review_reg.R

SANGHOOJEFFREY

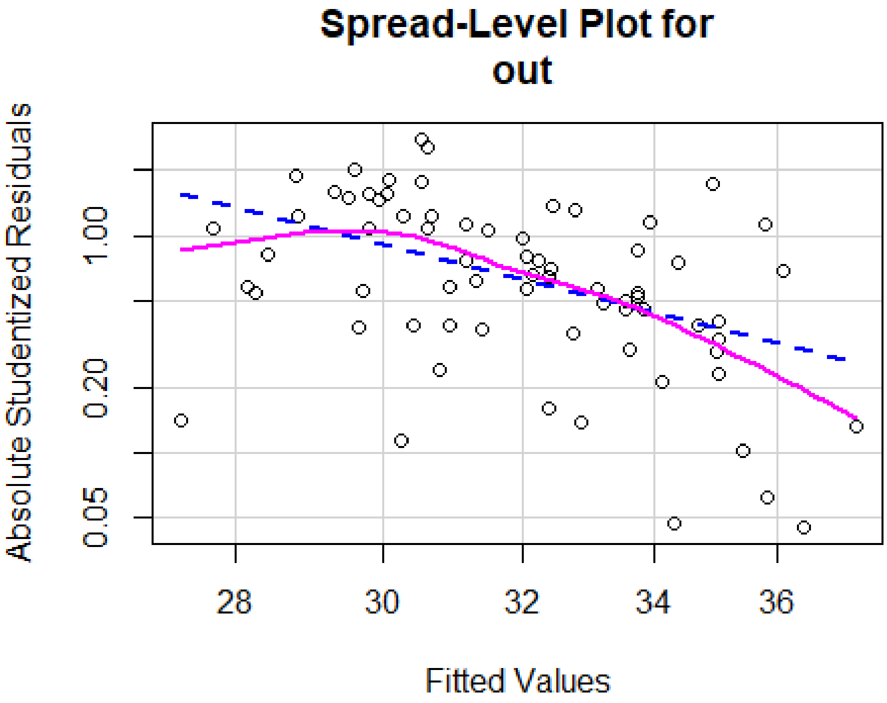
Wed Jun 27 23:13:13 2018

```
# 회귀분석을 위해 필요한 패키지 설치 또는 불러오기
if(!require(car)) install.packages("car",  repos = "http://cran.us.r-project.
org"); library(car)

## Loading required package: car

## Warning: package 'car' was built under R version 3.4.4

## Loading required package: carData

## Warning: package 'carData' was built under R version 3.4.4

if(!require(lmtest)) install.packages("lmtest", repos = "http://cran.us.r-pro
ject.org"); library(lmtest)

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 3.4.4

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.4.4

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

# 분석을 위한 자료 불러오기

load(file='data/data2.rda')
head(data2)

##   id     light        prec       rh     s_30     s_60     s_90   s_temp
## 1  1 0.3953990 0.000000000 74.87500 62.24833 65.37250 56.67375 20.07417
## 2  2 0.4716573 0.000000000 80.38417 59.39917 60.05708 56.08708 18.80713
## 3  4 0.3397830 0.002433239 80.80167 58.39167 62.56333 57.47583 22.90226
## 4  5 0.4375250 0.000000000 71.94375 54.62958 61.21750 59.27583 20.60042
## 5  1 0.4288470 0.000000000 83.62865 64.94625 70.97458 51.79833 19.93833
## 6  2 0.4387887 0.046666667 89.72399 62.46333 57.33125 64.06708 18.71500
##    s_trans     temp        ws    l
## 1  0.00125 24.19333 1.3340909 31.7
## 2  0.00000 23.74822 1.8795652 27.6
## 3 23.75000 23.24781 1.1700000 30.6
## 4  0.00000 24.53833 1.7139130 24.0
## 5  0.00625 18.44750 0.6579167 33.4
## 6  0.00000 19.25958 1.6591667 31.5
```

```
# 위 자료는 엽장(l)과 관측된 기상요소 간 관계성을 규명하고자 수집된 자료입니다.
# 다중회귀모형을 통해 엽장(l)을 위한 최적회귀모형을 찾으세요

# Q1. 산점도를 통한 설명변수와 반응변수 간 관계 파악

# Q2. lm()함수를 이용한 회귀모형 세우기
out=lm(l~.-id,data=data2)
summary(out) # 얻어진 결과를 해석해보세요
```

```
## 
## Call:
## lm(formula = l ~ . - id, data = data2)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -9.720 -2.180  1.049  2.171  7.093 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 41.51504    9.06888   4.578 2.47e-05 ***
