0   3rd Qu.: 1725   3rd Qu.: 605.0
 Max.   :500001   Max.   :15.0001   Max.   :52.00   Max.   :39320   Max.   :6445.0   Max.   :35682   Max.   :6082.0
```

중위수에 비해 MAX값들이 너무 크다.



Boxplot 으로 확인해본 결과 이상치가 굉장히 많은 것을 볼 수 있다.

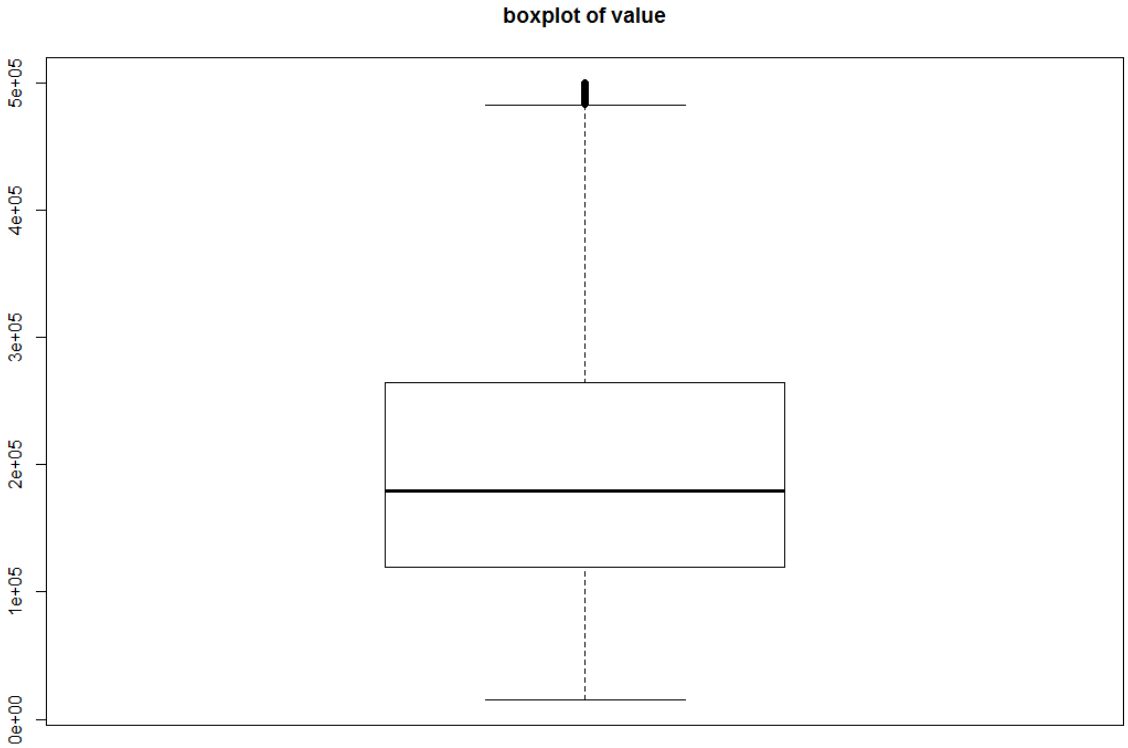
B.주택가격의 중위수를 캘리포니아 지도 위에 표현하시오. 생성된 그림을 간단히 셜멍하시오



중위수 fitting결과 파란색 점들이 Latitude=34, longitude=-119 부근에 몰려있는 것을 확인할 수 있다.

