김광일, 권태윤, 김동욱, 정해원, 허필재, 김태민

1. 프로젝트 배경과 주제

해결할 문제 정의 : 인구고령화가 심화되고 있다. 이에 대해서 일본 주택시작과 비교하고 한국 주택시장에 대한 전망을 예측하고자 한다.

해당 주제를 선정하게 된 이유. 왜 이 문제가 중요한가?

대한민국 사람이라면 누구나 고민하게 되는 재테크의 수단이 부동산 시장이다. 현재 정부정책으로 주택시장이 하락하고있다. 부동산 재테크 의사결정보조도구로서 주택가격을 추정하고자한다.

팀원 역할분담 소개 (임의로 넣어봤습니다.)

A : 시각화분석, 주택가격 예측모델 생성 (인구감소 예상값 포함)
B : 회귀분석, 발표자료 작성, 발표
C : 두집단 비교분석 (한국,일본)
D : 회귀분석, 잔차검증 분석
E : 가설기획, 시각화분석
F : 데이터 수집, 가설 기획

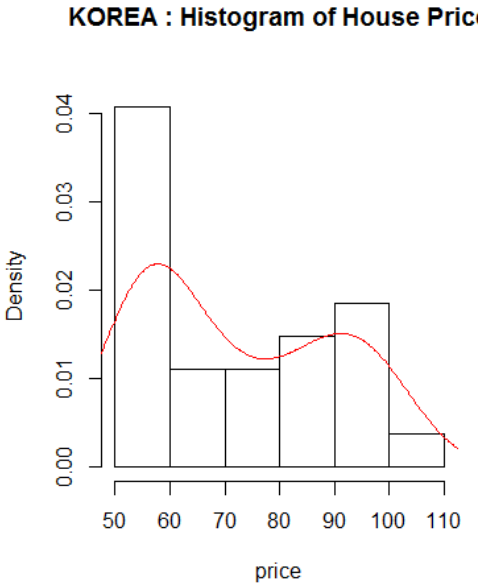
2. 데이터 소개와 변수 정의

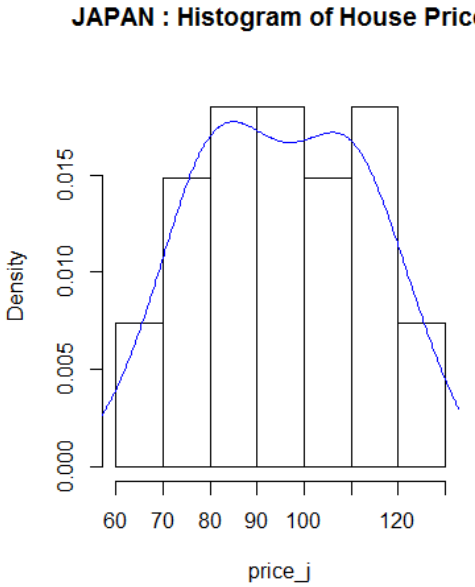
국가통계포탈 : http://kosis.kr (인구주택총조사)
DBpia : 한국은행 금융안정국 연구자료

변수 정의 :
price : 주택가격(%), 2015년도 평균 통계치를 100%기준으로 환산한 평균가격 (일본은 2000 년도 평균 통계치를 100% 기준으로함.)
population : 생산가능인구비중(%), 15~64세 범위에 속하는 인구수의 비율
apt : 아파트구성비중(%), 주택유형별 구성에서 아파트가 차지하는 비율
own : 자가비중(%), 점유형태별 유형중 자가(自家)에 속한 비율 (인구주택총조사)
rent : 임대차시장내월세비중(%), 임대점유형태 중 전세를 제외한 나머지 비율

