05. R 정형데이터 분석 01
평균 차이, 상관관계, 회귀와 예측

성현곤

충북대학교 도시공학과
Dept.of URBAN ENGINEERING

목차

평균차이 검정 방법론 개요

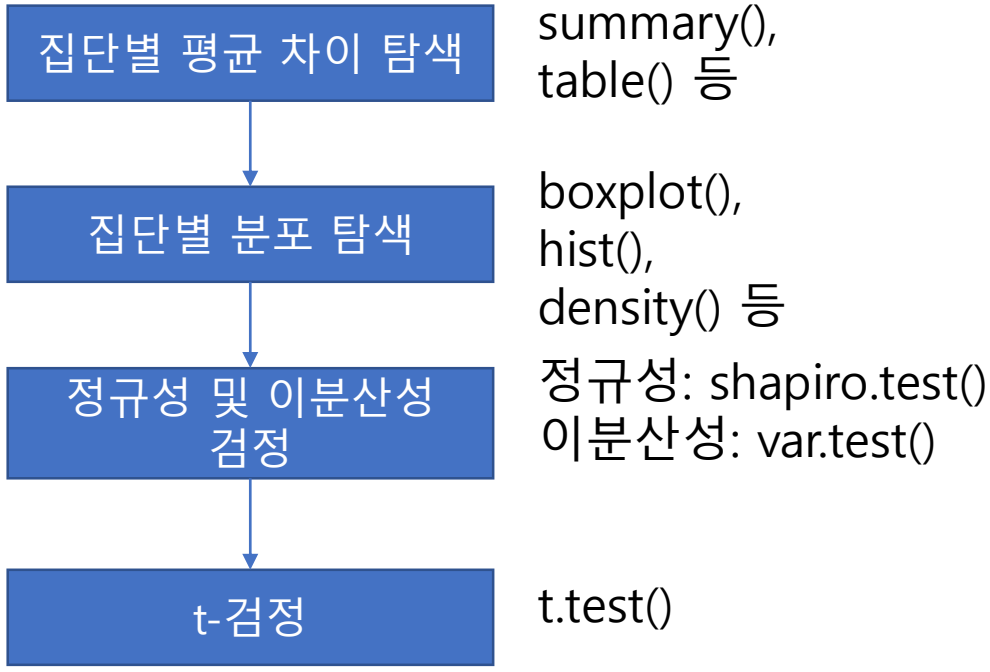
- 집단(표본)간 평균 차이 검정방법의 종류
  - 한 집단 또는 두 집단 차이 검정: t-test
  - 두 집단 이상의 차이 검정: ANOVA
- t-test 와 ANOVA의 비교

	t-test	ANOVA (ANalysis Of VAriance)
검정 방법	▪ 독립된 두 표본 평균(비율) 차이 검정 ▪ 대응(paired) 두 표본 평균(비율) 차이 검정 ▪ 한 집단 평균(비율) 차이 검정	▪ 서로 독립된 집단이 셋 이상인 경우의 평균의 차이 검정
가설 설정	▪ 귀무가설(H0): 평균의 차이가 동일함(보다 크거나 작음) ▪ 대립가설(Ha): 평균의 차이가 동일함(보다 크거나 작음)	▪ 귀무가설(H0): 세 집단간 평균의 차이가 없음 ▪ 대립가설(Ha): 적어도 하나 이상의 집단은 다른 집단과 평균의 차이가 있음
적용 예시	▪ 남자와 여자 간 소득의 차이 비교 ▪ 도시와 농촌지역의 주택가격은 차이가 있을까? ▪ 중간고사에 비하여 기말고사의 성적은 올랐을까? ▪ 과외를 하기 전과 후의 반 학생들의 성적 변화 ▪ 서울시의 집값 평균은 평당 1,000만원 보다 높을까?	▪ 계절(봄, 여름, 가을, 겨울)별 아파트 거래가격의 평균은 동일한가? ▪ 1, 2, 3, 4학년의 학생 성적의 평균의 차이는 없는가?