아래의 hist는 heat.color에서 제공되는 색상 중 그래프에 사용된 11개를 뽑은 건데 빨갈수록 floor(calif$value/50000) 의 값이 11에 가까워 상대적으로 value값이 크다고 해석할 수 있고 반대

로 노랄수록 value값이 작다고 해석할 수 있다. 그래프에서 7번~9번 정도에 해당하는 색깔에 중위수 점이 찍힌 것을 보아 value가 비교적 낮은 값들에 몰려있다고 해석할 수 있고 이는 boxplot의 해석 결과와 동일하다.

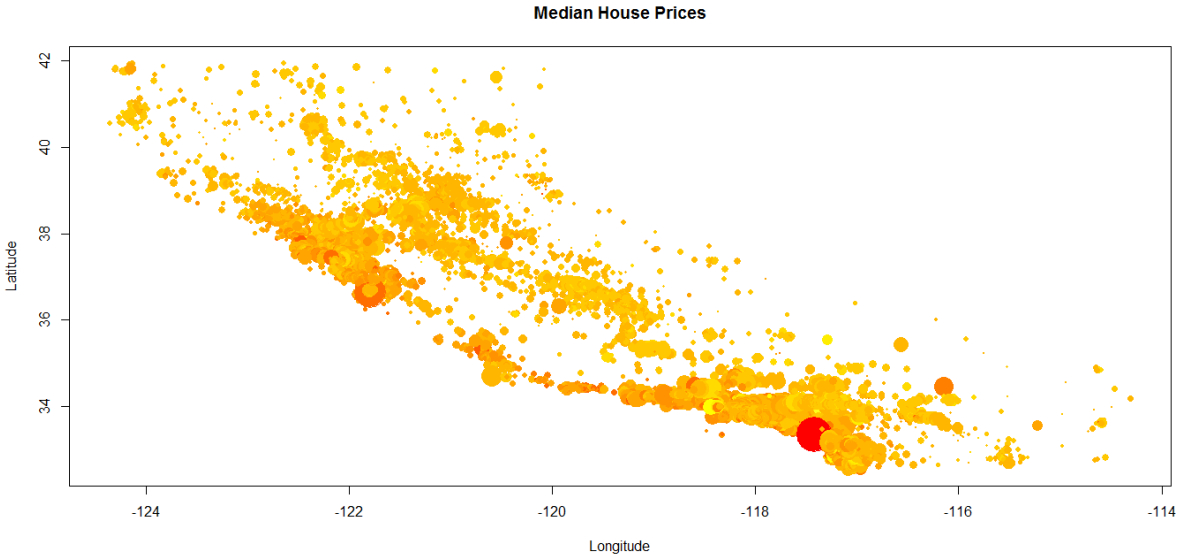


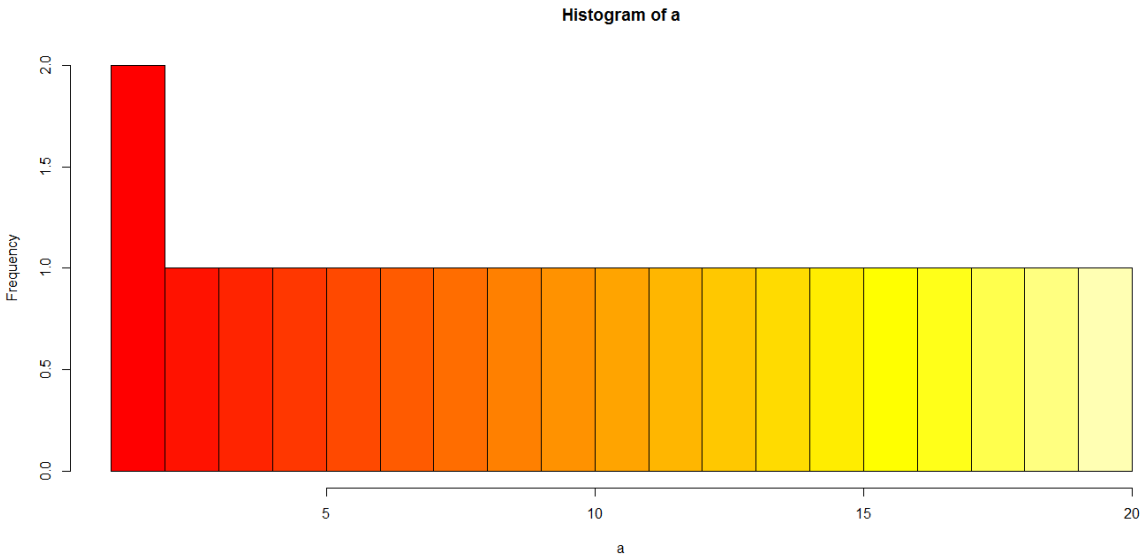
C.위도, 경도를 제외한 모든 변수를 이용하여 주택가격의 분위수를 설명하는 선형회귀모형을 적합하고, 그 결과를 설명하시오.

