# 산업공학특론I_10주차_주성분회귀&부분최소제곱회귀_실습

**Munwon Lim**

**5/8/2024**

# [데이터 분석]

(https://www.kaggle.com/datasets/sohommajumder21/appliances-energy-prediction-data-set
(https://www.kaggle.com/datasets/sohommajumder21/appliances-energy-prediction-data-set))

저에너지 건물의 가전제품 에너지 사용량에 관한 회귀 모델을 만들기 위한 실험 데이터

- date: 날짜 (시간 연-월-일 시:분:초)
- Appliances: 총 에너지 사용량 (Wh)
- lights: 집안의 조명 장치의 에너지 사용량 (Wh)
- T1: 주방 온도 (섭씨)
- RH_1: 주방 습도 (%)
- T2: 거실 온도 (섭씨)
- RH_2: 거실 습도 (%)
- T3: 세탁실 온도 (섭씨)
- RH_3: 세탁실 습도 (%)
- T4: 사무실 온도 (섭씨)
- RH_4: 사무실 습도 (%)
- T5, 화장실 온도 (섭씨)
- RH_5, 화장실 습도 (%)
- T6: 건물 외부 온도 - 북쪽 (섭씨)
- RH_6, 건물 외부 습도 - 북쪽 (%)
- T7: 다리미실 온도 (섭씨)
- RH_7, 다리미실 습도 (%)
- T8: 십대방 온도 (섭씨)
- RH_8: 십대방 습도 (%)
- T9: 부모방 온도 (섭씨)
- RH_9: 부모방 습도 (%)
- T_out: 외부 온도 (섭씨)
- Press_mm_hg: 기압 (mm Hg)
- RH_out: 외부 습도 (%)
- Windspeed: 풍속 (m/s)
- Visibility:가시도 (km)
- Tdewpoint: 이슬점 (°C)

# 1. 데이터 탐색 (EDA) 및 전처리

```
# 데이터 로드 및 요약
dat <- read.csv('산업공학특론I_10주차_실습 데이터.csv')
head(dat)
```

```
##                date Appliances lights    T1     RH_1   T2     RH_2    T3
## 1 11-01-2016 17:00         60     30 19.89 47.59667 19.2 44.79000 19.79
## 2 11-01-2016 17:10         60     30 19.89 46.69333 19.2 44.72250 19.79
## 3 11-01-2016 17:20         50     30 19.89 46.30000 19.2 44.62667 19.79
## 4 11-01-2016 17:30         50     40 19.89 46.06667 19.2 44.59000 19.79
## 5 11-01-2016 17:40         60     40 19.89 46.33333 19.2 44.53000 19.79
## 6 11-01-2016 17:50         50     40 19.89 46.02667 19.2 44.50000 19.79
##       RH_3       T4     RH_4       T5  RH_5       T6     RH_6       T7
## 1 44.73000 19.00000 45.56667 17.16667 55.20 7.026667 84.25667 17.20000
## 2 44.79000 19.00000 45.99250 17.16667 55.20 6.833333 84.06333 17.20000
## 3 44.93333 18.92667 45.89000 17.16667 55.09 6.560000 83.15667 17.20000
## 4 45.00000 18.89000 45.72333 17.16667 55.09 6.433333 83.42333 17.13333
## 5 45.00000 18.89000 45.53000 17.20000 55.09 6.366667 84.89333 17.20000
## 6 44.93333 18.89000 45.73000 17.13333 55.03 6.300000 85.76667 17.13333
##       RH_7   T8     RH_8       T9  RH_9 T_out Press_mm_hg RH_out Windspeed
## 1 41.62667 18.2 48.90000 17.03333 45.53  6.60       733.5     92  7.000000
## 2 41.56000 18.2 48.86333 17.06667 45.56  6.48       733.6     92  6.666667
## 3 41.43333 18.2 48.73000 17.00000 45.50  6.37       733.7     92  6.333333
## 4 41.29000 18.1 48.59000 17.00000 45.40  6.25       733.8     92  6.000000
## 5 41.23000 18.1 48.59000 17.00000 45.40  6.13       733.9     92  5.666667
## 6 41.26000 18.1 48.59000 17.00000 45.29  6.02       734.0     92  5.333333
##   Visibility Tdewpoint
## 1   63.00000       5.3
## 2   59.16667       5.2
## 3   55.33333       5.1
## 4   51.50000       5.0
## 5   47.66667       4.9
## 6   43.83333       4.8
```

