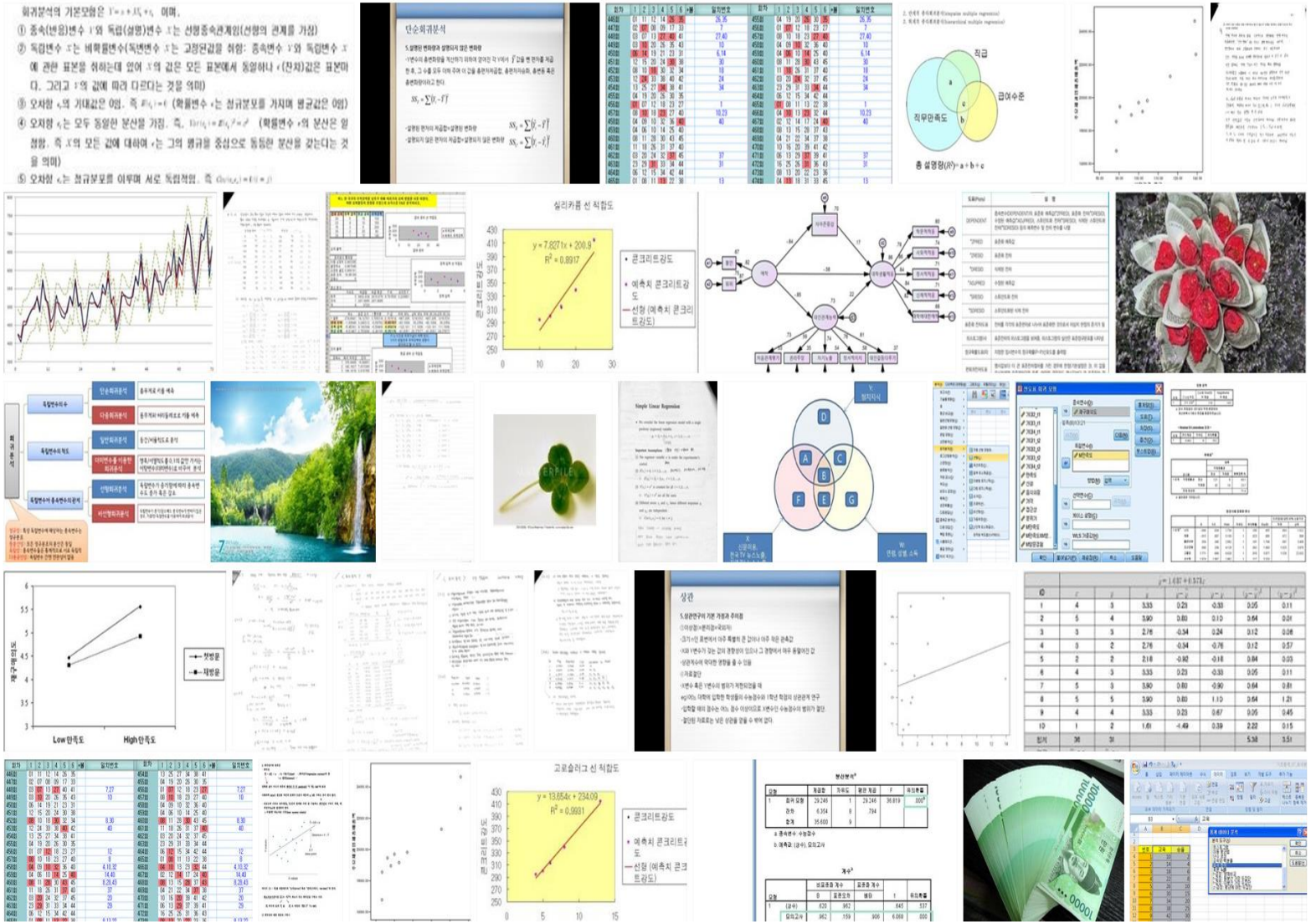
통계 알고리즘 8

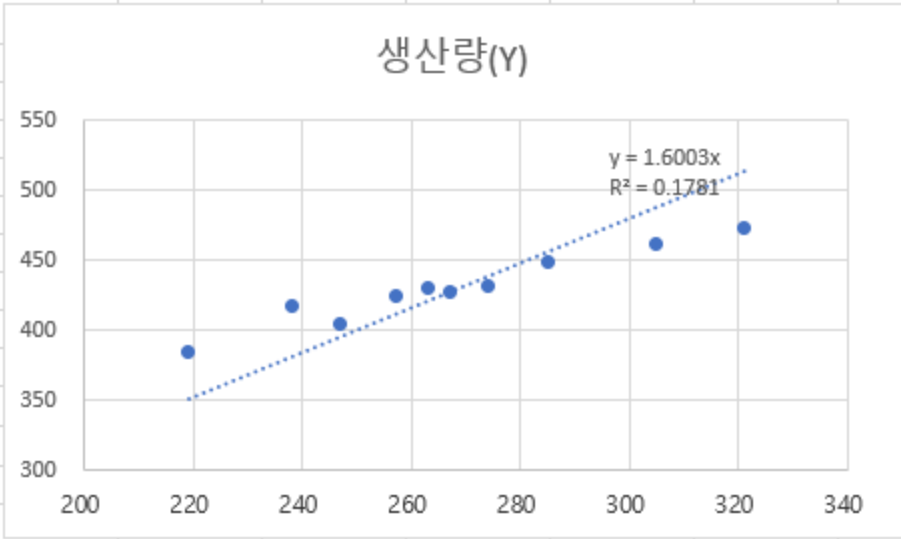
2019년 5월 정화민 교수

Regression (통계의 꽃)



단순 회귀식?
노동력이 350이면 생산량? 예측정확도 ?

번호	노동력(X)	생산량(Y)	X^2	XY	추정치	잔차
1	267	428	71,289	114,276	430	-2
2	263	430	69,169	113,090	427	3
3	238	417	56,644	99,246	406	11
4	219	384	47,961	84,096	390	-6
5	274	432	75,076	118,368	436	-4
6	257	425	66,049	109,225	422	3
7	321	474	103,041	152,154	476	-2
8	305	462	93,025	140,910	462	0
9	285	449	81,225	127,965	445	4
10	247	405	61,009	100,035	413	-8
합계	2,676	4,306	724,488	1,159,365		



Y 분산	626.0400	←	=VARP(C2:C11)
추정치분산	597.3244	←	=VARP(F2:F11)
잔차분산	28.7156	←	=VARP(G2:G11)
R^2	0.9519	←	=1 −(E19/E17)

N =	10
a =	204.8125
b =	0.8438

Mutiple Regression

함수는 특정한 작업을 수행하기 위해 일련의 구문들을 체계적으로 묶은 것, R은 수많은 내장 함수를 가지고 있음.(4주차 데이터 이용)

사용 Data : R 내장 데이터 mtcars:

1974년 미국 Motor trend US magazine에 나오는 data이다.
32개의 차량에 대해서 각 자료들을 기재되어 있음

mpg - Miles/gallon (연비, 1갤런당 몇 마일을 가는가)
cyl - Number of cylinders (차량 엔진의 실린더의 개수, 펌프같이 움직이는 것)
disp - Displacement (배기량)
hp - Gross horsepower (마력)
drat - Rear axle ratio (후방 축 비율
wt - Weight (1000lbs) 파운드 기준 차량무게
qseq - 1/4 mile time 1/4 마일 간 시간?
vs - V/S ??? versus??
am - Transmission(0 = automatic, 1 = manual) 변속기가 자동이냐 아니냐

gear - Number of forward gears 전진기어의 수? (1,2,3)
carb - Number of carburetors 카뷰레이터 수 (기화기수)