평균차이 검정 방법론 개요

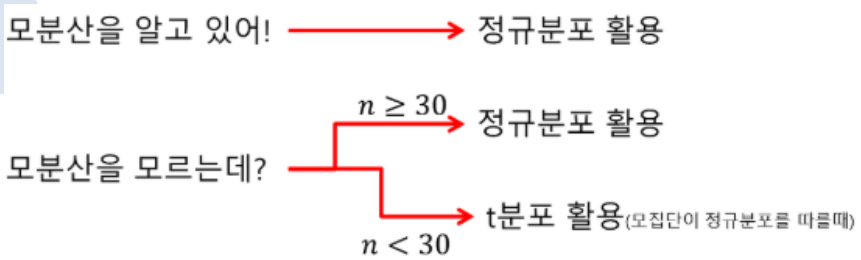
- t-test의 종류
  - 독립된 두 표본 평균(비율)의 차이
  - 쌍으로 된(대응) 두 표본 평균(비율)의 차이
  - 한 집단의 특정 평균값과의 차이
- t-test의 기본가정
  - 집단의 분포는 정규성과 등분산성을 충족
    - 중심극한 정리: 표본의 수가 일정 이상 보다 크면 (n>30) 정규성 검정 불필요
- t-test의 대립가설(Ha)
  - 집단간 차이는 있다.
  - A에 비하여 B집단의 평균이 크다.
  - A에 비하여 B집단의 평균이 작다.

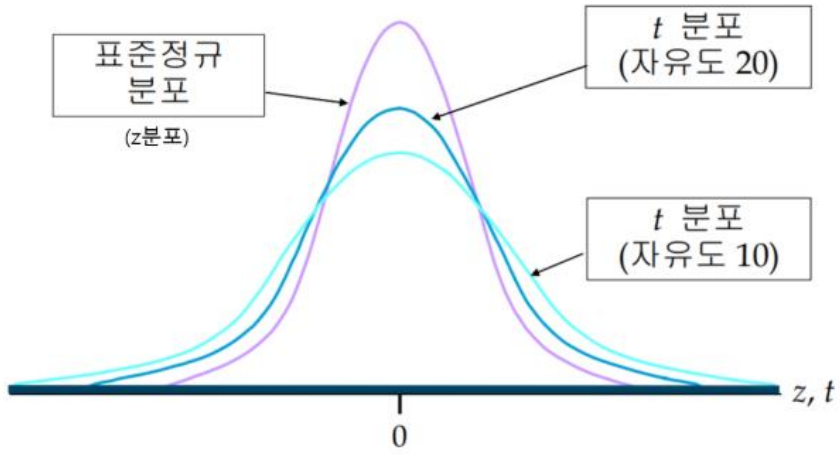
- t-test의 절차



평균차이 검정 방법론 개요

- t-분포와 신뢰구간
  - t-분포 vs. z-분포



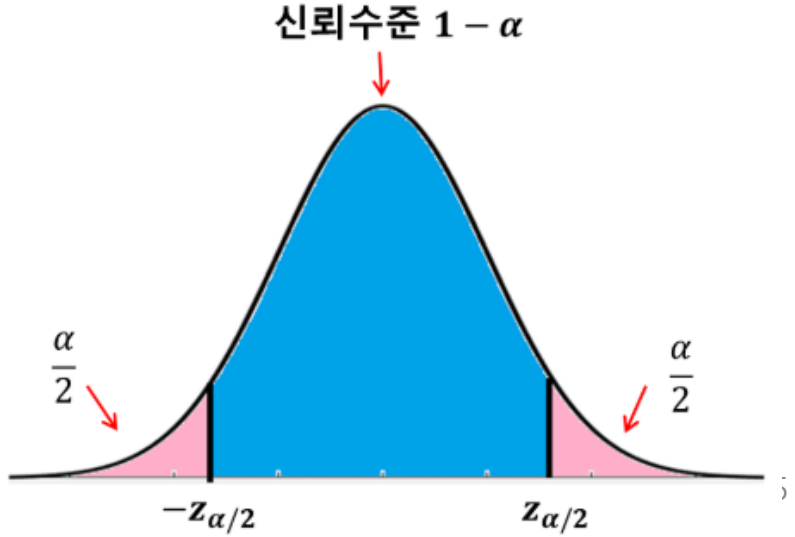


  - t-값(통계량)