## light       -2.01961    5.01354  -0.403   0.6885    
## prec         1.97701    3.75025   0.527   0.6001    
## rh           0.01977    0.07401   0.267   0.7902    
## s_30         0.04690    0.03066   1.529   0.1315    
## s_60        -0.04912    0.02634  -1.865   0.0671 .  
## s_90         0.03698    0.02720   1.359   0.1792    
## s_temp      -0.13621    0.38781  -0.351   0.7267    
## s_trans     -0.01297    0.17610  -0.074   0.9416    
## temp        -0.46174    0.27273  -1.693   0.0957 .  
## ws           0.50876    0.56152   0.906   0.3686    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.97 on 59 degrees of freedom
## Multiple R-squared:  0.2917, Adjusted R-squared:  0.1717 
## F-statistic:  2.43 on 10 and 59 DF,  p-value: 0.01689
```

```
# Q3. vif()함수를 이용한 다중 공선성 확인

# Q4. 정규성 검정 (정규성을 만족하지 않음)

# Q5. 잔차도표와 DurbinWatson 을 통한 독립성 검정(독립성을 만족하지 않음)

# Q6. 등분산성 검토 (등분산성을 만족하지 않음)

power=spreadLevelPlot(out)$PowerTransformation # 적합한 power
```



```
# 반응변수 변환
out2=lm(l^power~.-id,data=data2)
summary(out2)
```

```
##
## Call:
## lm(formula = l^power ~ . - id, data = data2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.953e+10 -8.601e+09  9.618e+08  6.251e+09  2.191e+10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.561e+10  2.461e+10   3.072  0.00321 **
## light       -3.552e+09  1.360e+10  -0.261  0.79492
## prec         1.253e+09  1.018e+10   0.123  0.90245
## rh          -6.859e+07  2.008e+08  -0.342  0.73392
## s_30         1.081e+08  8.321e+07   1.299  0.19886
## s_60        -1.138e+08  7.146e+07  -1.592  0.11674
## s_90         7.612e+07  7.381e+07   1.031  0.30660
## s_temp      -6.539e+08  1.052e+09  -0.621  0.53678
## s_trans     -1.261e+08  4.779e+08  -0.264  0.79278
## temp        -1.885e+09  7.401e+08  -2.547  0.01349 *
## ws           9.915e+08  1.524e+09   0.651  0.51778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.077e+10 on 59 degrees of freedom
## Multiple R-squared:  0.3513, Adjusted R-squared:  0.2414
## F-statistic: 3.196 on 10 and 59 DF,  p-value: 0.00244
```

```
shapiro.test(out2$residuals) #정규성 만족
```

```
##
##  Shapiro-Wilk normality test
##
## data:  out2$residuals
## W = 0.96828, p-value = 0.07279
```

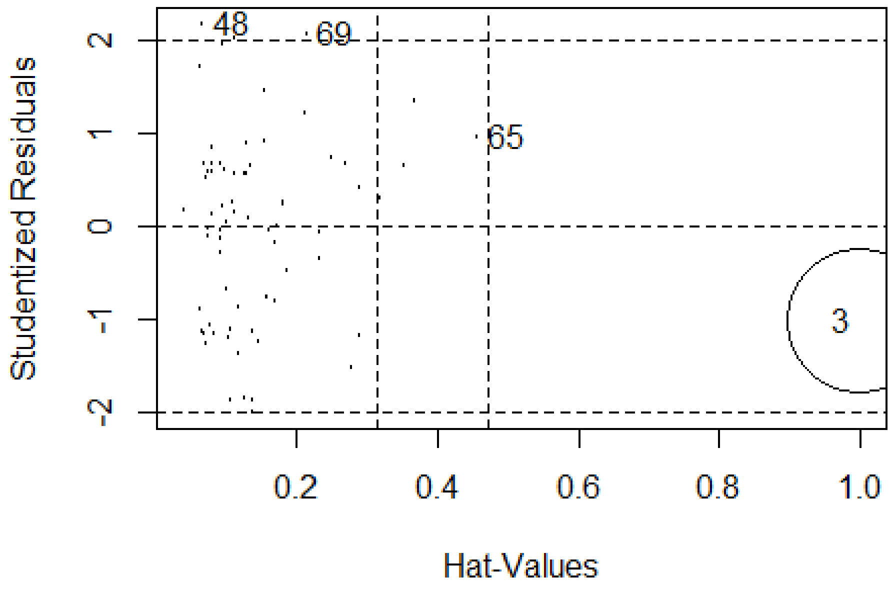
```
durbinWatsonTest(out2) # 독립성은 만족되지 않음
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5242338     0.9507617       0
##  Alternative hypothesis: rho != 0
```

```
ncvTest(out2) #등분산성 만족
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.956769    Df = 1     p = 0.1618593
```

```
influencePlot(out2)
```



```
##       StudRes        Hat        CookD
## 3  -1.0207943 0.99987199 739.40050763
## 48  2.1716285 0.06764959   0.02926430
## 65  0.9562592 0.45754418   0.07021948
## 69  2.0687561 0.21542444   0.10120250
```