Mutiple Regression

연구문제 : 다중회귀분석을 이용하여 주행연비 계산하기 ?
회귀식 만들기 ? y = a + b1x1 + b2x2 +...bnxn
종속변수: 연비(mpg), 독립변수: 배기량(disp), 마력(hp), 무게(wt)

```
input <- mtcars[,c("mpg","disp","hp","wt")]
print(head(input))
```

```
                   mpg disp  hp    wt
Mazda RX4         21.0  160 110 2.620
Mazda RX4 Wag     21.0  160 110 2.875
Datsun 710        22.8  108  93 2.320
Hornet 4 Drive    21.4  258 110 3.215
Hornet Sportabout 18.7  360 175 3.440
Valiant           18.1  225 105 3.460
```

```
input <- mtcars[,c("mpg","disp","hp","wt")]
 model <- lm(mpg~disp+hp+wt, data = input)
print(model)
```

```
Call:
lm(formula = mpg ~ disp + hp + wt, data = input)

Coefficients:
(Intercept)         disp           hp           wt
  37.105505    -0.000937    -0.031157    -3.800891
```

y = a + b1x1 + b2x2 +...bnxn

아래 공식을 완성하세요

Y = 37.10 +(-0.000937*200)+
(-0.031157*120)+(-3.800891*2.91)

문제 disp =200, hp=120, wt =2.91 인 자동차의 예상주행거리는 얼마인가 ?