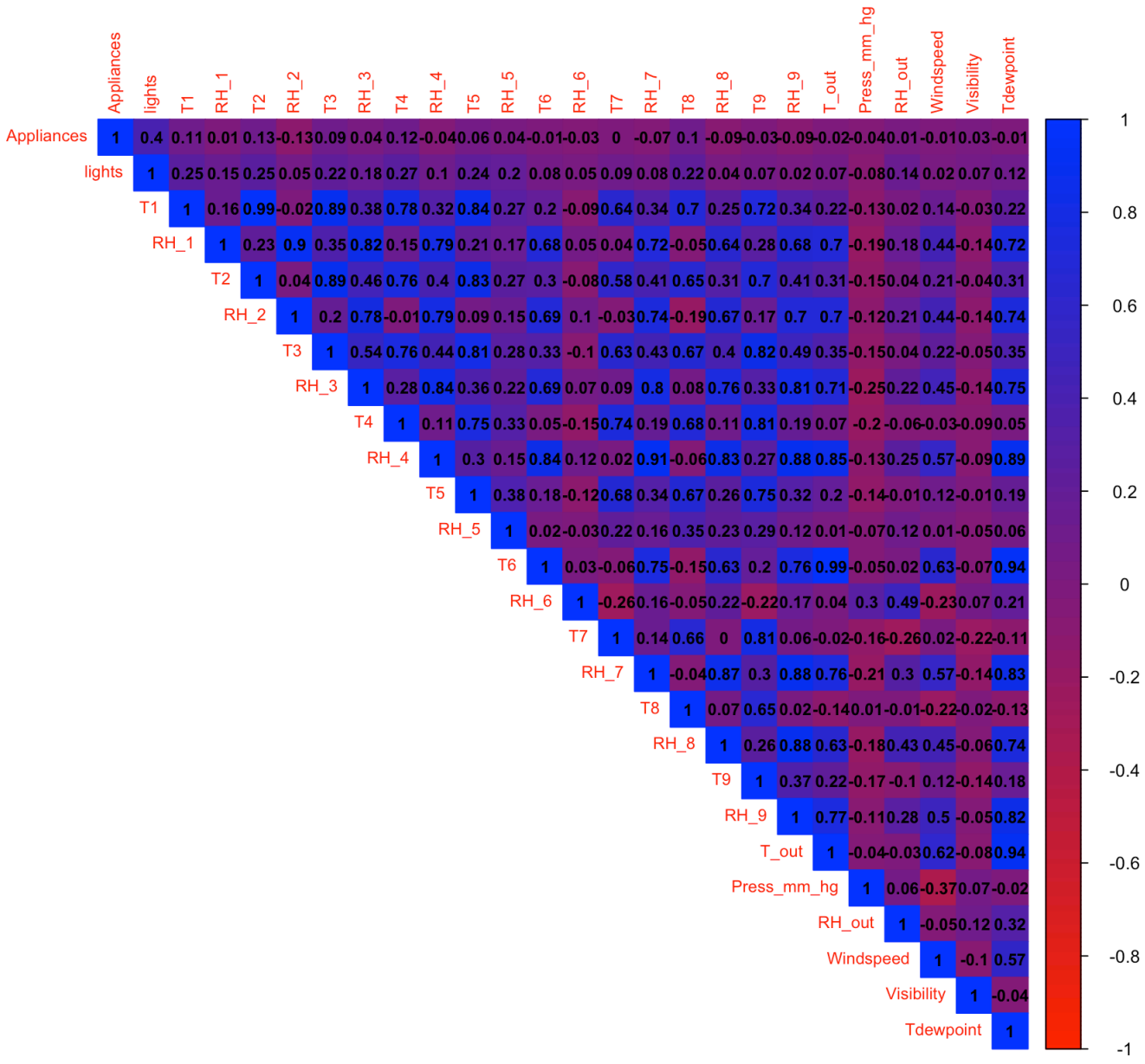
```
# 데이터 전처리
dat <- dat[,-1] #불필요한 변수 제거
dat <- scale(dat)#단위 통일
dat <- as.data.frame(dat)

# 학습, 테스트셋 분할
set.seed(0)
trainidx <- sample(1:nrow(dat), 0.7*nrow(dat))
train <- dat[trainidx,]
test <- dat[-trainidx,]
```

# 2. 상관분석

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
# 상관계수 테이블 생성
corr <- cor(train, method='pearson')

