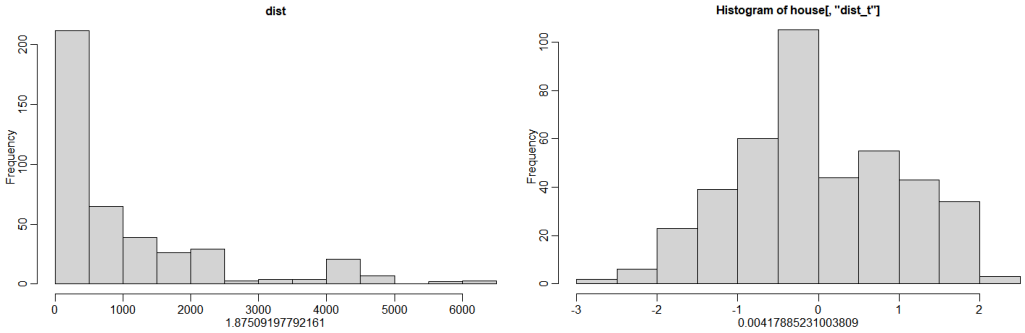
[Statistical Modeling and Machine Learning HW 2]

2017311974 통계학과 진수정

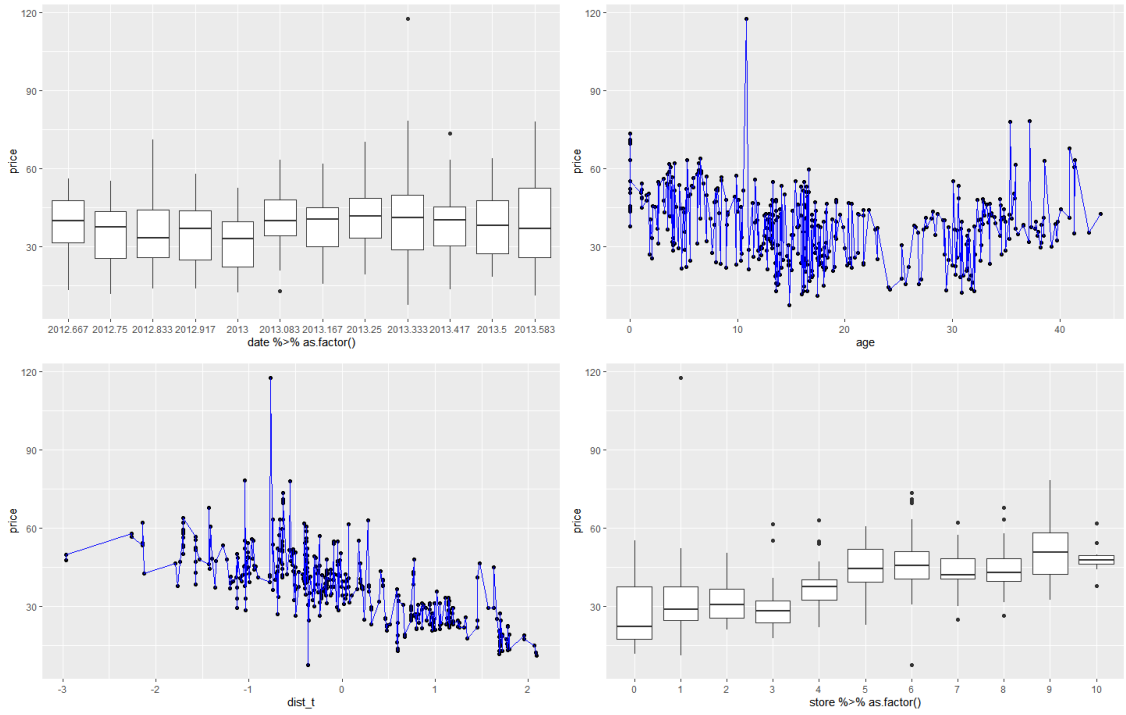
1. (1) AIC = 2904.761

1) Data Transformation



- dist변수가 Right-skewed 되어 있으므로 Yeo-Johnson Transformation 시켜주었다.

2) Data Visualization



- Age가 대략 20 이하일 때는 감소하다가 그 이후로는 약간 증가하는 패턴을 확인할 수 있다.