Logistic Regression

"mtcars"데이터 세트에서 자동차 이진 값 (0 또는 1) 인 am을 종속변수 hp, wt 및 cyl를 독립변수로한 logistic regression

```
input <- mtcars[,c("am","cyl","hp","wt")]
print(head(input))
```

```
                  am cyl  hp    wt
Mazda RX4          1   6 110 2.620
Mazda RX4 Wag      1   6 110 2.875
Datsun 710         1   4  93 2.320
Hornet 4 Drive     0   6 110 3.215
Hornet Sportabout  0   8 175 3.440
Valiant            0   6 105 3.460
```

```
input <- mtcars[,c("am","cyl","hp","wt")]
am.data = glm(formula = am ~ cyl + hp + wt, data = input, family = binomial)
print(summary(am.data))
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 19.70288    8.11637   2.428   0.0152 *
cyl          0.48760    1.07162   0.455   0.6491
hp           0.03259    0.01886   1.728   0.0840 .
wt          -9.14947    4.15332  -2.203   0.0276 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ARIMA

ARIMA (Auto-regressive Integrated Moving Average) 모델: 과거의 관측값과 오차를 사용해서 현재의 시계열 값을 설명하는 BOX-JENKINS 모델(ARMA)을 일반화한 것으로 시계열 데이터의 과거 치들이 설명변수인 AR과 과거의 오차항들이 설명변수인 MA 모델의 합성어이다.

ARIMA 모델

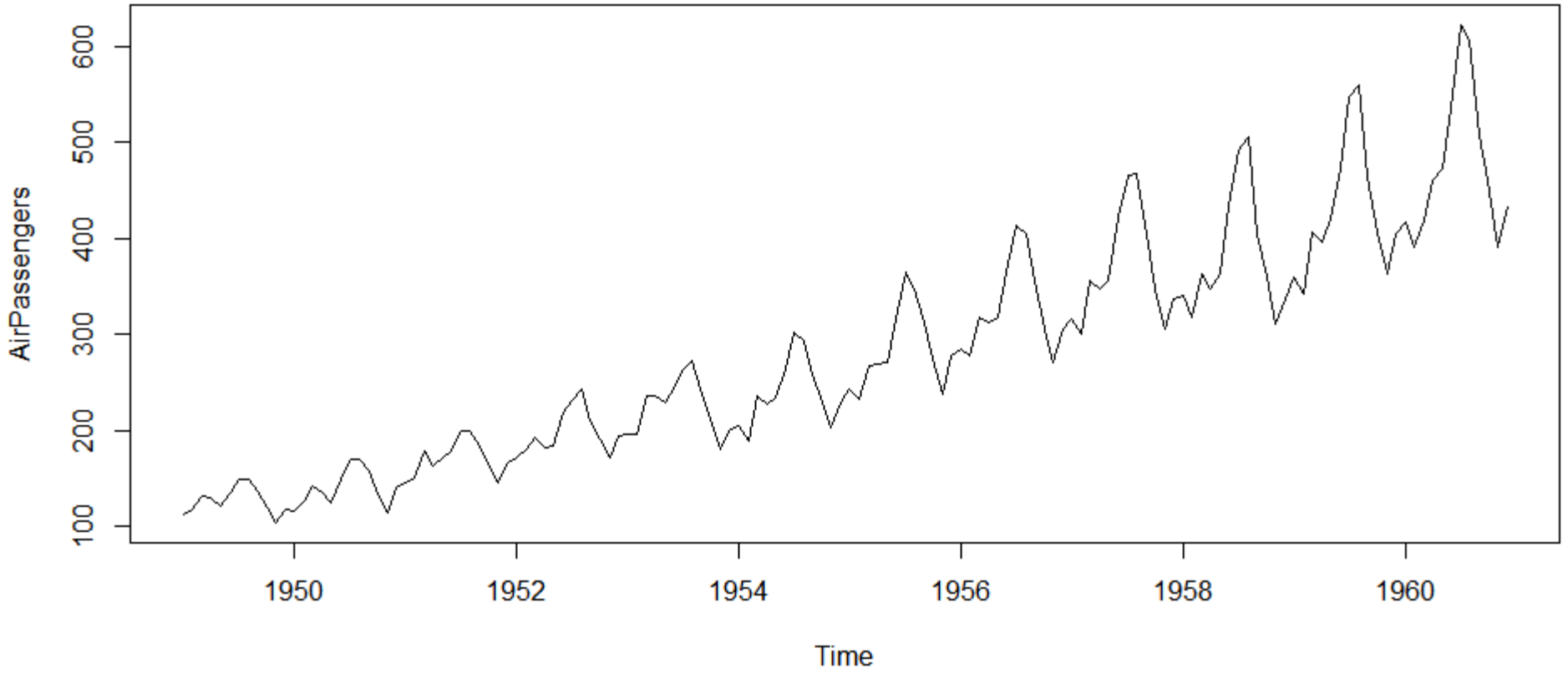
데이터는 AirPassengers 데이터로 1949년부터 1960년 사이에 매월 항공기 탑승승객수를 나타낸 데이터 세트.

```
> data(AirPassengers)
> AirPassengers
     Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1949 112 118 132 129 121 135 148 148 136 119 104 118
1950 115 126 141 135 125 149 170 170 158 133 114 140
1951 145 150 178 163 172 178 199 199 184 162 146 166
1952 171 180 193 181 183 218 230 242 209 191 172 194
1953 196 196 236 235 229 243 264 272 237 211 180 201
1954 204 188 235 227 234 264 302 293 259 229 203 229
1955 242 233 267 269 270 315 364 347 312 274 237 278
1956 284 277 317 313 318 374 413 405 355 306 271 306
1957 315 301 356 348 355 422 465 467 404 347 305 336
1958 340 318 362 348 363 435 491 505 404 359 310 337
1959 360 342 406 396 420 472 548 559 463 407 362 405
1960 417 391 419 461 472 535 622 606 508 461 390 432
```

ARIMA

