경제자료분석

프로젝트

212STG05 김현민

A. 어떤 상품의 수요탄력성 계수를 추정하려고 한다. 아래 절차를 따라 계수의 추정치와 그 신뢰구간을 구하고 결과를 설명하시오.

(0) 데이터를 설명하고 분석의 목적을 기술하시오.

데이터는 미국의 2000년 1월부터 2022년 3월까지의 월별 천연가스 가격, 소비량, 생산자물가지수에 관한 자료로, 각 변수에 대한 설명은 다음과 같다.

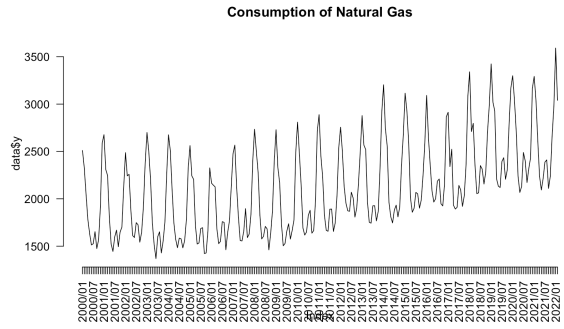
· Date : 2000년 1월부터 2022년 3월까지의 월별 데이터
· Price(X) : 천연가스 가격(US Dollors per MMBTU) (출처 : IndexMundi)
· Consumption(Y) : 천연가스 소비량(Billion Cubic Feet) (출처 : Federal Reserve Economic Data)
· PPI(Z) : 생산자물가지수(Index 1982=100) (출처 : Federal Reserve Economic Data)

일반적으로 천연가스 가격이 증가할수록 천연가스 소비량이 감소할 것이라고 생각한다. 이를 확인해보기 위해 천연가스의 수요탄력성을 추정해보고, 새로운 변수도 추가하여 개선된 수요탄력성을 추정해보고자 한다. 또한, ARIMA 모형, ADL 모형, VAR 모형을 구축하여 향후 5개월의 천연가스 소비량을 예측해보고, 세 모형의 예측력을 비교해보는 것을 분석 목적으로 한다.

```
library(readxl)
setwd("/Users/kimhyunmin")
gasdata <- read_excel("gasdata.xlsx", sheet = 'Sheet1')
names(gasdata) <- c("Date","x","y","z")
data <- gasdata[1:266,]
```

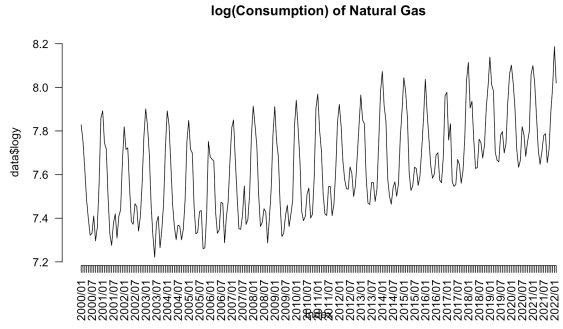
(1) 그 상품 로그수요량 Y의 시도표를 그려보시오. 시간 축에 실제 시간이 표시되도록 한다.

수요량 Y의 시도표




```
plot(data$y,axes=F,type="l",main="Consumption of Natural Gas")
axis(1,at=1:266,lab=c(data$Date),las=2)
axis(2,las=1)
```

로그수요량 Y의 시도표

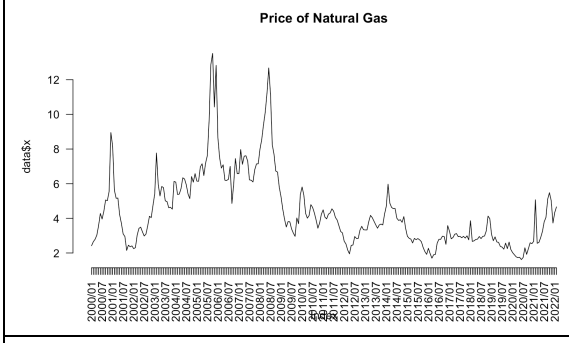



```
data$logy <- log(data$y)
plot(data$logy,axes=F,type="l",main="log(Consumption) of Natural Gas")
axis(1,at=1:266,lab=c(data$Date),las=2)
axis(2,las=1)
```

수요량 Y의 시도표와 로그수요량 Y의 시도표를 그려본 결과, 수요량 Y보다 로그수요량 Y가 조금 더 등분산이므로 로그수요량 Y를 사용하는 것이 더 좋다.