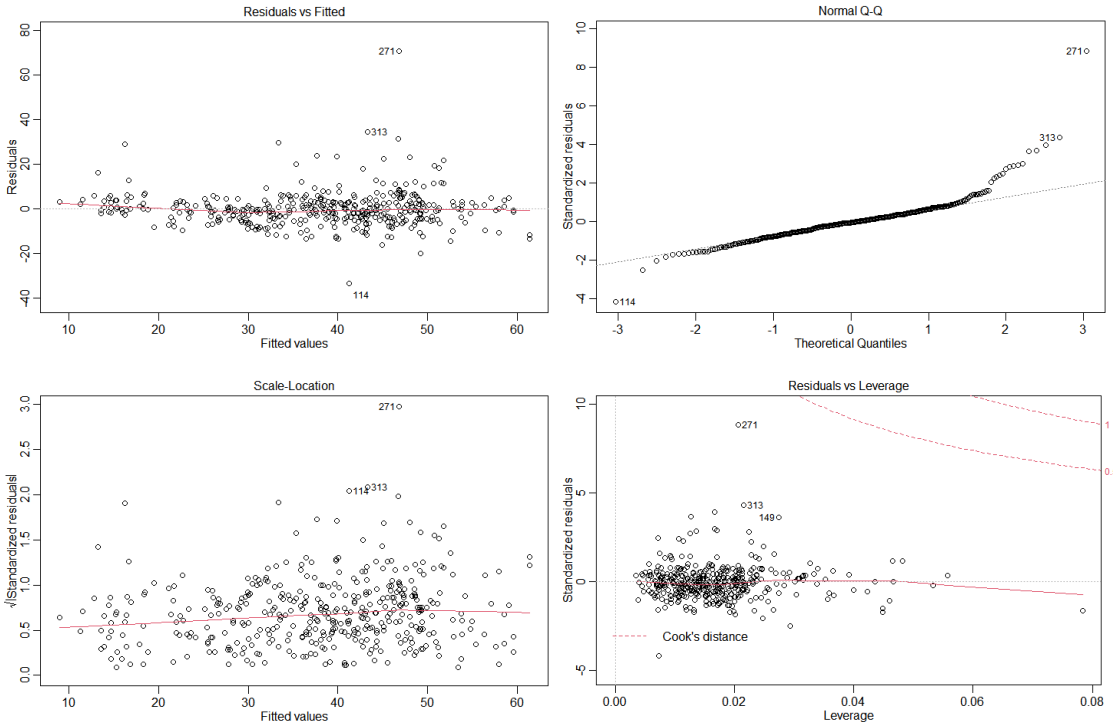
- Dist_t에 대해서는 감소하는 패턴을, store에 대해서는 증가하는 패턴을 확인할 수 있다.

3) Modeling

- Linear Regression 적용

```
> fit1_1 = lm(price ~ date + age + dist_t + store + lat + lon, data = house)
> summary(fit1_1)
Call:
lm(formula = price ~ date + age + dist_t + store + lat + lon,
    data = house)

Residuals:
    Min      1Q  Median      3Q     Max
-33.664  -4.290  -0.467   2.957  70.668

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.310e+04  4.951e+03  -4.666 4.18e-06 ***
date         6.550e+00  1.429e+00   4.584 6.08e-06 ***
age         -2.295e-01  3.540e-02  -6.483 2.60e-10 ***
dist_t      -7.409e+00  6.524e-01 -11.355  < 2e-16 ***
store        3.644e-01  1.915e-01   1.903   0.0578 .
lat          2.856e+02  3.750e+01   7.617 1.84e-13 ***
lon          2.326e+01  3.442e+01   0.676   0.4997
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.081 on 407 degrees of freedom
Multiple R-squared:  0.6524,    Adjusted R-squared:  0.6473
F-statistic: 127.3 on 6 and 407 DF,  p-value: < 2.2e-16
```

```
> plot(fit1_1)
```