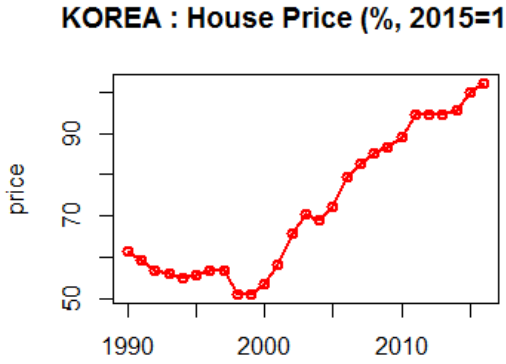
3.데이터 탐색 및 분석

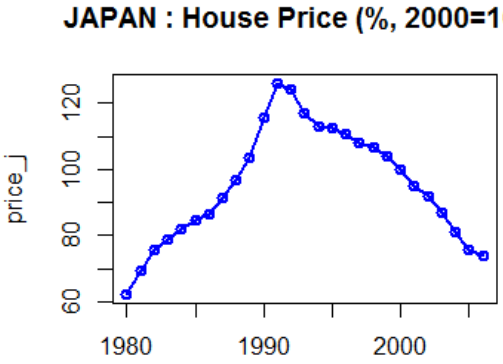
A. 일본과 한국 주택시장, 두집단에 대한 차이 분석

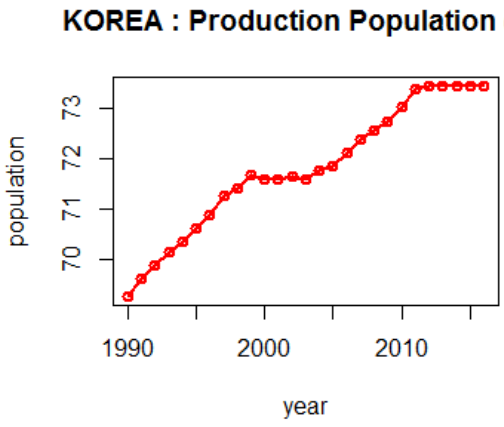


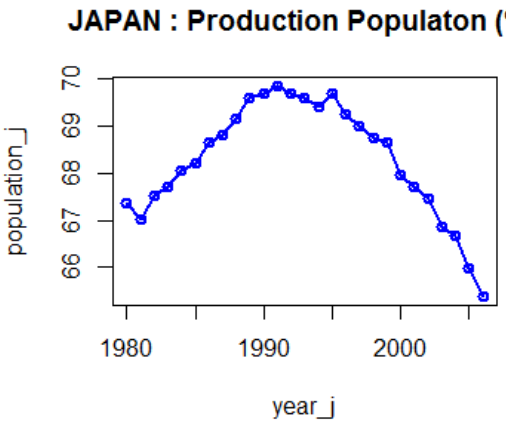


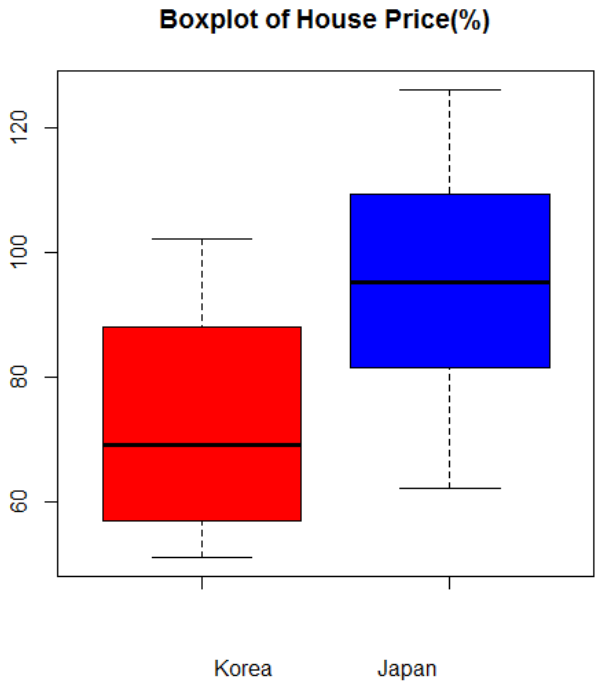
종속변수(한국 주택가격)에 대한 정규분포 가정이 위배된다.
데이터 특성상 모수를 증가할 수 없고, log 변환을 하여도 정규분포가 되지 않기 때문에 최종 회귀모형을 통한 분석 신뢰성이 떨어질 수 있다.