```
outlierTest(out2)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## 48 2.171628           0.033987           NA
```

```
# 3 영향점, 48 이상점, 69 번째는 영향점이면서 이상점으로 의심됨

k=10
#hat(값의 임계값)
2*(k+1)/nrow(data2)
```

```
## [1] 0.3142857
```

```
#cookD 임계값
4/(nrow(data2)-k-1)
```

```
## [1] 0.06779661
```

```r
# 자료를 제거하고 다시 회귀모형 세우기
data3=data2[-c(3,48, 69),]

rownames(data3)<-1:nrow(data3)
out3=lm(l^power~.-id,data3)
summary(out3)
```

```
##
## Call:
## lm(formula = l^power ~ . - id, data = data3)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.069e+10 -7.993e+09  1.565e+09  6.932e+09  1.752e+10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.752e+10  2.394e+10   1.985   0.0521 .
## light       -6.691e+09  1.263e+10  -0.530   0.5983
## prec         3.824e+09  9.331e+09   0.410   0.6835
## rh          -1.176e+07  1.970e+08  -0.060   0.9526
## s_30         1.966e+08  8.495e+07   2.314   0.0243 *
## s_60        -9.089e+07  6.639e+07  -1.369   0.1765
## s_90         1.480e+08  7.270e+07   2.036   0.0465 *
## s_temp      -7.332e+08  9.921e+08  -0.739   0.4630
## s_trans      7.613e+10  3.869e+10   1.967   0.0541 .
## temp        -1.142e+09  7.340e+08  -1.556   0.1253
## ws           8.875e+08  1.403e+09   0.633   0.5295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.85e+09 on 56 degrees of freedom
## Multiple R-squared:  0.4204, Adjusted R-squared:  0.3169
## F-statistic: 4.062 on 10 and 56 DF,  p-value: 0.0003193
```

```r
shapiro.test(out3$residuals)
```

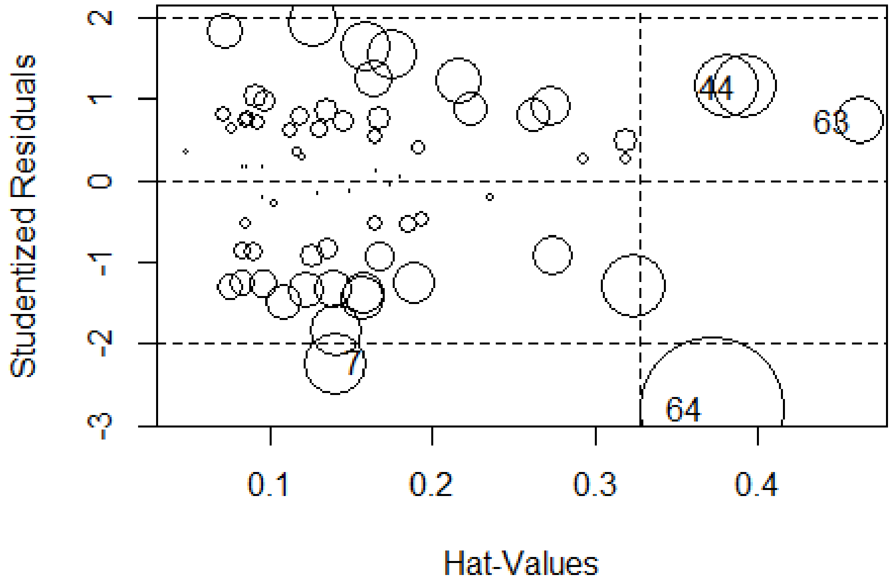
```
##
##  Shapiro-Wilk normality test
##
## data:  out3$residuals
## W = 0.96454, p-value = 0.05302
```

```r
durbinWatsonTest(out3)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.4124547      1.162901   0.002
##  Alternative hypothesis: rho != 0
```

```r
ncvTest(out3)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.02417289    Df = 1     p = 0.8764458

influencePlot(out3)
```



```
##        StudRes       Hat      CookD
## 7  -2.2336454 0.1404548 0.06918609
## 44  1.1457497 0.3924529 0.07666123
## 63  0.7223982 0.4633754 0.04131874
## 64 -2.8103449 0.3723649 0.37926165

# 분석결과 64 번째 자료가 영향점이면서 이상점으로 보임


# 64 번째를 다시 제거하고 회귀모형 세우
data4=data3[-c(64),]
rownames(data4)<-1:nrow(data4)
out4=lm(l^power~.-id,data4)
summary(out4)

##
## Call:
## lm(formula = l^power ~ . - id, data = data4)
##
## Residuals:
```

```
##         Min         1Q     Median         3Q        Max
## -1.847e+10 -7.064e+09  1.286e+09  6.206e+09  1.577e+10
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.944e+10  2.260e+10   2.188  0.03297 *