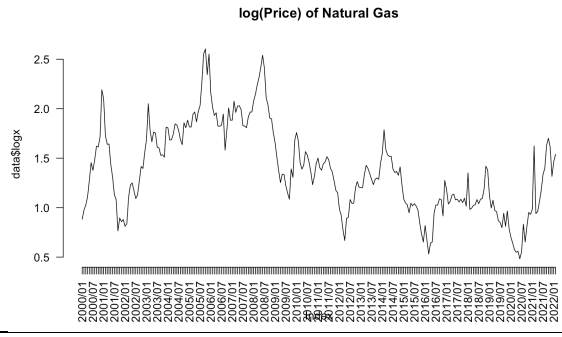
(2) 그 상품의 로그가격 X의 시도표를 그려보시오.

가격 X의 시도표




```
plot(data$x,axes=F,type="l",main="Price of Natural Gas")
axis(1,at=1:266,lab=c(data$Date),las=2)
axis(2,las=1)
```

로그가격 X의 시도표




```
data$logx <- log(data$x)
plot(data$logx,axes=F,type="l",main="log(Price) of Natural Gas")
axis(1,at=1:266,lab=c(data$Date),las=2)
axis(2,las=1)
```

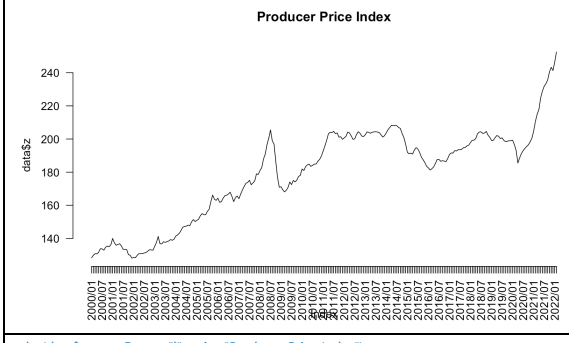
가격 X의 시도표와 로그가격 X의 시도표를 그려본 결과, 가격 X보다 로그가격 X가 조금 더 등분산이므로 로그가격 X를 사용하는 것이 더 좋다.

(3) 위 (1), (2) 자료를 이용하여 수요탄력성 β_1의 추정치 $\widehat{\beta_1}$을 계산하시오. $\widehat{\beta_1}$ = -0.1473

```
lm(logy ~ logx, data)
```

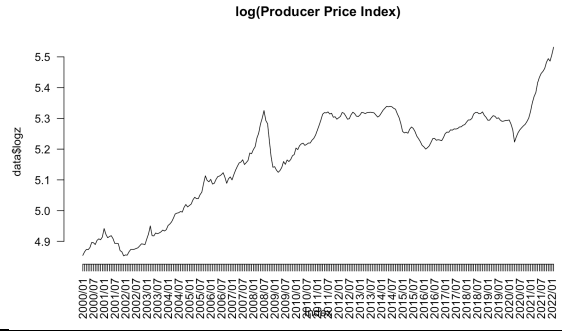
(4) 위 (3)의 추정치를 개선하기위해 다른 추가 변수 Z를 하나 찾아보고, 그 시도표를 그리시오. 개선된 추정치 $\widehat{\beta_1}$을 구하시오. $\widehat{\beta_1}$ = -0.1051

Z의 시도표




```
plot(data$z,axes=F,type="l",main="Producer Price Index")
axis(1,at=1:266,lab=c(data$Date),las=2)
axis(2,las=1)
```