# 상관계수 테이블 시각화
col <- colorRampPalette(c('red','blue'))
corrplot(corr, method='color', col=col(200), addCoef.col = 'black', type='upper',
         number.cex=0.75, tl.cex=0.75)
```




## 3. 일반회귀모형 적합

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: Can't find generic `sew` in package knitr to register S3 method.
## This message is only shown to developers using devtools.
## Do you need to update knitr to the latest version?
```

```
library(Metrics)
```

```
##
## Attaching package: 'Metrics'
```

```
## The following objects are masked from 'package:caret':
##
##     precision, recall
```

```
# 모델 수립
reg <- train(Appliances~ ., data=train, method = 'lm', trControl = trainControl(metho
d = 'cv'))

# 모델 수립 결과
summary(reg)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2512 -0.3903 -0.1298  0.0955  6.8266
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.004063   0.017142   0.237 0.812660
## lights       0.333347   0.019976  16.687  < 2e-16 ***
## T1          -0.581495   0.153665  -3.784 0.000158 ***
## RH_1         0.410062   0.049162   8.341  < 2e-16 ***
## T2           0.285206   0.157681   1.809 0.070605 .
## RH_2        -0.750580   0.063081 -11.899  < 2e-16 ***
## T3           0.187660   0.059190   3.170 0.001540 **
## RH_3         0.225175   0.053621   4.199 2.77e-05 ***
## T4           0.084617   0.044019   1.922 0.054677 .
## RH_4         0.126180   0.072046   1.751 0.079999 .
## T5          -0.042518   0.037876  -1.123 0.261738
## RH_5         0.025272   0.020417   1.238 0.215897
## T6           0.118888   0.121232   0.981 0.326846
## RH_6        -0.022047   0.023321  -0.945 0.344556
## T7           0.143978   0.053094   2.712 0.006738 **
## RH_7        -0.007378   0.075132  -0.098 0.921777
## T8           0.027382   0.039643   0.691 0.489805
## RH_8        -0.162755   0.056862  -2.862 0.004239 **
## T9          -0.219223   0.058415  -3.753 0.000179 ***
## RH_9        -0.023261   0.059993  -0.388 0.698244
## T_out       -1.029762   0.587682  -1.752 0.079850 .
## Press_mm_hg  0.023117   0.022561   1.025 0.305623
## RH_out      -0.295052   0.211987  -1.392 0.164089
## Windspeed    0.020835   0.029272   0.712 0.476671
## Visibility   0.007103   0.019205   0.370 0.711513
## Tdewpoint    1.042612   0.617959   1.687 0.091688 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8763 on 2595 degrees of freedom
## Multiple R-squared:  0.2612, Adjusted R-squared:  0.2541
## F-statistic: 36.69 on 25 and 2595 DF,  p-value: < 2.2e-16
```

```
# 예측력 평가
err <- function(actual,pred){
  result <- c(mae(actual,pred), mse(actual,pred), rmse(actual,pred))
  names(result) <- c('MAE','MSE','RMSE')
  print(result)
}

pred_reg <- predict(reg, test)
err(test$Appliances, pred_reg)
```

```
##       MAE       MSE      RMSE
## 0.4902237 0.7064951 0.8405326
```