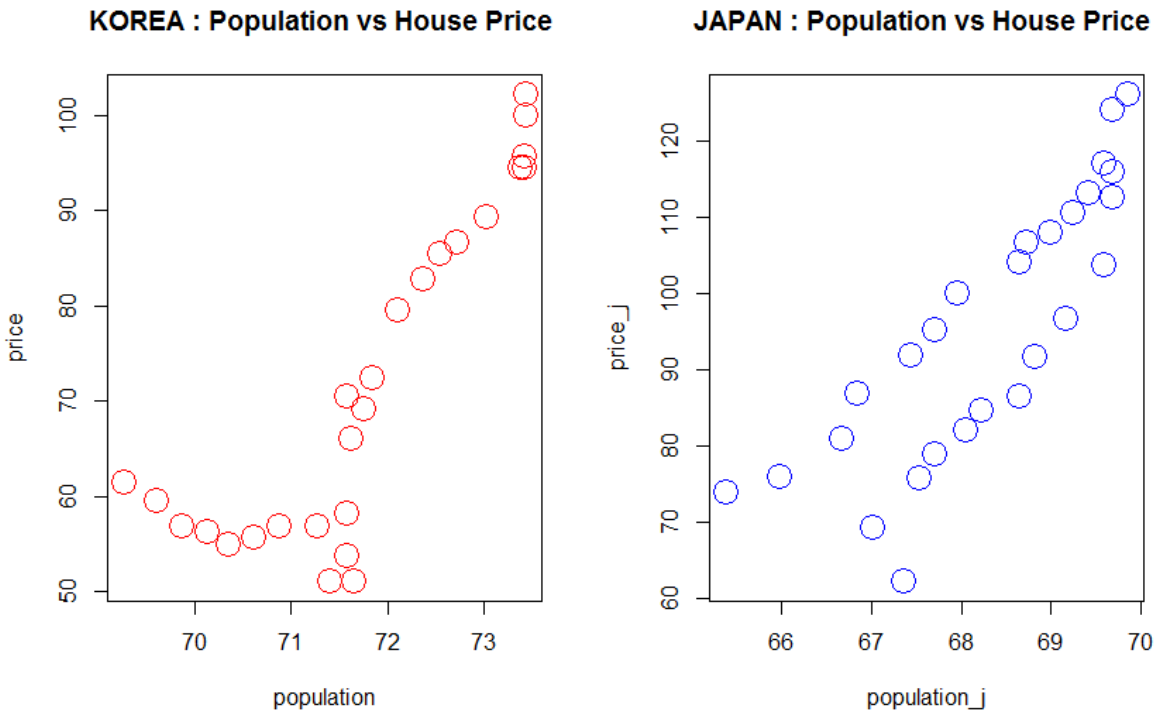








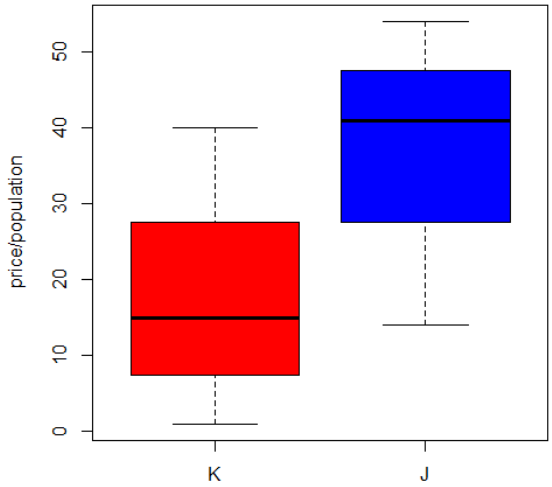




일본의 주택가격의 경우 생산인구비중에 따라 주택가격이 강한 선형관계를 보인다.
하지만 한국의 주택가격의 경우 비선형의 관계가 보인다.