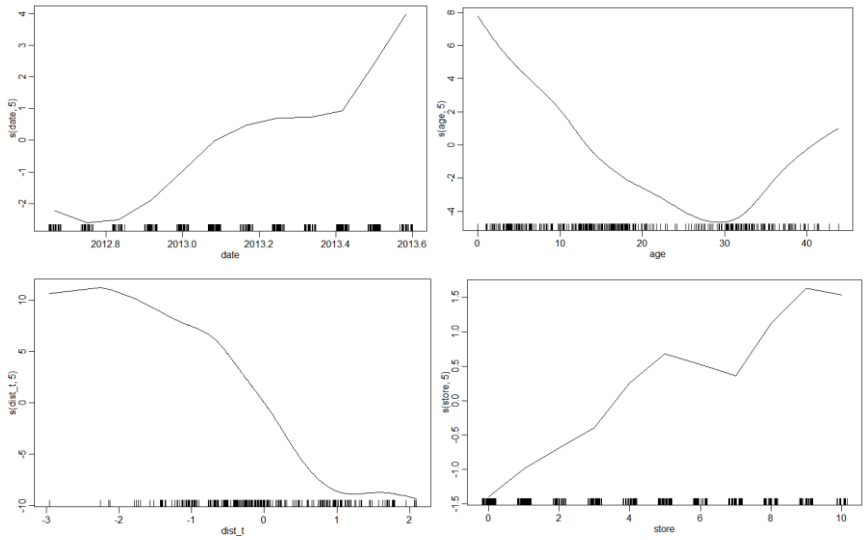

```
> dwtest(fit1_1)
```

```
        Durbin-Watson test

data:  fit1_1
DW = 2.1447, p-value = 0.9312
alternative hypothesis: true autocorrelation is greater than 0
```

: Error Assumption을 크게 위배하고 있지 않는 것으로 보인다.

- GAM 적용

```
> fit2_1 = gam(price ~ s(date,5) + s(age,5) + s(dist_t,5) + s(store,5) + lat +
lon, data = house)
> plot(fit2_1)
```




```
> fit2_1 = lm(price ~ date + poly(age,2) + dist_t + store + lat + lon, data =
house)
> summary(fit2_1)
Call:
lm(formula = price ~ date + poly(age, 2) + dist_t + store + lat +
    lon, data = house)

Residuals:
    Min      1Q  Median      3Q     Max
-31.734  -4.003  -0.148   3.332  72.177

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.383e+04  4.825e+03  -4.939 1.15e-06 ***
date           6.718e+00  1.392e+00   4.826 1.98e-06 ***
poly(age, 2)1 -5.445e+01  7.987e+00  -6.818 3.36e-11 ***
poly(age, 2)2  4.108e+01  8.563e+00   4.798 2.26e-06 ***
dist_t        -6.525e+00  6.616e-01  -9.863  < 2e-16 ***
store          3.928e-01  1.866e-01   2.105   0.0359 *
lat            2.872e+02  3.652e+01   7.864 3.39e-14 ***
lon            2.608e+01  3.353e+01   0.778   0.4372
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.871 on 406 degrees of freedom
Multiple R-squared:  0.6711,    Adjusted R-squared:  0.6654
F-statistic: 118.3 on 7 and 406 DF,  p-value: < 2.2e-16
```

- SAR 적용

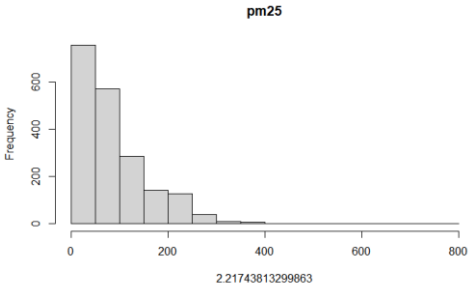
```
> glst = lapply(dists,function(d) exp(-100*d))
> lw = nb2listw(dnb,glist = glst,style = 'C')
```