로그 Z의 시도표




```
data$logz <- log(data$z)
plot(data$logz,axes=F,type="l",main="log(Producer Price Index)")
axis(1,at=1:266,lab=c(data$Date),las=2)
axis(2,las=1)
```

Z의 시도표와 로그 Z의 시도표를 그려본 결과, Z보다 로그 Z가 조금 더 등분산이므로 로그가격 Z를 사용하는 것이 더 좋다. 또한, (3)의 추정치보다 다른 변수가 추가된 (4)의 추정치가 좀 더 신뢰성있다.

```
lm(logy ~ logx+logz, data=data)
```

(5) 위 (4)에서의 회귀모형에서 오차항의 등분산성에 대한 검정을 하고, 검정 결과를 설명하시오.

오차항의 등분산성을 검정하고자 White 검정을 수행한 결과 W=10.06261 > $X_{0.05}{}^2(4)$=9.49 이므로 등분산성 가정이 기각된다.

```
lm.fit <- lm(logy ~ logx+logz, data=data)
e.square <- lm.fit$residuals^2
data$logx.square <- data$logx^2
data$logz.square <- data$logz^2
white.test <- lm(e.square ~ logx+logz+logx.square+logz.square, data=data)
summary(white.test)
W = summary(white.test)$r.square*nrow(data)
W
```

(6) 위 (4)에서의 회귀모형에서 오차항의 무자기상관성에 대한 검정을 하고, 검정 결과를 설명하시오.

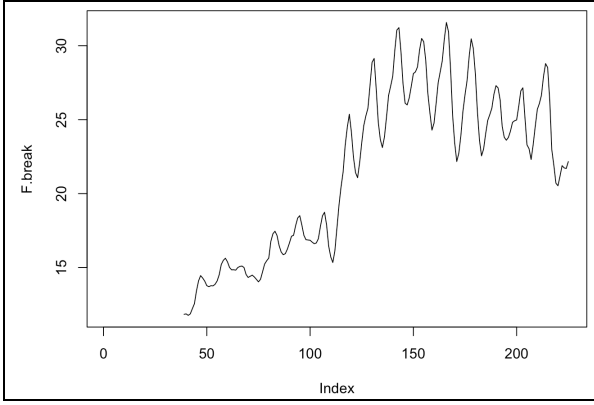
```
        Durbin-Watson test

data:  lm.fit
DW = 0.49509, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

```
library(lmtest)
dwtest(lm.fit)
```

오차항의 무자기상관성을 검정하기 위해 Durbin-Watson 검정 결과, DW = 0.49509 < 2이고 p-value가 거의 0 값이므로 오차에 강한 양의 자기상관이 있음을 알 수 있다.

(7) 위 (4)에서의 회귀모형에서 상수항을 β_0라 하였을 때, 모수 (β_0, β_1)이 시간에 따라 변했는지 QLR 검정을 통해 판단하시오.




```
n = length(data$y)
ols = lm(logy ~ logx, data)
t = seq(1:n)
F.break = c()
for (tau in (0.15*n):(0.85*n)){
  D = ifelse(t <= tau, 0, 1)
  OLS.break = lm(logy ~ 1+logx+D+logx:D, data=data)
  F.break[tau] = anova(ols, OLS.break)$F[2]
}
plot(F.break, type="l")
max(F.break[(0.15*n):(0.85*n)])
which.max(F.break)
```

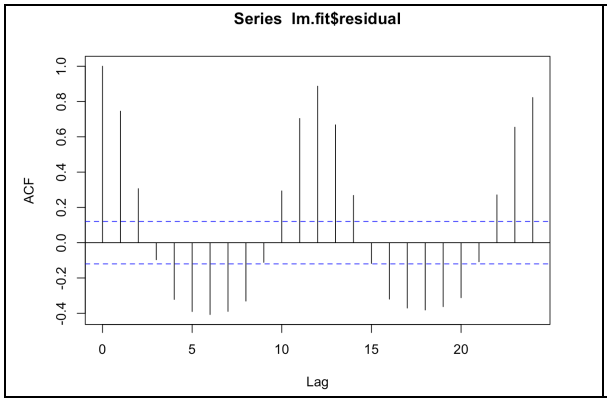
supWald = 31.57335 > 1% critical value = 7.78 이므로 브래이크가 존재한다는 것을 알 수 있다. 즉, 모수는 시간에 따라 변하였고, $\hat{\tau}$=166(2013년 10월)을 기점으로 모수가 바뀌었다.

(8) 추정치 $\widehat{\beta_1}$의 OLS 표준오차, HC 표준오차, HC 표준오차를 구하시오

OLS 표준오차	HC 표준오차	HC 표준오차
0.02779277	0.02774755	0.02986575