한국의 주택시장과 일본의 주택시장이 동일한 집단의 특성을 가진다면
생산인구비중의 변화에 따라 한국의 주택가격이 일본과 동일한 패턴으로
인구노령화에 따라 급격히 주택가격이 하락할 것인지를 유추할 수 있다.

이를 위해서 "생산인구비중에 대한 주택가격"이라는 변수를 생성해서
두 국가의 주택시장 모집단이 차이가 있는지를 two sample t-test 를 하고자 한다.
(※ 생산인구비중에 대한 주택가격 = 주택가격 / 생산인구비중)
정규분포를 따르지 않더라도 두 집단의 과측수가 54 이므로 (n1+n2>30)
일반적으로 two sample t-test 를 사용할 수 있다.



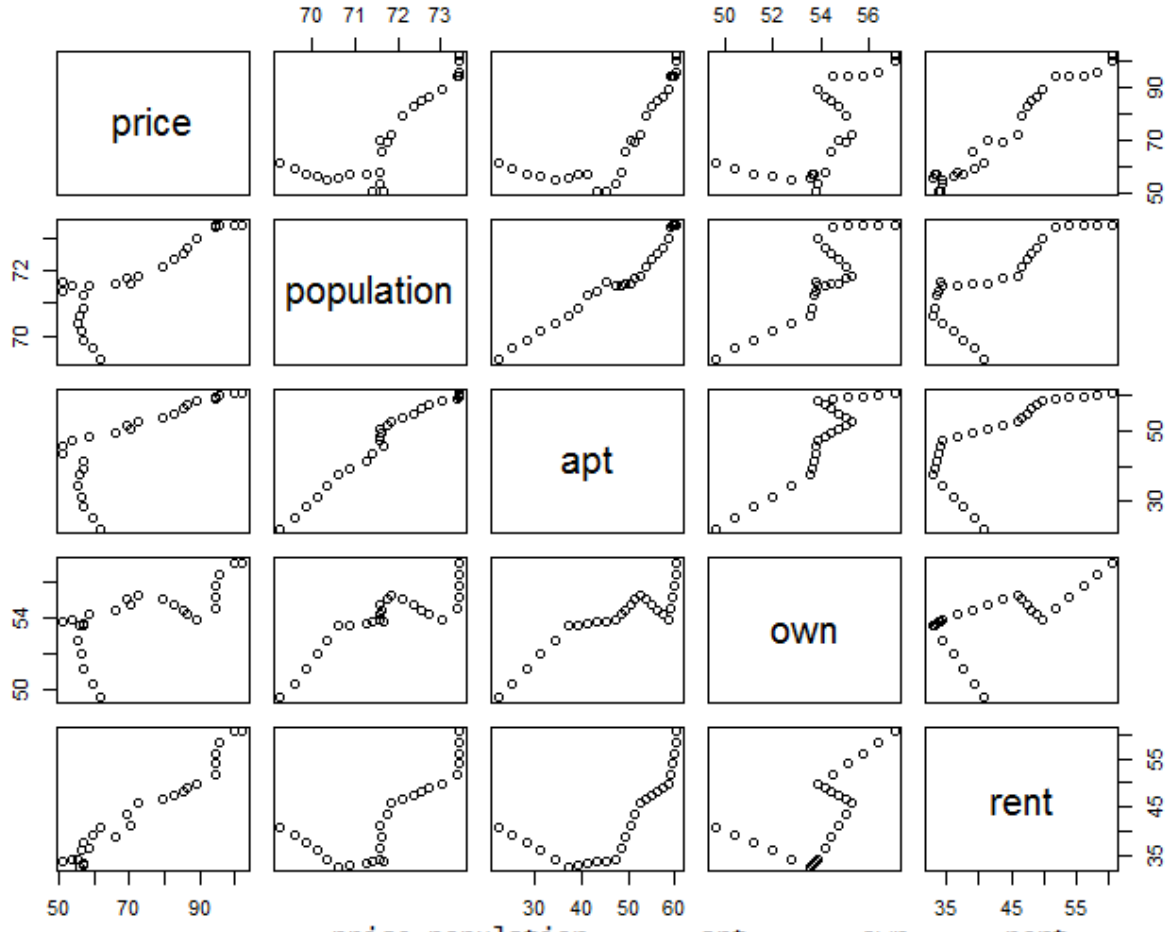

```
> var.test(p~g,data=n)

        F test to compare two variances

data:  p by g
F = 1.0271, num df = 26, denom df = 26, p-value = 0.9462
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4680782 2.2537868
sample estimates:
ratio of variances
          1.027107
```