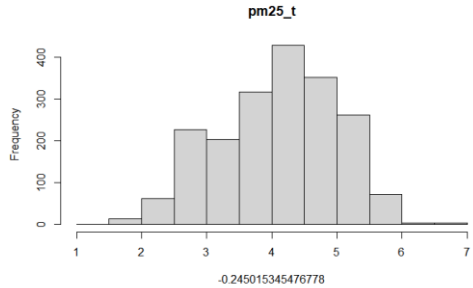
```
> fit = lagsarlm(price ~ date + poly(age,2) + dist_t + store, data = house,
listw = lw)
> summary(fit)
> AIC(fit)
[1] 2904.761
Call:lagsarlm(formula = price ~ date + poly(age, 2) + dist_t + store,
    data = house, listw = lw)

Residuals:
      Min        1Q    Median        3Q       Max
-33.83235  -4.63963  -0.50944   2.96144  72.70120

Type: lag
Coefficients: (numerical Hessian approximate standard errors)
                 Estimate  Std. Error   z value  Pr(>|z|)
(Intercept)   -1.3890e+04  1.4979e+01 -927.2901 < 2.2e-16
date           6.9128e+00  7.4412e-03  928.9853 < 2.2e-16
poly(age, 2)1 -6.8398e+01  8.4108e+00   -8.1321 4.441e-16
poly(age, 2)2  3.3959e+01  8.6541e+00    3.9241 8.706e-05
dist_t        -4.4104e+00  7.3997e-01   -5.9603 2.518e-09
store          2.7388e-01  1.9193e-01    1.4270    0.1536

Rho: 0.25643, LR test value: 49.403, p-value: 2.0846e-12
Approximate (numerical Hessian) standard error: 0.035329
    z-value: 7.2585, p-value: 3.9124e-13
Wald statistic: 52.686, p-value: 3.9124e-13

Log likelihood: -1444.38 for lag model
ML residual variance (sigma squared): 62.766, (sigma: 7.9225)
Number of observations: 414
Number of parameters estimated: 8
AIC: 2904.8, (AIC for lm: 2952.2)
```

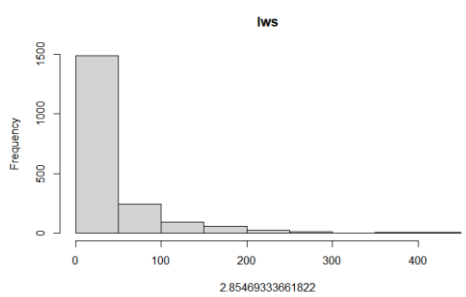
1. (2)
   - ✓ 2012 ~ 2013 년에 걸쳐 집값이 전반적으로 올랐다.
   - ✓ 새로 지은 집일수록 집값이 비싸지만, 특정 지점 이후에는 집의 연식이 있을수록 집값이 오르기도 한다.
   - ✓ 지하철역에서의 거리가 가까울수록 집값이 비싸다.
   - ✓ (통계적으로 유의한 수준은 아니지만) 집 근처에 편의점의 수가 많을수록 집값이 비싸다.
   - ✓ 집과 집 사이의 공간적인 거리가 멀 때 집값의 상관관계는 지수적으로 감소한다.

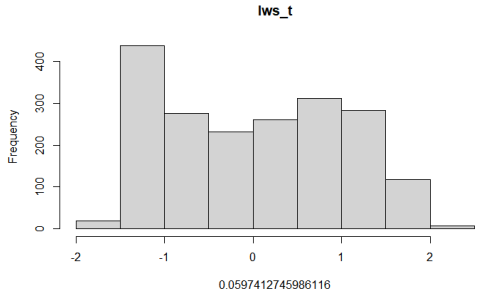
2. (1)

1) Data Transformation







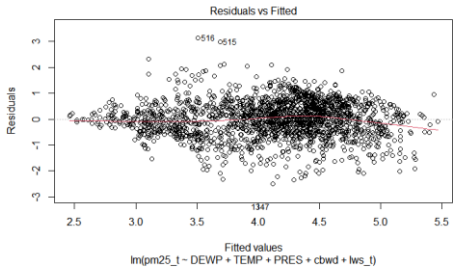


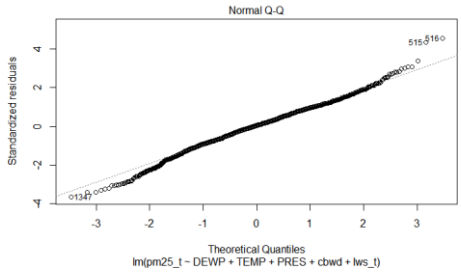
- Pm25 와 Iws 가 Right-skewed 되어 있어서 Transform 시켰다. 이 때 pm25 는 MSE 계산의 편의성을 위해서 로그를 취했고, Iws 는 Yeo-Johnson Transformation 을 이용하였다.