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

  - 유의수준(a), 신뢰구간, p-value
    - 유의수준(a)= 0.05, 0.01, 0.001, 0.1 등 사용
      - 모수가 신뢰구간 안에 포함되지 않을 확률(a)
      - 모수가 신뢰구간에 포함될 확률(1-a)
    - P-value

$$P(a \leq \mu \leq b) = 1 - \alpha$$



평균차이 검정 방법론 개요

- t-test
  - 데이터가 서로 다른 두 모집단, 또는 하나의 모집단에서 반복 추출되었을 때의 평균의 차이 검정
  - T-test로부터의 분석결과의 해석은 t-값, p-값, 신뢰구간 등을 활용

집단	평균 ①	t-value ②	d.f	p-value ③	Conf. interval ④	
집단 A	674.3	4.94	9246.3	0.002	14.73	34.10
집단 B	649.9					

① 집단별 평균의 차이를 확인
② T-값이 높으면 통계적으로 평균의 차이가 유의할 확률이 높음
③ P-값이 0.05(0.01)보다 적으면, 95%(99%) 신뢰수준에서 통계적을 유의하다고 판단
④ 신뢰구간의 값이 0이 포함되어 있지 않으면, 또는 집단간 평균의 차이(=674.3-649.9)가 해당 신뢰구간에 있으면 통계적으로 유의한 차이가 있다고 판단

평균차이 검정 방법론 개요

- 분산분석(ANOVA: Analysis of Variance)
  - 서로 다른 집단(표본)이 3개 이상일 경우
  - 비교하고자 하는 값(종속변수)은 연속변수이고, 비교하고자 하는 집단은 범주형 자료(categorical data)또는 요인(factor)인 경우
  - 분산의 개념을 이용하여 분석
    - 분산을 계산할 때처럼 편차의 각각의 제곱합을 해당 자유도로 나누어서 얻게 되는 값을 이용하여 수준의 평균들간의 차이가 존재하는 지를 판단
- 분산분석의 종류
  - 비교하고자 하는 집단의 변수(설명변수)의 수가 개수에 따라 구분
    - 일원 분산분석(One-way ANOVA): 1개 비교 집단
    - 이원 분산분석 (Two-way ANOVA): 2개 비교 집단
    - 다원 분산분석 (N-way ANOVA): 3개 이상 비교 집단