p-value >0.05, 두 집단의 분산이 같다는 결과이므로, t.test(, var.equal=TRUE) 진행한다.

```
> t.test(p~g,data=n,var.equal=TRUE)

        Two Sample t-test

data:  p by g
t = -5.9754, df = 52, p-value = 2.106e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.56799 -13.20979
sample estimates:
mean in group K mean in group J
       17.55556        37.44444
```

p-value < 0.05 이므로 두 집단 모평균은 다르다.
한국의 주택시장과 일본의 주택시장이 동일한 집단은 동일한 특성을 가지지
않기 때문에 일본처럼 인구 고령화에 따라 급격하게 주택가격이 하락한다고 할 수 없다.

B. 한국 주택가격을 시뮬레이션/추정 (다중회귀분석 모형)



	price	population	apt	own	rent
price	1.0000000	0.8576903	0.8022654	0.6791088	0.9774744
population	0.8576903	1.0000000	0.9766257	0.8797401	0.8084991
apt	0.8022654	0.9766257	1.0000000	0.9092230	0.7443079
own	0.6791088	0.8797401	0.9092230	1.0000000	0.6565795
rent	0.9774744	0.8084991	0.7443079	0.6565795	1.0000000

변수간의 산점도와 상관계수 표를 통해서 선형관계를 판단하였다.
종속변수는 price, 나머지는 독립변수로서 회귀분석을 진행하였다. (아래는 변수 설명)

price : 주택가격(%), 2015년도 평균 통계치를 100%기준으로 환산한 평균가격 (일본은 2000 년도 평균 통계치를 100% 기준으로함.)
population : 생산가능인구비중(%), 15~64세 범위에 속하는 인구수의 비율
apt : 아파트구성비중(%), 주택유형별 구성에서 아파트가 차지하는 비율
own : 자가비중(%), 점유형태별 유형중 자가(自家)에 속한 비율 (인구주택총조사)
rent : 임대차시장내월세비중(%), 임대점유형태 중 전세를 제외한 나머지 비율

population 과 apt 간의 변수가 매우 상호관계가 높아서 다중 공선성의 문제가 의심된다.
시각화를 통해 price 와 rent 간의 선형관계가 매우 강함을 확인할 수 있다.

```
> c<-lm(price~.,data=a)
> summary(c)

Call:
lm(formula = price ~ ., data = a)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3965 -1.6002 -0.2167  1.7632  3.3056


> vif(c)
population        apt        own       rent
 32.095156  31.430223   5.822674   3.314079
```

분산팽창인수(VIF; variance inflation factor) 는 가능한 가장 작은 값=1이며, 공선성이 전혀 없는 것을 나타낸다. VIF가 5 또는 10을 초과하면 공선성 존재로 판단하는데
population 과 apt 변수에 대한 VIF 지수가 높기 때문에 apt 변수를 제거한다.