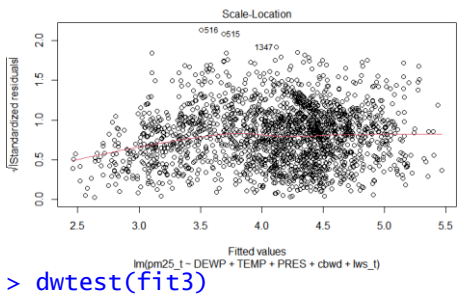
2) Modeling

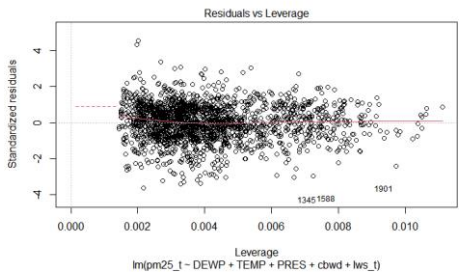
- Linear Regression 적용

```
> fit3 = lm(pm25_t ~ DEWP + TEMP + PRES + cbwd + Iws_t, data = pm25_tr2)
> summary(fit3)
Call:
lm(formula = pm25_t ~ DEWP + TEMP + PRES + cbwd + Iws_t, data = pm25
_tr2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50994 -0.43654  0.03104  0.45840  3.12807

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.133008   2.743975  10.253  < 2e-16 ***
DEWP         0.040663   0.002386  17.041  < 2e-16 ***
TEMP        -0.036468   0.002772 -13.154  < 2e-16 ***
PRES        -0.023214   0.002686  -8.642  < 2e-16 ***
cbwdNE      -0.145931   0.068377  -2.134 0.032950 *
cbwdNW      -0.207274   0.058883  -3.520 0.000441 ***
cbwdSE       0.365023   0.055603   6.565 6.67e-11 ***
Iws_t       -0.204403   0.020285 -10.077  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6874 on 1936 degrees of freedom
Multiple R-squared:  0.4188,    Adjusted R-squared:  0.4167
F-statistic: 199.3 on 7 and 1936 DF,  p-value: < 2.2e-16
> plot(fit3)
```










```
> dwtest(fit3)
```

```
        Durbin-Watson test

data:  fit3
DW = 0.23926, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

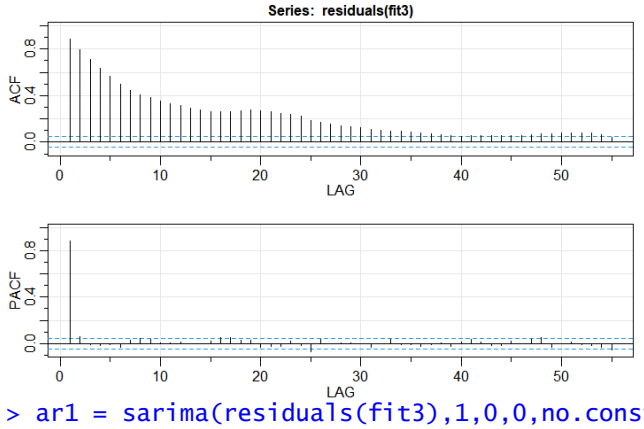
: Error 가 independent 하지 않는 문제가 발생하는데, 이를 AR 모델을 사용하여 해결하고자 한다.

- GAM 적용

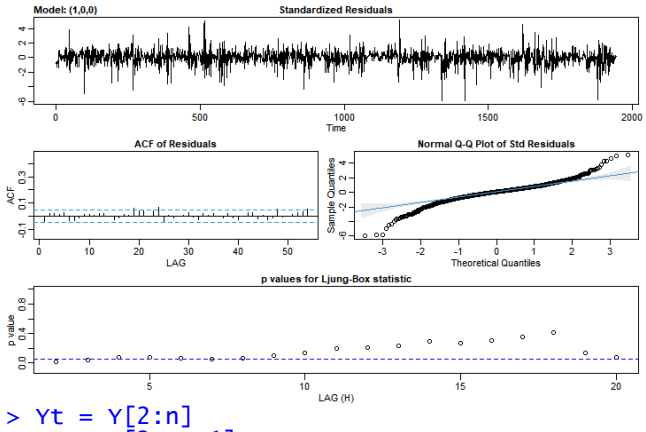
```
> fit4 = gam(pm25_t ~ s(DEWP,5) + s(TEMP,5) + s(PRES,5) + cbwd + s(Iws_t,5), data = pm25_tr2)
> fit3 = lm(pm25_t ~ DEWP + TEMP + poly(PRES,2) + cbwd + poly(Iws_t,2), data = pm25_tr2)
> summary(fit3)
Call:
lm(formula = pm25_t ~ DEWP + TEMP + poly(PRES, 2) + cbwd + poly(Iws_
t,
    2), data = pm25_tr2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.67892 -0.40899  0.04728  0.45039  2.95381