```
lm.fit <- lm(logy ~ logx+logz, data=data)
ols.se = summary(lm.fit)$coef[2,2]
ols.se
library(sandwich)
HC.se = sqrt(vcovHC(lm.fit)[2,2])
HC.se
HAC.se = sqrt(vcovHAC(lm.fit)[2,2])
HAC.se
```

(9) 위 (7)의 표준오차 중 가장 적절한 것을 선택하고 선택한 이유를 설명하시오.




```
acf(lm.fit$residual)
```

ols 잔차의 acf를 그려본 결과, ±2표준오차 밴드를 벗어나는 부분이 있기 때문에 오차항이 자기상관을 가짐을 알 수 있다. 따라서 HAC 표준오차를 사용하는 것이 적절하다.

(10) 수요탄력성 β_1의 95% 신뢰구간을 구하시오.

수요탄력성 β_1의 95% 신뢰구간 : (-0.16359005, -0.04651631)

```
CI = c(lm.fit$coefficients[[2]]-1.96*HAC.se, lm.fit$coefficients[[2]]+1.96*HAC.se)
CI
```

B. 위 A의 종속변수 Y를 예측하고자 한다. 마지막 시점을 T라 했을 때, 예측 대상은 Y_{T+h}, h=1,2,3,4,5이다.

(1) ARIMA 모형에 의한 예측치를 구하고 예측치, 95% 예측 구간을 시도표에 이어서 그리시오. 모형 차수는 BIC 기준으로 정하시오.

step1) Y와 로그 Y의 시도표 확인

A. (1)에서 수요량 Y보다 로그수요량 Y가 좀 더 등분산이므로 로그수요량 Y에 대해 ARIMA 분석을 한다.

step2) 차분 검토

```
Title:
Augmented Dickey-Fuller Test

Test Results:
  PARAMETER:
    Lag Order: 14
  STATISTIC:
    Dickey-Fuller: -0.0065
  P VALUE:
    0.9545
```

```
library(forecast)
aic = c()
for (p in 1:20){
  ar.fit = Arima(data$logy, order=c(p,0,0), method="ML")
  aic[p] = ar.fit$aic
}
which.min(aic) #15

library(fUnitRoots)
adfTest(data$logy, type="c", lags=14)
```

p-value = 0.9545로 유의수준 5%에서 단위근 가설 기각하지 못하므로 차분이 필요하다.

step3) ARIMA(p,1,q) 모형 식별

```
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] -317.9472 -361.1587 -364.1593 -404.9519 -429.9128
[2,] -355.7794 -357.8024 -360.3098 -420.0440 -424.3370
[3,] -368.8149 -440.5192 -419.9378 -476.2076 -418.7583
[4,] -386.8277 -438.6820 -433.1470 -470.6281 -467.4085
[5,] -382.8009 -380.3317 -429.2244 -469.7548 -463.8061
```