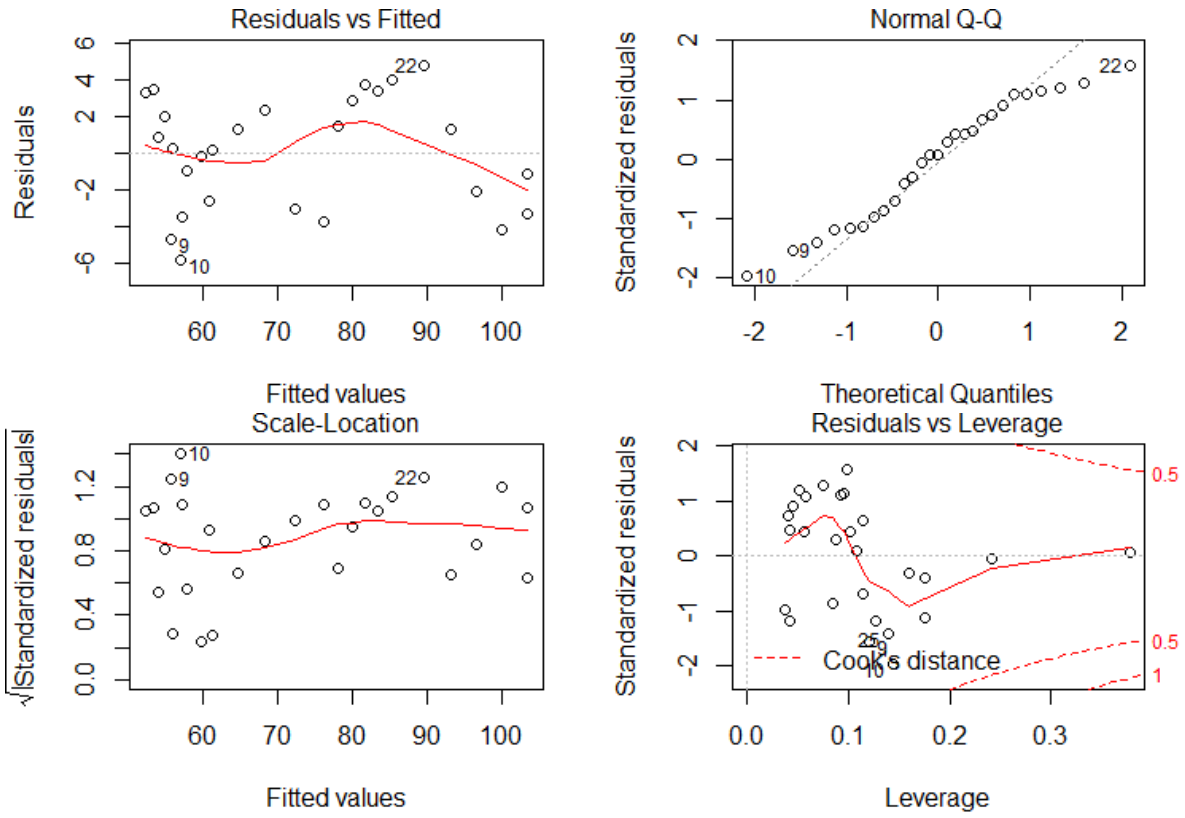
```
> d <-subset(b,select=-c(apt))
> c2<-lm(price~.,data=d)
> summary(c2)

Call:
lm(formula = price ~ ., data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3638 -1.8835  0.2214  2.1540  4.4576

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -250.8009    62.9182  -3.986 0.000582 ***
population     4.6154     1.2953   3.563 0.001653 **
own           -1.3846     0.7249  -1.910 0.068686 .
rent           1.5300     0.1141  13.410 2.33e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.038 on 23 degrees of freedom
Multiple R-squared:  0.9729,    Adjusted R-squared:  0.9693
F-statistic:   275 on 3 and 23 DF,  p-value: < 2.2e-16
```

H0 : X와 Y가 상관관계가 없다.
H1 : X와 Y가 상관관계가 있다.
95% 신뢰구간에서 p-value 가 0.05 보다 작기 때문에 은 population, rent 두 지표가 귀무가설을 기각하고 유의미하다. 즉, 상관관계가 있다.

상기 분석결과를 가지고, 두 지표만을 고려하여 새로운 다중회귀식을 생성하였다.

```
> c3<-lm(price~population+rent, data=d)
> summary(c3)

Call:
lm(formula = price ~ population + rent, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8430 -2.8574  0.2453  2.5827  4.8193

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -187.9357    56.5030  -3.326  0.00283 **
population     2.6708     0.8438   3.165  0.00418 **
rent           1.5726     0.1179  13.338 1.36e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.201 on 24 degrees of freedom
Multiple R-squared:  0.9686,    Adjusted R-squared:  0.966
F-statistic: 369.9 on 2 and 24 DF,  p-value: < 2.2e-16
```

모델의 유의성 평가인 F검정값을 보면 F=1이 아니기 때문에((F=369.6 , p< 0.05) 적어도 하나의 변수는 주택가격인 price 변수를 설명하는데 상관관계가 있다. (F=570 이면 p ≈ 0)
모든 계수의 p-value 는 0.05 이하이므로 상기 모델의 변수들은 유의미하다.