```
> lm(value~., data=calif[,1:7])

Call:
lm(formula = value ~ ., data = calif[, 1:7])

Coefficients:
(Intercept)       income          age      .  rooms      bedrooms   population   households
  -45951.59     47697.74      1880.94        -19.64       100.43       -35.50       125.46


> summary(lm(value~., data=calif[,1:7]))

Call:
lm(formula = value ~ ., data = calif[, 1:7])

Residuals:
    Min      1Q  Median      3Q     Max
-636230  -48172  -11755   34485  709402

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.595e+04  2.241e+03  -20.51   <2e-16 ***
income       4.770e+04  3.422e+02  139.37   <2e-16 ***
age          1.881e+03  4.532e+01   41.51   <2e-16 ***
rooms       -1.964e+01  8.332e-01  -23.57   <2e-16 ***
bedrooms     1.004e+02  7.489e+00   13.41   <2e-16 ***
population  -3.550e+01  1.165e+00  -30.47   <2e-16 ***
households   1.255e+02  8.055e+00   15.57   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75970 on 20633 degrees of freedom
Multiple R-squared:  0.5667,    Adjusted R-squared:  0.5666
F-statistic:  4497 on 6 and 20633 DF,  p-value: < 2.2e-16
```

전체적인 P – value 값이 굉장히 작은 것을 보아 회귀분석의 결과가 유의함을 알 수 있고 사용된 변수들 모두가 유의미한 변수임을 확인 할 수 있다. Adjusted R-square 값을 통해 이 회귀식은 56% 정도의 설명력 지님을 확인할 수 있다.

D. 위에서 적합된 결과에서 구한 잔차를 지도위에 표시하는 그림을 그리고 그 결과를 설명하시오





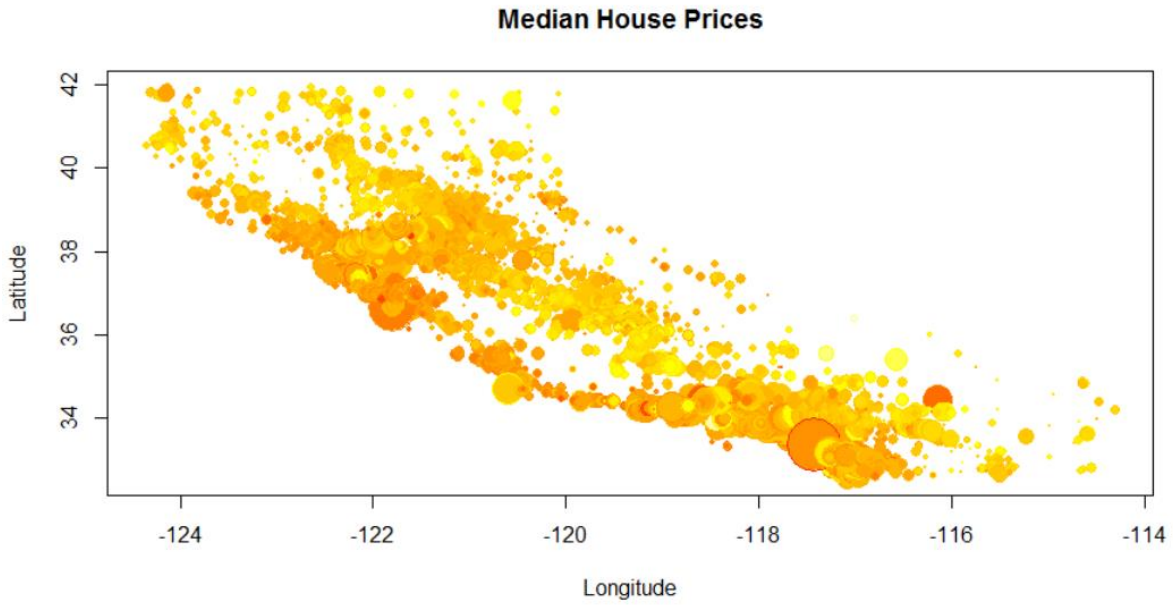
아래의 hist는 heat.color에서 제공되는 색상 중 그래프에 사용된 20개를 뽑은 건데 빨갈수록 위에 B번과 동일한 이유로 상대적으로 잔차값이 크다고 해석할 수 있고 반대로 노랄수록 잔차값이 작다고 해석할 수 있다. 10의 주변인 주황색이 가장 많이 눈에 띄고 노랑과 빨강이 많지 않게 고른 분포를 띄는 것을 보아 잔차가 0주변에 잘 분포되어 있고, C의 회귀분석 결과와 동일하게 회귀식이 잘 적합 되었다고 해석할 수 있다.

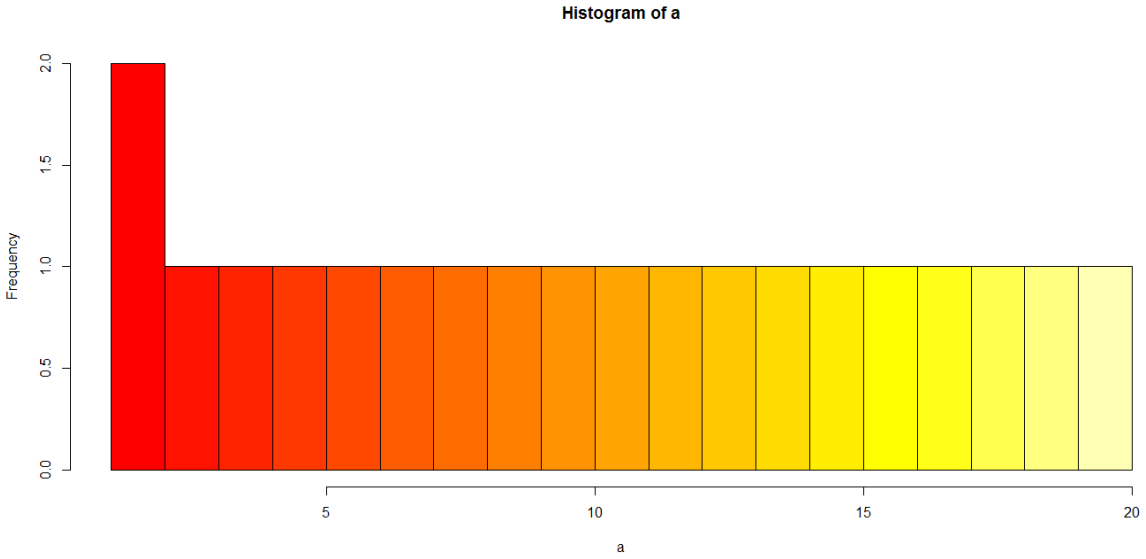
E. 반응변수를 로그 변환한 자료를 C와 D를 반복하고 그 결과를 비교하시오