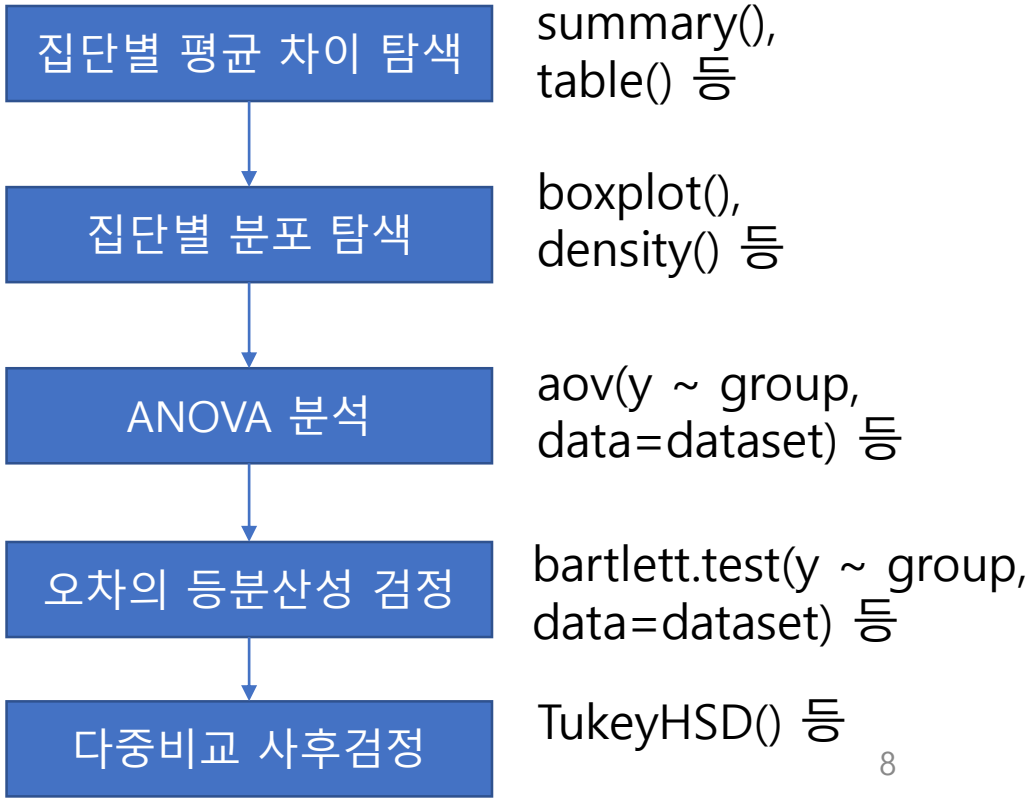
1 Factor	One-way ANOVA
2 Factors	Two-way ANOVA
n Factors	n-way ANOVA

평균차이 검정 방법론 개요

- 분산분석(ANOVA: Analysis of Variance)
  - 함수: aov()
  - 영향요인의 수에 따른 분산분석 모형

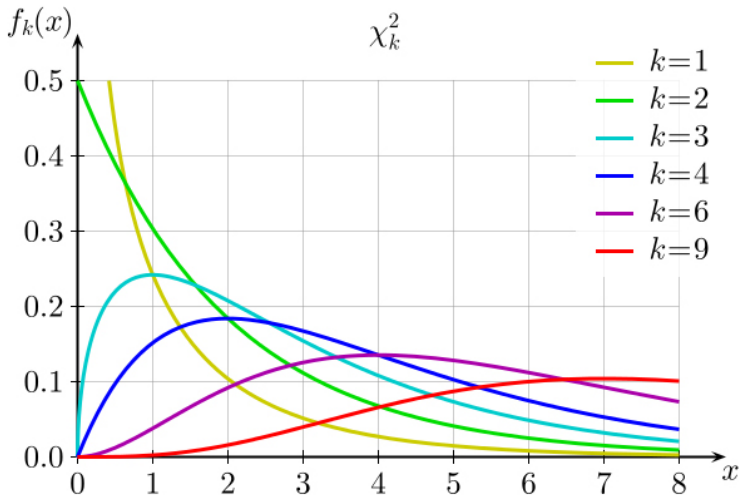
n-way ANOVA	Model	Description
one-way ANOVA	y ~ x1	y is explained by x1 only
two-way ANOVA	y ~ x1 + x2	y is explained by x1 and x2
two-way ANOVA	y ~ x1 * x2	y is explained by x1, x2 and the interaction between them
three-way ANOVA	y ~ x1 + x2 + x3	y is explained by x1, x2 and x3