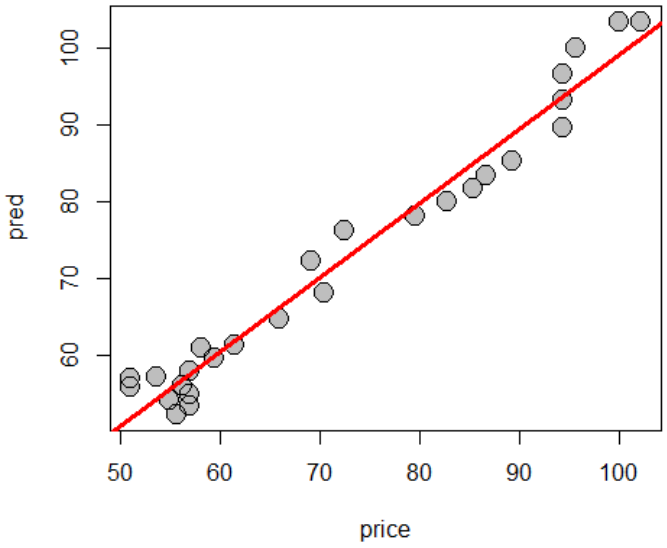
5. 결론 도출

"생산인구비중에 대한 주택가격"이라는 변수를 생성해서
일본과 한국의 주택시장 모집단이 차이가 있는지를 two sample t-test 를 하고, ,
두 집단 모평균이 다르다고 결론지었다.

한국의 주택시장과 일본의 주택시장이 동일한 집단은 동일한 특성을 가지지
않기 때문에 일본처럼 인구 고령화에 따라 급격하게 주택가격이 하락한다고 할 수 없다.

한국 주택가격을 추정하는 최종 모델은 아래와 같다.

Price = 187.9357 + 2.6708 * population +1.5726 * rent



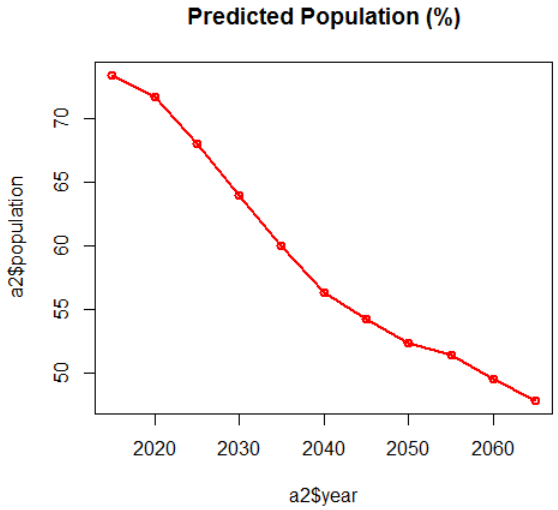
X축 : 한국 주택가격, Y축 : 다중회귀식으로 계산한 주택가격

모델에 대한 성능을 확인하기 위해서 모델값과 실제값을 산점도로 fitting 하였다.
생산인구비중(population)과 임대차시장내 월세비중(rent) 변수로 생성된 주택가격(price)을
시뮬레이션 하는 모델은 R^2= 0.9686 로 약 97% 주택 가격을 설명할 수 있다.

6. 부 록

2016년 12월 24일 연합뉴스(이재윤 기자)가 작성한 통계청 출처의 생산가능인구 정망 data를 Plot 하면 아래와 같다.