## light       -6.757e+09  1.191e+10  -0.567  0.57294
## prec         1.364e+10  9.472e+09   1.440  0.15552
## rh          -2.630e+07  1.860e+08  -0.141  0.88806
## s_30         2.397e+08  8.161e+07   2.938  0.00482 **
## s_60        -1.112e+08  6.306e+07  -1.763  0.08349 .
## s_90         1.800e+08  6.953e+07   2.588  0.01232 *
## s_temp      -1.333e+09  9.601e+08  -1.389  0.17052
## s_trans      1.156e+11  3.912e+10   2.955  0.00460 **
## temp        -7.752e+08  7.048e+08  -1.100  0.27617
## ws           5.409e+08  1.329e+09   0.407  0.68565
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.294e+09 on 55 degrees of freedom
## Multiple R-squared:  0.4921, Adjusted R-squared:  0.3998
## F-statistic: 5.329 on 10 and 55 DF,  p-value: 1.858e-05
```

```
shapiro.test(out4$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  out4$residuals
## W = 0.96552, p-value = 0.06334
```

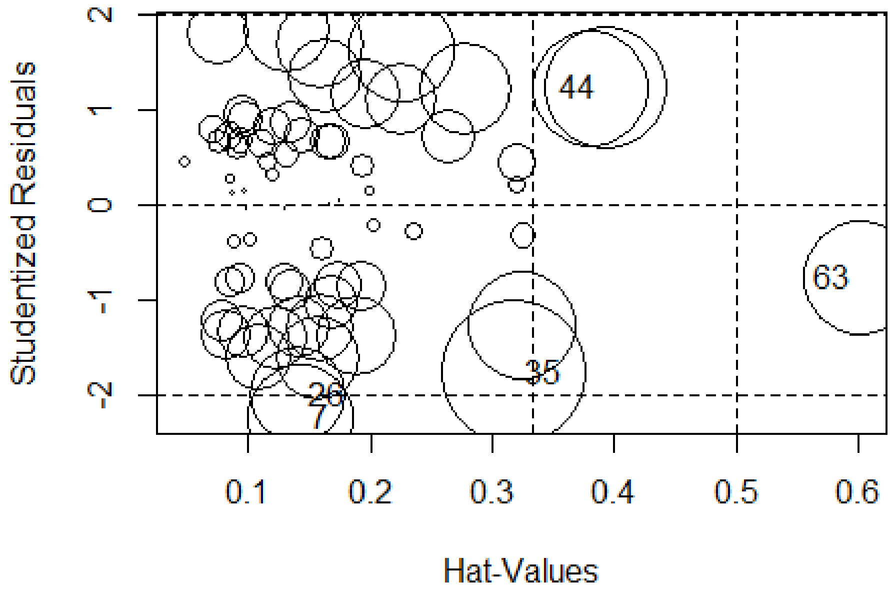
```
durbinWatsonTest(out4)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3968569      1.187187       0
##  Alternative hypothesis: rho != 0
```

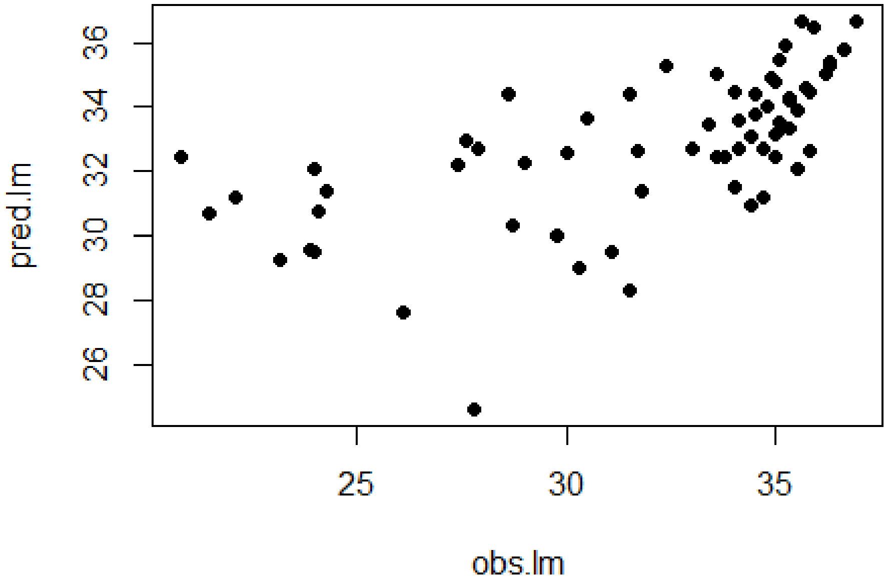
```
ncvTest(out4)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.5683629    Df = 1     p = 0.4509102
```