# 4. PCR

```
library(pls)
```

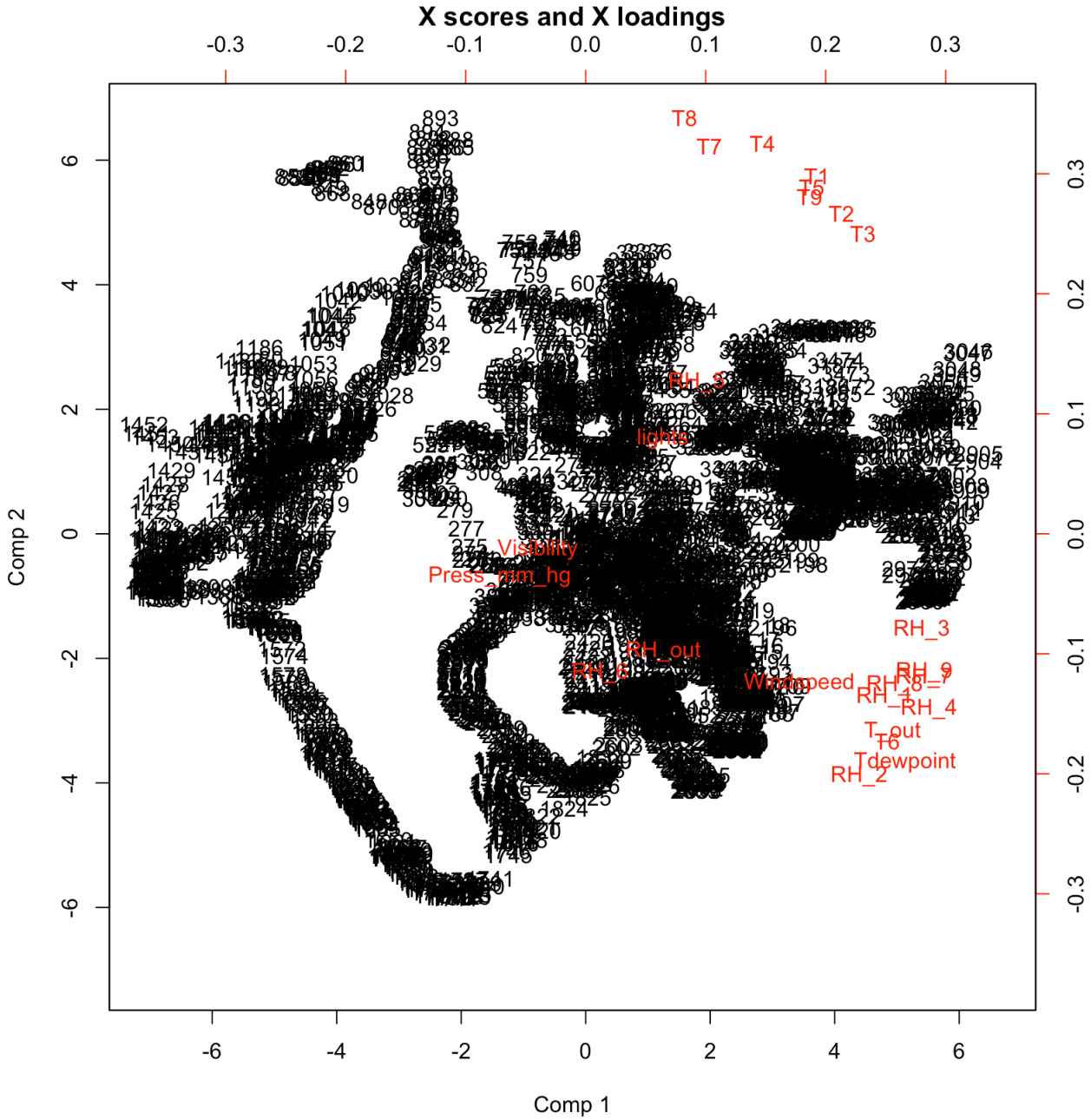
```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:caret':
##
##     R2
```

```
## The following object is masked from 'package:corrplot':
##
##     corrplot
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

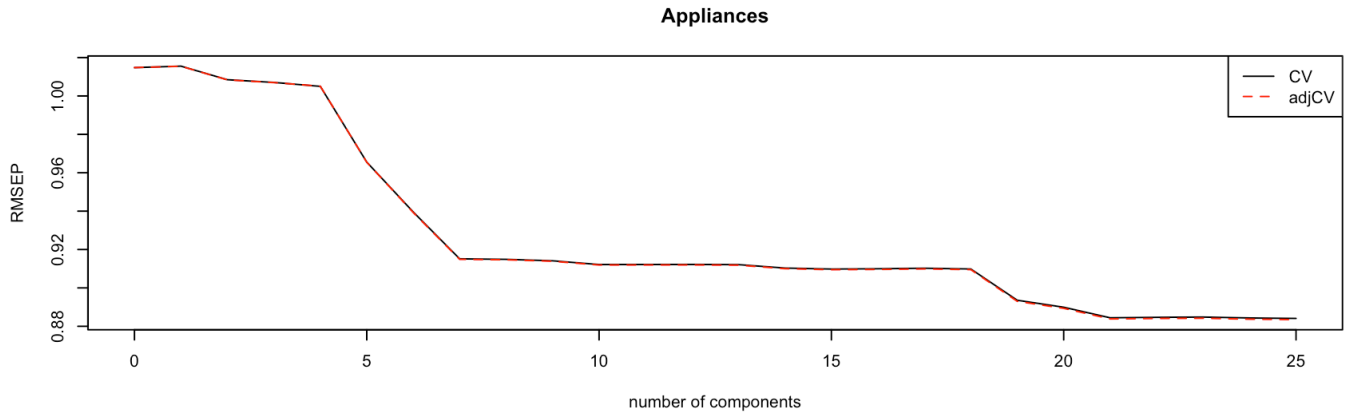
```
# 모델 수립
pcr_model <- pcr(Appliances ~ ., data = train, validation = 'CV')