```
bic = matrix(rep(0, 5*5), 5, 5)
for (p in 1:5){
  for (q in 1:5){
    ari.fit = Arima(data$logy, order=c(p-1,1,q-1))
    bic[p,q] = ari.fit$bic
  }
}
bic
```

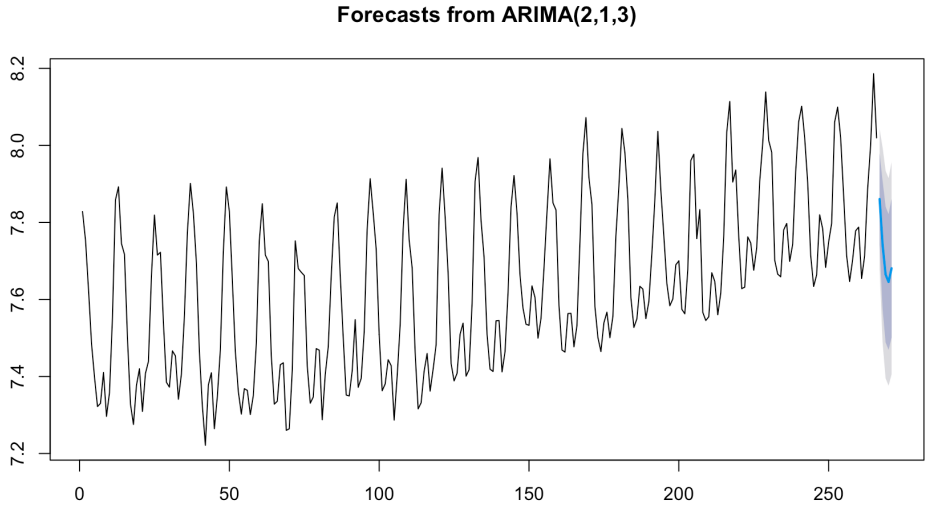
BIC 결과 중 가장 작은 값은 -476.2076이므로 차수 p와 q의 order는 각각 2,3이다.

step4) ARIMA(2,1,3) 모형으로 구축한 후 예측한 5개월 예측치 및 95% 신뢰구간

Date	예측치	95% 신뢰구간
2022/04	7.860805	(7.679370, 8.042239)
2022/05	7.741805	(7.489116, 7.994493)
2022/06	7.664817	(7.395665, 7.933969)
2022/07	7.645431	(7.376189, 7.914673)
2022/08	7.681006	(7.405781, 7.956230)

```
arima.fit <- Arima(data$logy, order=c(2,1,3))
arima.hat = forecast(arima.fit, h=5)
arima.hat
```

step5) 향후 5개월을 예측한 전체 시도표