```
5  plot(AirPassengers)
6
```

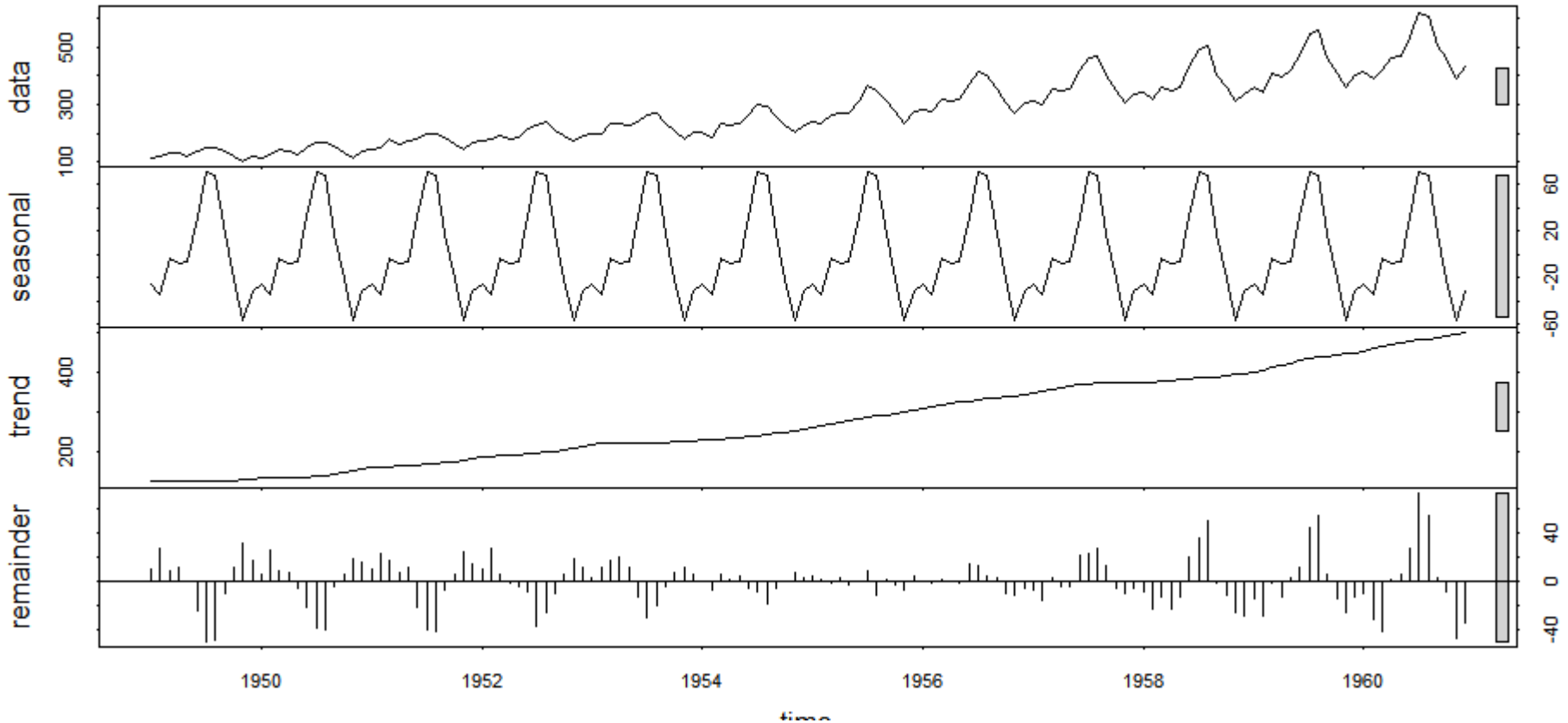
AirPassengers 시계열 데이터에 대해 시간에 따른
항공기 탑승승객 수에 대한 플롯 (plot() 함수 수행 결과



ARIMA

계절성(seasonality), 추세 (trend), 불확실성 (random) 요소로 분해한 그래프

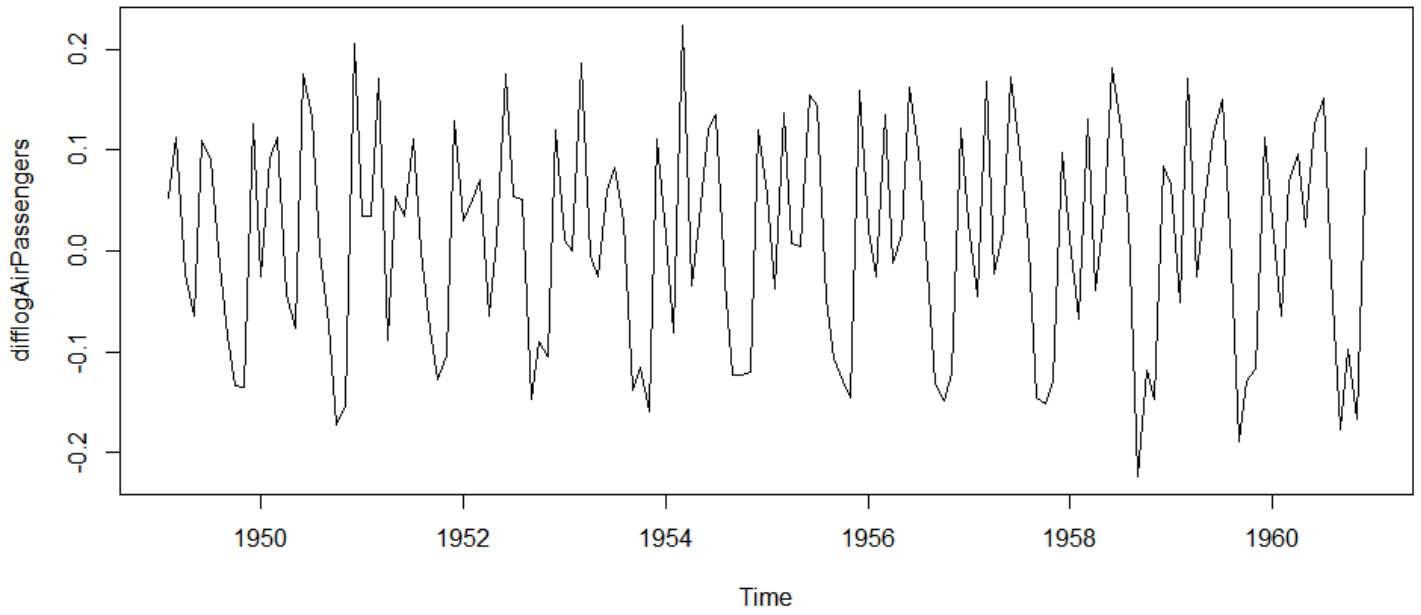
```
plot(stl(AirPassengers, s.window='periodic'))
```



ARIMA

이상의 시계열 데이터를 diff()함수와 log()함수 등을 활용하여 안정적인 시계열 데이터로 변환

```
install.packages("tseries")
library(tseries)
difflogAirPassengers <- diff(log(AirPassengers))
plot(difflogAirPassengers)
```



AirPassengers 시계열 데이터에 diff 및 log함수를 적용한 데이터 플롯

ARIMA

```
install.packages("forecast")
library(forecast)
auto.arima(difflogAirPassengers)

#구한 파라미터를 기준으로 log(AirPassengers) 데이터 세트를 대상으로 하는 ARIMA 모델을
생성한다.

fitted <- arima(log(AirPassengers), c(1, 0, 1), seasonal =list(order = c(0, 1, 1), period =
12))
fitted

#ARIMA 모델을 생성한 뒤 향후 10년(120개월)간의 데이터를 예측해본 후 (predicted 변수에 저장됨
), 기존 데이터 (AirPassengers)와 예측치를 이어서 그래프로 표현한다

predicted <- predict(fitted, n.ahead = 120)
ts.plot(AirPassengers, exp(predicted$pred), lty = c(1,2))
#predicted$pred 항목에 log(AirPassengers)의 예측치 값이 저장
```