```
> rg_log_calif

Call:
lm(formula = log_value ~ ., data = log_calif)

Coefficients:
(Intercept)       income          age        rooms     bedrooms   population   households
  2.392e+00    1.881e-02    6.478e-04   -9.143e-06    3.378e-05   -1.212e-05    6.180e-05
```

```
> summary(rg_log_calif)

Call:
lm(formula = log_value ~ ., data = log_calif)

Residuals:
      Min        1Q    Median        3Q       Max
-0.248942 -0.018875  0.002289  0.020820  0.224452

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.392e+00  9.798e-04 2441.23   <2e-16 ***
income       1.881e-02  1.497e-04  125.67   <2e-16 ***
age          6.478e-04  1.982e-05   32.69   <2e-16 ***
rooms       -9.143e-06  3.644e-07  -25.09   <2e-16 ***
bedrooms     3.378e-05  3.275e-06   10.31   <2e-16 ***
population  -1.212e-05  5.095e-07  -23.80   <2e-16 ***
households   6.180e-05  3.522e-06   17.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03322 on 20633 degrees of freedom
Multiple R-squared:  0.5089,    Adjusted R-squared:  0.5088
F-statistic:  3563 on 6 and 20633 DF,  p-value: < 2.2e-16
```

Adjusted R square 값이 0.50 으로 log를 취하지 않았을 때보다 값이 더 작아진 것을 볼 수 있다. Log를 취한 것이 오히려 설명력이 좋지 않다는 것을 알 수 있다.





아래의 hist는 heat.color에서 제공되는 색상 중 그래프에 사용된 20개를 뽑은 건데 빨갈수록 위에 B번과 동일한 이유로 상대적으로 잔차값이 크다고 해석할 수 있고 반대로 노랄수록 잔차값이 작다고 해석할 수 있다. 위에 D의 그림과 비교했을 때 조금 흩어진 것을 볼 수 있으나 log를 취함으로써 극단치 (빨간색)가 사라진 것을 확인 할 수 있다