# 주성분 검토
par(mfrow=c(1,1))
biplot(pcr_model)
```

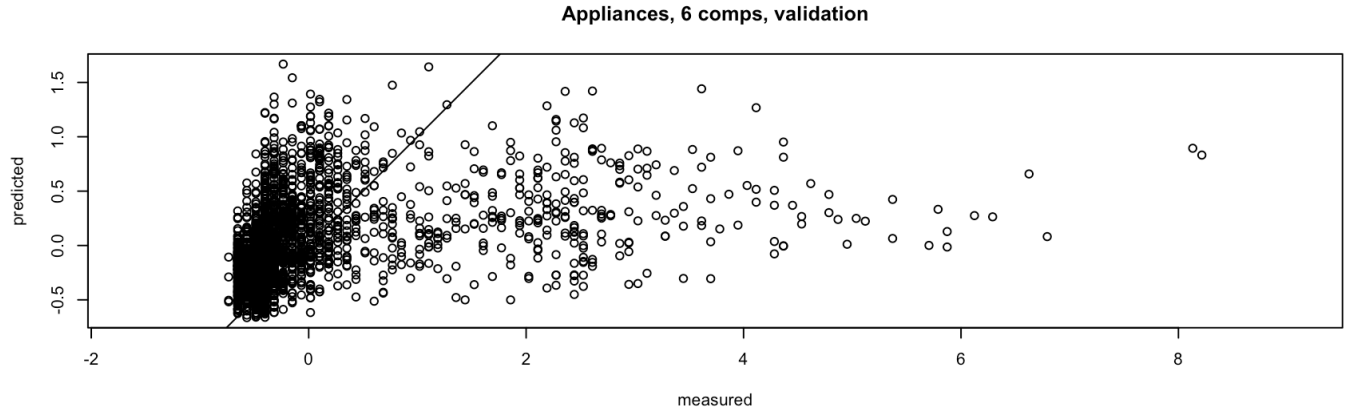


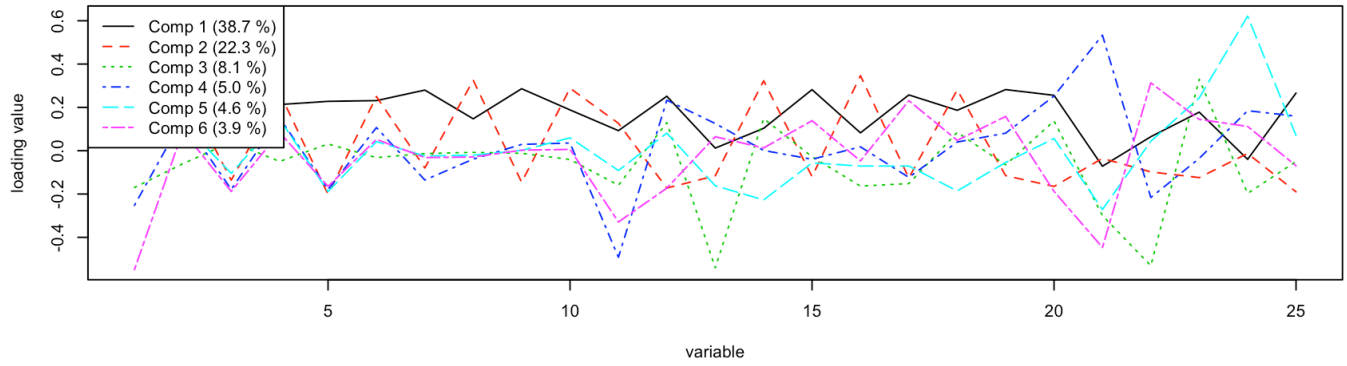

```
summary(pcr_model)
```

```
## Data:    X dimension: 2621 25
##  Y dimension: 2621 1
## Fit method: svdpc
## Number of components considered: 25
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.015    1.016    1.008    1.007    1.005   0.9655   0.9396
## adjCV        1.015    1.015    1.008    1.007    1.005   0.9655   0.9393
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV      0.9151   0.9149   0.9141    0.9122    0.9122    0.9122    0.9121
## adjCV   0.9149   0.9146   0.9139    0.9119    0.9119    0.9119    0.9118
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## CV       0.9103    0.9098    0.9100    0.9102    0.9099    0.8936
## adjCV    0.9100    0.9095    0.9096    0.9099    0.9095    0.8930
##        20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
## CV       0.8899    0.8844    0.8847    0.8848    0.8843    0.8841
## adjCV    0.8893    0.8838    0.8841    0.8842    0.8837    0.8835
##
## TRAINING: % variance explained
##               1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X           38.736356   61.049   69.109   74.113    78.68    82.61
## Appliances   0.006548    1.455    1.804    2.418    10.08    14.94
##             7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
## X             85.99    88.83    90.73     92.48     93.93     95.16
## Appliances    19.38    19.46    19.61     20.01     20.02     20.08
##             13 comps  14 comps  15 comps  16 comps  17 comps  18 comps
## X              96.35     97.24     97.92     98.50     98.93     99.20
## Appliances     20.15     20.56     20.72     20.75     20.85     21.08
##             19 comps  20 comps  21 comps  22 comps  23 comps  24 comps
## X              99.44     99.64     99.80     99.93     99.97    100.00
## Appliances     24.17     24.86     25.85     25.85     25.87     26.03
##             25 comps
## X             100.00
## Appliances     26.12
```

```
par(mfrow=c(3,1))
plot(RMSEP(pcr_model), legendpos = "topright")
plot(pcr_model, ncomp = 6, asp = 1, line = TRUE)
plot(pcr_model, "loadings", comps = 1:6, legendpos = "topleft")
```