```
plot(arima.hat)
```

(2) 변수 X, Z를 추가적으로 고려한 ADL 모형에 의한 예측치를 구하고 예측치, 95% 예측 구간을 시도표에 이어서 그리시오. 모형 차수는 BIC 기준으로 정하시오.

step1) Y와 로그 Y의 시도표 확인

A. (1)에서 수요량 Y보다 로그수요량 Y가 좀 더 등분산이므로 로그수요량 Y에 대해 ARIMA 분석을 한다.

step2) 공적분 검토

Title:
Augmented Dickey-Fuller Test

Test Results:
PARAMETER:
Lag Order: 15
STATISTIC:
Dickey-Fuller: -1.6487
P VALUE:
0.4429

Asymptotic critical values for cointegration tests

k	Test statistic	1%	5%	10%
2	c	−3.90	−3.34	−3.04
	ct	−4.32	−3.78	−3.50
3	c	−4.29	−3.74	−3.45
	ct	−4.66	−4.12	−3.84
4	c	−4.64	−4.10	−3.81
	ct	−4.97	−4.43	−4.15
5	c	−4.96	−4.42	−4.13
	ct	−5.25	−4.72	−4.43
6	c	−5.25	−4.71	−4.42
	ct	−5.52	−4.98	−4.70

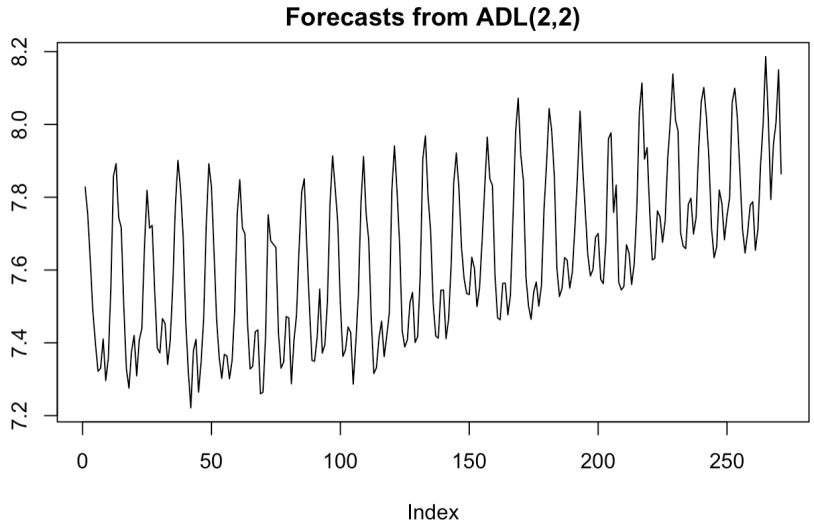
```
reg <- lm(logy ~ logx+logz, data=data)
aic = c()
for (p in 1:20){
  ar.z = Arima(reg$residual, order=c(p,0,0))
  aic[p] = ar.z$aic
}
which.min(aic)
plot(reg$residual, type="l")
adfTest(reg$residual, type="c", lags=15)
```

잔차에 대한 ADF 결과, ADF=-1.6487은 10%에서도 유의하지 않으므로 공적분 관계가 없다.

step3) B. (1)의 BIC order가 2이므로 ADL(2,2) 모형으로 구축한 후 예측한 5개월 예측치 및 전체 시도표

Date	예측치
2022/04	7.793934
2022/05	7.945394
2022/06	8.004816
2022/07	8.149869
2022/08	7.864140

향후 5개월을 예측한 전체 시도표




```
y0 = data$logy[4:266]; y1 = data$logy[3:265]; y2 = data$logy[2:264]; y3 = data$logy[1:263]
x0 = data$logx[4:266]; x1 = data$logx[3:265]; x2 = data$logx[2:264]; x3 = data$logx[1:263]
z0 = data$logz[4:266]; z1 = data$logz[3:265]; z2 = data$logz[2:264]; z3 = data$logz[1:263]

dy0 = y0-y1; dy1 = y1-y2; dy2 = y2-y3
dx0 = x0-x1; dx1 = x1-x2; dx2 = x2-x3
dz0 = z0-z1; dz1 = z1-z2; dz2 = z2-z3

y.hat = c()
n = length(dy0)
for (k in 1:5){
  N = n-5+k
  adl.fit = lm(dy0[1:N] ~ dy1[1:N]+dy2[1:N]+dx1[1:N]+dx2[1:N]+dz1[1:N]+dz2[1:N])
  beta = adl.fit$coef
  dy.hat = beta[1] + beta[2]*dy0[N] + beta[3]*dy1[N] + beta[4]*dx0[N] + beta[5]*dx1[N] + beta[6]*dz0[N] + beta[7]*dz1[N]
  y.hat[k] = log(gasdata[[N+3,3]])+dy.hat
}
y.hat

logy.data <- data[['logy']]
list <- c(logy.data, y.hat)
plot(list, type="l", main="Forecasts from ADL(2,2)")
```

(3) 위 (2)에서 X가 Y를 Granger Cause 하는지 검정하시오.

Analysis of Variance Table

Model 1: dy0 ~ dy1 + dy2
Model 2: dy0 ~ dy1 + dy2 + dx1 + dx2

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	260	3.6038				
2	258	3.4482	2	0.15552	5.8182	0.003377 **

```
reduced = lm(dy0 ~ dy1+dy2)
full = lm(dy0 ~ dy1+dy2+dx1+dx2)
anova(reduced, full)
```

분산분석 결과 p값이 유의하여 X가 Y를 Granger Cause 한다. 즉, X는 Y에 대해 추가 설명력이 있다.

(4) 변수 X, Y, Z에 대한 단위근 검정을 수행하고 결과를 설명하시오.

단위근 검정 결과

	lag	ADF	p-value
Y	14	-0.0065	0.9545
X	0	-2.8229	0.05913
Z	5	-1.7923	0.6634

단위근 검정 결과 변수 X, Y, Z는 p-value 값이 0.05보다 모두 크므로 단위근 가설을 기각하지 못한다. 따라서 모두 단위근 계열이다.