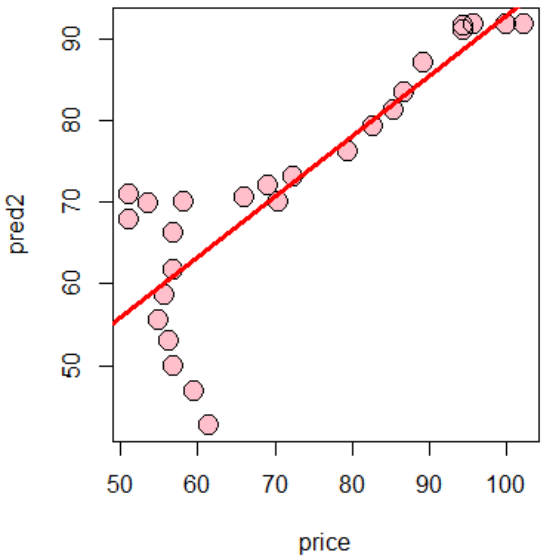
```
# prediction
a2<-read.csv("population_future.csv")
plot(a2$year,a2$population,main="Predicted Population (%)",type="o", lwd=2, col="red")
```

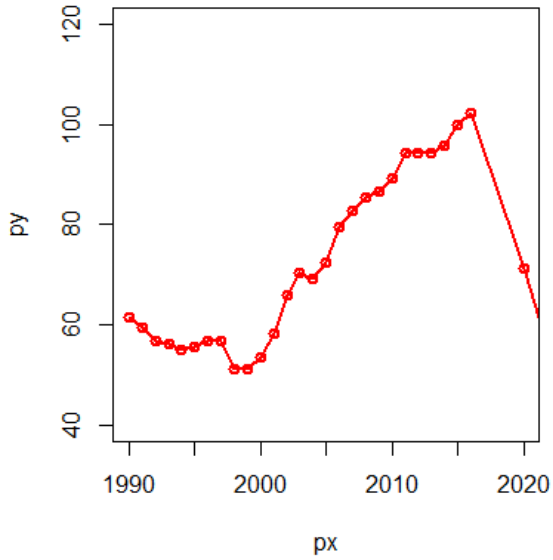


이 데이터를 통해 (한가지 변수만 고려함, 생산가능인구비중) 선형회귀식을 생성하고, 미래 추정치의 생산가능 인구비중만을 가지고 추정값을 도출하고자 한다.

```
> c4<-lm(price~population, data=d)
> summary(c4)

Call:
lm(formula = price ~ population, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-19.796  -3.765   2.803   3.772  18.665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -772.574    101.325  -7.625 5.58e-08 ***
population    11.770      1.411   8.341 1.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.097 on 25 degrees of freedom
Multiple R-squared:  0.7356,    Adjusted R-squared:  0.7251
F-statistic: 69.57 on 1 and 25 DF,  p-value: 1.09e-08
```

새로 생성된 회귀식은 R^2= 0.7356 으로 약 73% 주택 가격을 설명할 수 있다





좌 그림 : y축 – 추정 주택가격 , x축 – 실제 주택가격
우 그림 : y축 – 주택가격 , x축 – 년도 (※ 2017년 이후 추정치 fitting)

모델에 대한 학습데이터 fitting 산점도를 보면 주택가격이 70% 이하에서는 비선형적이며, 잔차가 커지는 것을 알 수 있다. 따라서 위 모형은 주택가격이 70% 이상 범위 내에서만 추정하는 것이 올바르다고 판단된다.

따라서 우 그림에서 나타난 시계열 주택가격 그래프에서 2020년도 이후의 주택가격 추정치는 제외하였다.

추정 주택가격은 한국의 평균 주택가격이며, 단일 변수만 고려된 낮은 차원의 회기식이므로 신뢰도가 매우 낮다. 하지만 일반적으로 일본 주택시장에서 미리 경험한 인구 고령화에 의한 주택가격평균치의 하락은 통계적 접근에서도 유사하다.

사회적 현상은 통계적으로 설명하는데 한계가 있기 때문에
상기 분석 결과는 통계적 관점에서 주택가격을 접근했다는데 의의를 두며, 분석 내용은 크게 신뢰성이 있다고 생각하지 않는다.

- 이 상 -