```
pc <- pcr_model$scores[,1:6] #최적 주성분 추출


# 모델 수립 결과
reg_pc <- lm(train$Appliances ~ pc)
summary(reg_pc) #회귀모델 요약
```

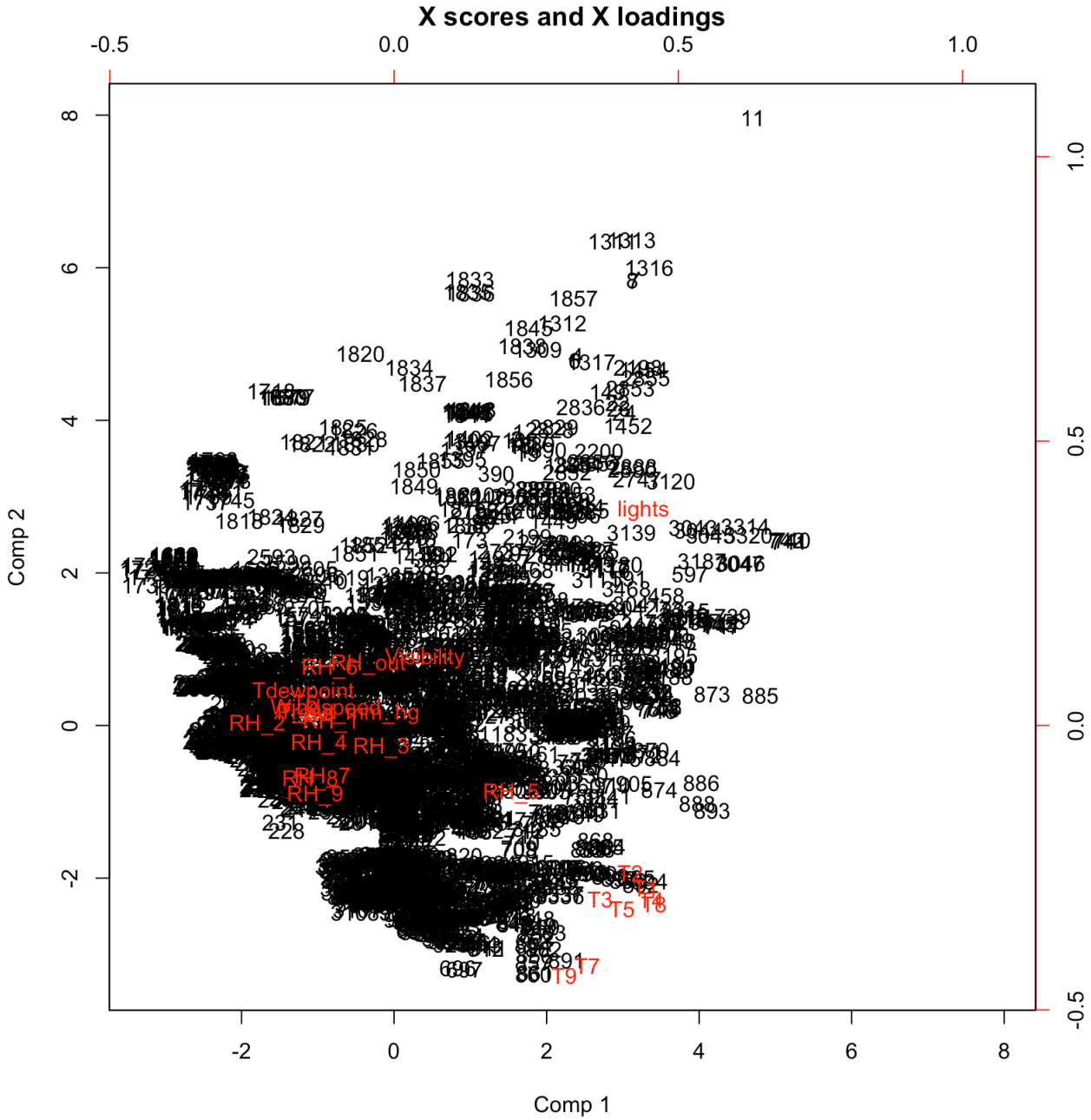
```
##
## Call:
## lm(formula = train$Appliances ~ pc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8139 -0.4231 -0.2134  0.0220  7.4491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.006914   0.018298   0.378  0.70558
## pcComp 1     0.002641   0.005886   0.449  0.65376
## pcComp 2     0.051745   0.007756   6.672 3.07e-11 ***
## pcComp 3    -0.042289   0.012904  -3.277  0.00106 **
## pcComp 4    -0.071092   0.016378  -4.341 1.48e-05 ***
## pcComp 5     0.262939   0.017140  15.341  < 2e-16 ***
## pcComp 6    -0.225921   0.018477 -12.227  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9368 on 2614 degrees of freedom
## Multiple R-squared:  0.1494, Adjusted R-squared:  0.1475
## F-statistic: 76.52 on 6 and 2614 DF,  p-value: < 2.2e-16
```