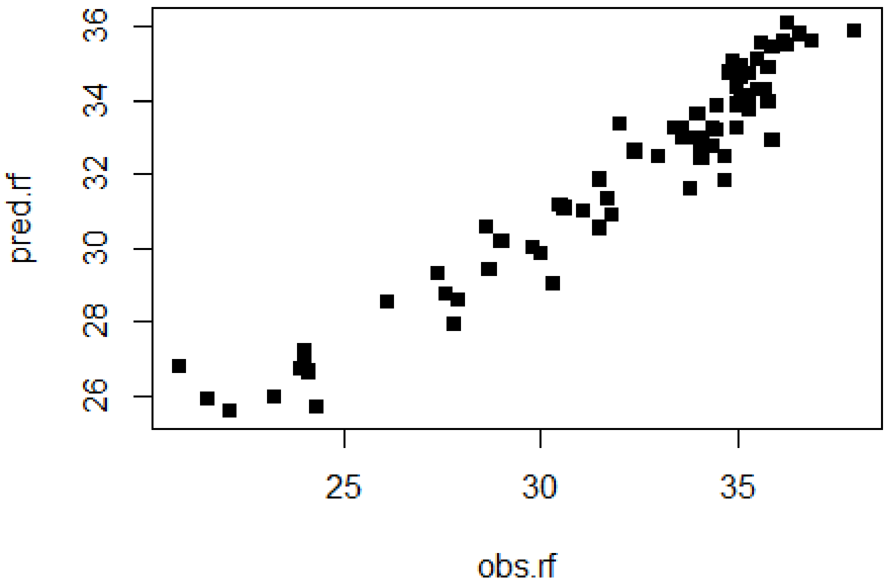
```
influencePlot(out4)
```



```
##        StudRes       Hat      CookD
## 7  -2.2216984 0.1427033 0.06970481
## 26 -1.9810046 0.1407694 0.05549810
## 35 -1.7619901 0.3178324 0.12665226
## 44  1.2421570 0.3925015 0.08974083
## 63 -0.7552603 0.6009124 0.07869534

# Q7. 변수선택 했을 때 결과는??




# 관측값과 예측값 간 그림 그려보기
obs.lm <- data4$l
pred.lm <- out4$fitted.values^(1/power)

(lm.rmse<-sqrt(sum((obs.lm-pred.lm)^2,na.rm=T)/length(out$residuals)))

## [1] 3.472486

plot(obs.lm, pred.lm, pch=16)
```



```
# randomforest 모형과 비교해보기
if(!require(randomForest)) install.packages("randomForest", repos = "http://c
ran.us.r-project.org");library(randomForest)

## Loading required package: randomForest

## Warning: package 'randomForest' was built under R version 3.4.4

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

colnames(data2)

##  [1] "id"      "light"   "prec"    "rh"      "s_30"    "s_60"    "s_90"
##  [8] "s_temp"  "s_trans" "temp"    "ws"      "l"

rf_l=randomForest(l~.-id,data2)
pred.rf<-predict(rf_l,newdata=data2)
obs.rf <- data2$l
(rf.rmse<-sqrt(sum((obs.rf-pred.rf)^2,na.rm=T)/length(obs.rf)))

## [1] 1.687784

plot(obs.rf, pred.rf, pch=15)
```


pred.rf
obs.rf
26
28
30
32
34
36
25
30
35