```
library(forecast)
aic = c()
for (p in 1:20){
  ar.fit = Arima(data$logy, order=c(p,0,0), method="ML")
  aic[p] = ar.fit$aic
}
which.min(aic)
library(fUnitRoots)
adfTest(data$logy, type="c", lags=14)

aic = c()
for (p in 1:10){
  ar.fit = Arima(data$logx, order=c(p,0,0))
  aic[p] = ar.fit$aic
}
which.min(aic) #1
adfTest(data$logx, type="c", lags=0)

aic = c()
for (p in 1:10){
  ar.fit = Arima(data$logz, order=c(p,0,0))
  aic[p] = ar.fit$aic
}
which.min(aic) #6
adfTest(data$logz, type="ct", lags=5)
```

(5) 변수 X, Y, Z에 대한 공적분 rank를 구하시오.

Values of teststatistic and critical values of test:

	test	10pct	5pct	1pct
r <= 2 |	6.56	7.52	9.24	12.97
r <= 1 |	13.68	13.75	15.67	20.20
r = 0 |	106.58	19.77	22.00	26.81

```
library(urca)
Data = data.frame(data$logy, data$logx, data$logz)
johanson.test = ca.jo(Data, type="eigen", ecdet="const")
summary(johanson.test)
```

귀무가설 r=0은 기각되지만 귀무가설 r<=1은 기각되지 않으므로 r=1이고 1개의 공적분 관계가 있다.

C. 위 B의 세 모형 (ARIMA 모형, ADL 모형, VAR(또는 VEC) 모형의 Y_{T+1}의 예측력을 비교해 보시오.

(1) ARIMA 모형의 예측오차 : 0.05829693

```
y.arima.1 = forecast(arima.fit, h=1)$mean[1]
e.arima.1 = log(gasdata[[267,3]])-y.arima.1
e.arima.1
```

(2) ADL 모형의 예측오차 : 0.0549614

```
beta = adl.fit$coef
N = length(dy0)
y.hat=c()
dy.hat=c()
e=c()
dy.hat[1] = beta[1] + beta[2]*dy0[N] + beta[3]*dy1[N] + beta[4]*dx0[N] + beta[5]*dx1[N] + beta[6]*dz0[N] + beta[7]*dz1[N]
y.hat[1] = log(gasdata[[266,3]])+dy.hat[1]
e.adl.1 = log(gasdata[[267,3]]) - y.hat[1]
e.adl.1
```

(3) VEC 모형의 예측오차 : 0.1038342

```
library(tsDyn)
bic = c()
for (p in 1:10){
  vecm.fit <- VECM(Data, lag=p, r=1, estim="ML", include="const")
  bic[p] = summary(vecm.fit)$bic
}
which.min(bic) #1

vecm.fit = VECM(Data, lag=1, r=1, estim="ML", include="const")
vecm.hat = predict(vecm.fit, n.ahead=1)
e.vecm.1 = log(gasdata[[267,3]])-vecm.hat[[1,1]]
e.vecm.1
```

모형별 예측오차

ARIMA(2,1,3) 모형	ADL(2,2) 모형	VEC 모형
0.05829693	0.0549614	0.1038342

ARIMA 모형, ADL 모형, VEC 모형으로 예측오차를 비교해본 결과 ADL 모형의 예측오차가 가장 작았다. 따라서 ADL 모형이 가장 잘 예측했다고 말할 수 있어 ADL 모형으로 미래의 값을 예측하는 것이 좋다고 할 수 있다.