```
# 예측력 평가
pred_pcr <- predict(pcr_model, test, ncomp = 6)
err(test$Appliances, pred_pcr)
```

```
##       MAE       MSE      RMSE
## 0.5395025 0.7820893 0.8843581
```

# 5. PLSR

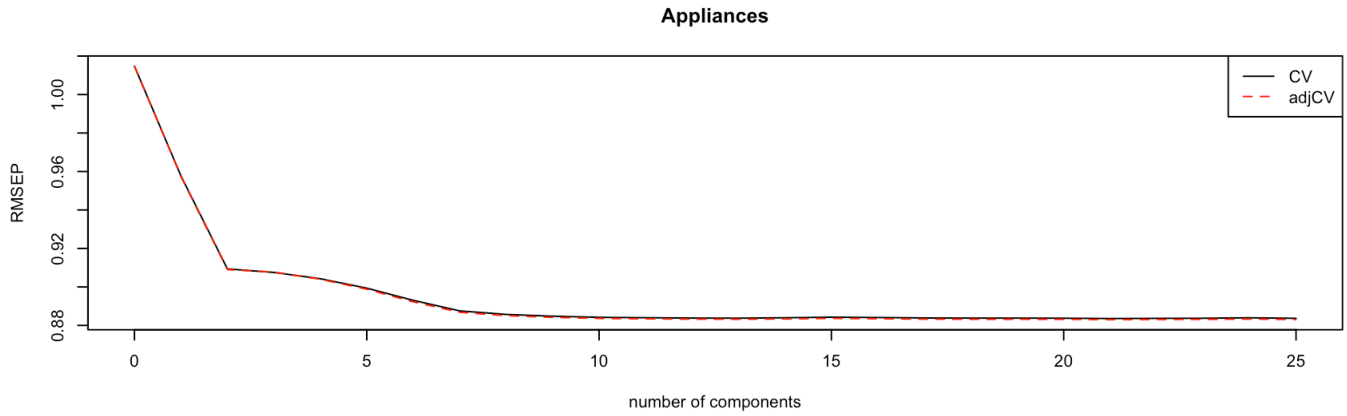
```
# 모델 수립
plsr_model <- plsr(Appliances ~ ., data = train, validation = 'CV')

#잠재변수 검토
par(mfrow=c(1,1))
biplot(plsr_model)
```

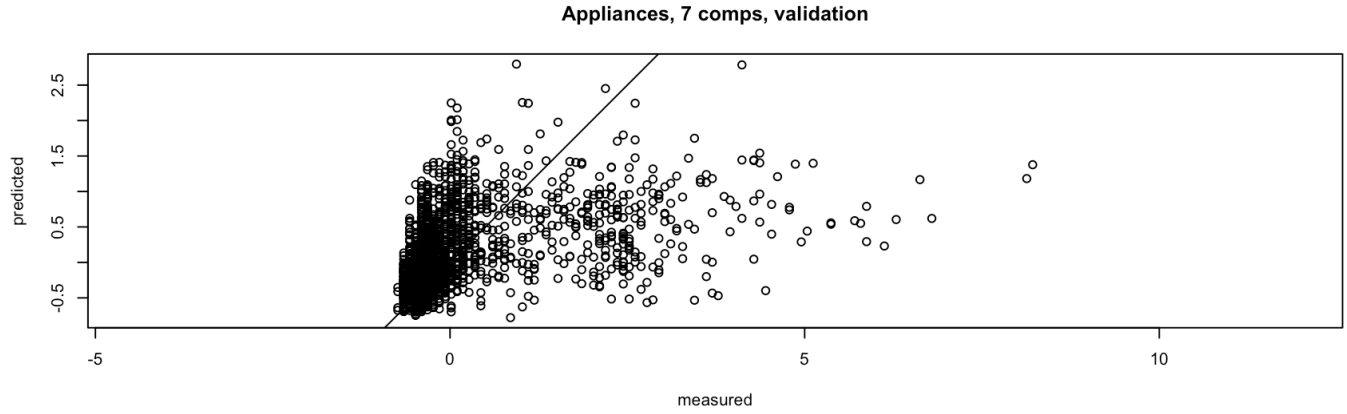


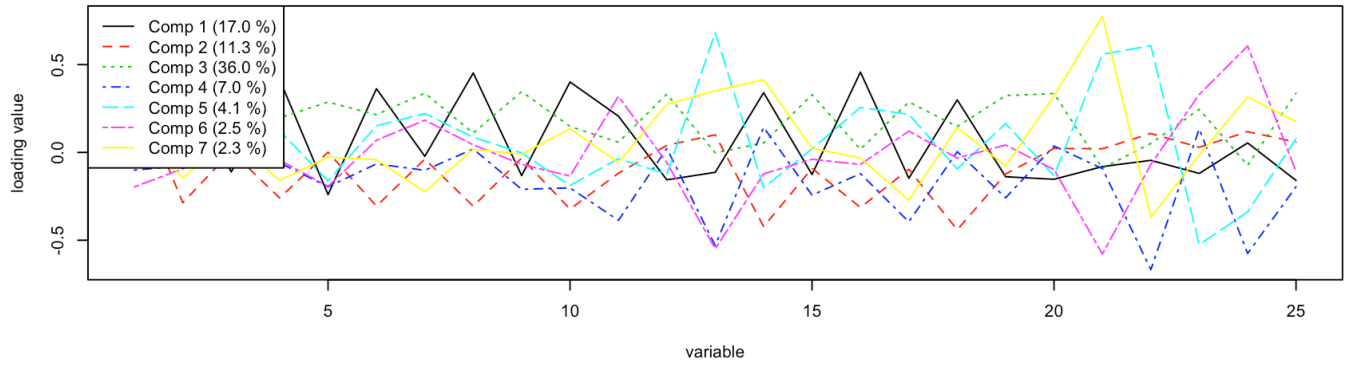