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.693182   0.060710  77.305  < 2e-16 ***
DEWP             0.039828   0.002357  16.898  < 2e-16 ***
TEMP            -0.038876   0.002757 -14.102  < 2e-16 ***
poly(PRES, 2)1  -9.897728   1.031125  -9.599  < 2e-16 ***
poly(PRES, 2)2  -2.876704   0.681344  -4.222 2.53e-05 ***
cbwdNE          -0.278415   0.070490  -3.950 8.10e-05 ***
cbwdNW          -0.349714   0.062050  -5.636 2.00e-08 ***
cbwdSE           0.180530   0.061712   2.925  0.00348 **
poly(Iws_t, 2)1 -7.597284   0.906906  -8.377  < 2e-16 ***
poly(Iws_t, 2)2 -5.010443   0.780290  -6.421 1.70e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6782 on 1934 degrees of freedom
Multiple R-squared:  0.4348,    Adjusted R-squared:  0.4322
F-statistic: 165.3 on 9 and 1934 DF,  p-value: < 2.2e-16
```

- AR 모델 적용

```
> acf2(residuals(fit3))
```




```
> ar1 = sarima(residuals(fit3),1,0,0,no.constant = T)
```




```
> Yt = Y[2:n]
> Xt = X[2:n,-1]
> et = residuals(fit3)[1:(n-1)]
> beta.old2 = rep(0,9)
> mdif = 10000
>
> while (mdif > 0.0000001) {
+   fit.temp = lm(Yt ~ Xt + et)
+   beta.new2 = fit.temp$coefficients
+   mdif = max(abs(beta.new2[1:8] - beta.old2[1:8]))
+
+   et = (Y - X %*% beta.new2[1:8])[1:(n-1)]
+   beta.old2 = beta.new2
+ }
> summary(fit.temp)
Call:
lm(formula = Yt ~ Xt + et)

Residuals:
     Min       1Q   Median       3Q      Max
-1.82317 -0.13145  0.02141  0.15369  1.74542

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.3804515  1.1480585  15.139   <2e-16 ***
XtDEWP       0.0424497  0.0009938  42.715   <2e-16 ***
XtTEMP      -0.0306556  0.0011566 -26.504   <2e-16 ***
XtPRES      -0.0126862  0.0011239 -11.288   <2e-16 ***
XtcbwdNE    -0.0646336  0.0284846  -2.269   0.0234 *
XtcbwdNW    -0.0399412  0.0245811  -1.625   0.1044
XtcbwdSE     0.1994812  0.0232240   8.589   <2e-16 ***
XtIws_t     -0.1858935  0.0084489 -22.002   <2e-16 ***
et           0.9139016  0.0095130  96.068   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2862 on 1934 degrees of freedom
Multiple R-squared:  0.8993,    Adjusted R-squared:  0.8989
F-statistic:  2159 on 8 and 1934 DF,  p-value: < 2.2e-16
> dwtest(fit.temp)
        Durbin-Watson test

data:  fit.temp
DW = 1.9561, p-value = 0.1352
alternative hypothesis: true autocorrelation is greater than 0
```

2. (2) test MSE = 1752.367

```
> errs2 = data.frame(
+   past_err = (Y - X %*% beta.new2[1:8])[1:(n-1)],
+   err = (Y - X %*% beta.new2[1:8])[2:n]
+ )
> fit6 = lm(err ~ past_err, data = errs2)
> err = (Y - X %*% beta.new2[1:8])[n]
> for (i in 1:120) {
+   err_fit = cbind(1,err) %*% fit6$coefficients
+   pred_errset2[i] = err_fit
```

```
+   err = err_fit
+ }
```

- (t-1) 시점에서의 error 를 이용하여 t 시점에서의 error 를 추정하였다.

```
> AR2_te = cbind(X_te,c((Y - X %*% beta.new2[1:8])[n],pred_errset2[-120])) %*%
beta.new2
> MSE(Y_te,AR2_te)
[1] 0.2477902
> MSE(exp(Y_te),exp(AR2_te))
[1] 1752.367
```

- Test data 의 X 정보와 predicted error 를 이용하여 test MSE 를 계산하였다. Transformed Y 를 기존의 Y 로 되돌리는 작업을 통해 MSE 를 계산한 결과 test MSE 는 1752.367 로 나왔다.