```
summary(plsr_model)
```

```
## Data:    X dimension: 2621 25
##  Y dimension: 2621 1
## Fit method: kernelpls
## Number of components considered: 25
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           1.015   0.9578   0.9094   0.9075   0.9043   0.8994   0.8930
## adjCV        1.015   0.9574   0.9090   0.9077   0.9040   0.8988   0.8923
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV      0.8875   0.8857   0.8848    0.8842    0.8840    0.8838    0.8838
## adjCV   0.8868   0.8851   0.8842    0.8836    0.8834    0.8833    0.8832
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## CV       0.8840    0.8843    0.8841    0.8839    0.8838    0.8838
## adjCV    0.8834    0.8837    0.8835    0.8833    0.8832    0.8832
##        20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
## CV       0.8838    0.8836    0.8837    0.8837    0.8839    0.8837
## adjCV    0.8832    0.8830    0.8831    0.8831    0.8833    0.8831
##
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X             16.95    28.21    64.24    71.25    75.40    77.91    80.19
## Appliances    11.55    20.24    20.69    21.65    23.01    24.56    25.39
##             8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X             83.32    85.85     88.74     90.21     92.08     93.27
## Appliances    25.68    25.82     25.87     25.91     25.93     25.95
##             14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## X              94.45     95.43     96.01     96.96     97.57     98.28
## Appliances     25.97     25.99     26.02     26.03     26.04     26.05
##             20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
## X              99.04     99.49     99.66     99.86     99.90    100.00
## Appliances     26.05     26.05     26.06     26.06     26.11     26.12
```

```
par(mfrow=c(3,1))
plot(RMSEP(plsr_model), legendpos = "topright")
plot(plsr_model, ncomp = 7, asp = 1, line = TRUE)
plot(plsr_model, "loadings", comps = 1:7, legendpos = "topleft")
```




```
lv <- plsr_model$scores[,1:7] #최적 주성분 추출

# 모델 수립 결과
reg_pls <- lm(train$Appliances ~ lv)
summary(reg_pls) #회귀모델 요약
```

```
##
## Call:
## lm(formula = train$Appliances ~ lv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1489 -0.3906 -0.1475  0.0680  6.9130
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.006914   0.017141   0.403    0.687
## lvComp 1    0.223178   0.011095  20.116  < 2e-16 ***
## lvComp 2    0.191004   0.010951  17.442  < 2e-16 ***
## lvComp 3    0.026004   0.006539   3.977 7.18e-05 ***
## lvComp 4    0.099612   0.017179   5.799 7.49e-09 ***
## lvComp 5    0.161375   0.023441   6.884 7.24e-12 ***
## lvComp 6    0.191109   0.025938   7.368 2.31e-13 ***
## lvComp 7    0.154734   0.028701   5.391 7.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8776 on 2613 degrees of freedom
## Multiple R-squared:  0.2539, Adjusted R-squared:  0.2519
## F-statistic:   127 on 7 and 2613 DF,  p-value: < 2.2e-16
```

```
# 예측력 평가
pred_plsr <- predict(plsr_model, test, ncomp = 7)
err(test$Appliances, pred_plsr)
```

```
##       MAE       MSE      RMSE
## 0.4942616 0.7135533 0.8447208
```