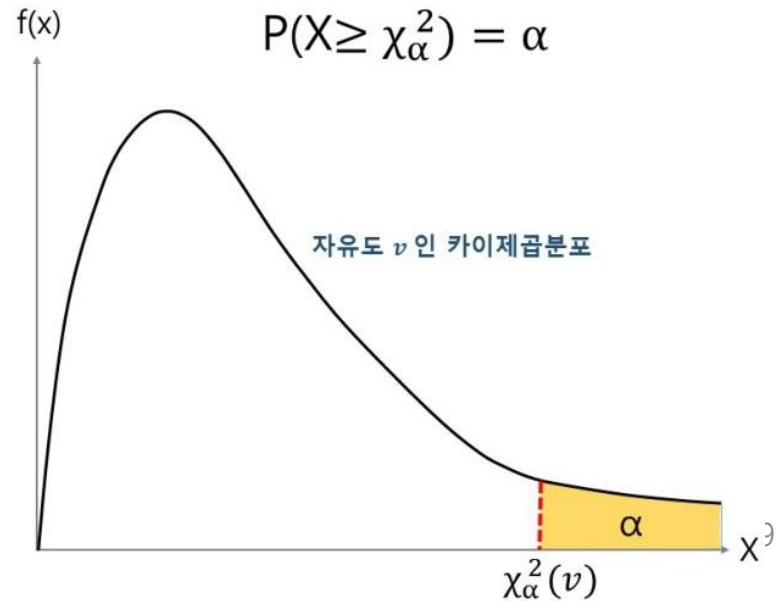
- 분산분석의 절차



평균차이 검정 방법론 개요

- F-분포(F-Distribution)
  - 정규분포의 모집단에서 독립적으로 추출한 표본들의 분산비율이 나타내는 연속적인 확률분포
    - k= 자유도(degree of freedom)
- 적용분야
  - 두 개 이상의 표본집단의 분산의 비교 또는 모집단의 분산의 추정시 활용
  - 범주형 자료에 대한 적합도 검정 및 복수 집단의 독립성 검정 등에 활용





평균차이 검정 방법론 개요

- 일원분산분석
  - 분석결과의 해석 방법은 f-값, p-값, 신뢰구간 등을 활용

집단	Obs.(샘플수)	평균①	분산
집단 A	3419	685	72004
집단 B	3004	658	72160
집단 C	3510	643	74187
집단 D	3376	678	77850

Df = degree of freedom
Sum Sq = deviance (within groups, and residual)
Mean Sq = variance (within groups, and residual)
F value = the value of the Fisher statistic test, so computed (variance within groups) / (variance residual)
Pr(>F) = p-value

	자유도(Df)	제곱합(Sum Sq)	평균제곱(Mean Sq)	F-값②	P-값③
집단 간	3	3794707	1264902	17.07 (=1264902 / 74098)	0.0001
집단 내	13305	985873109	74098		

① 집단별 평균의 차이를 확인
② 분산분석은 F 값이 클수록 귀무가설이 기각될 가능성이 커짐.
③ P-값이 0.05(0.01)보다 적으면, 95%(99%) 신뢰수준에서 통계적을 유의하다고 판단
※ 분산분석 결과만으로는 한 그룹만 평균의 차이가 있는지, 모든 그룹이 평균의 차이가 있는지는 확인할 수 없음.

평균차이 검정 방법론 개요

- 일원분산분석 원리
  - 요인수준별 종속변수의 평균과 오차
    - 오차의 정규성 가정

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, 2, \ldots, r; \quad j = 1, 2, \ldots, n$$

이때, Y_{ij} 는 i번째 수준에서 측정된 j번째 값,

μ_i 는 i번째 수준에서의 모평균

ε_{ij} 는 오차로서 서로 독립이며, 정규분포 $N(\mu_i, \sigma^2)$을 따름

- 종속변수의 측정값의 전체 변동

  = 요인수준 간 차이 + 그 밖의 설명되지 않는 요인에 의하여 발생하는 차이

① 측정값 Y_{ij}와 전체 측정값들의 평균 $\bar{Y}$ 간의 차이인 편차는 다음과 같이 분할 가능함

$$\underbrace{Y_{ij} - \bar{Y}}_{①} = \underbrace{(\bar{Y}_i - \bar{Y})}_{②} + \underbrace{(Y_{ij} - \bar{Y}_i)}_{③}$$

② 수준 i 에서의 측정값과 전체 평균 간의 차이, 즉 수준간의 차이에 따른 편차

③ 측정값과 수준 평균 간의 차이, 즉 수준 i에 의해서 설명될 수 없는 편차

- 집단간과 집단내의 변동의 제곱하여 그 분산의 비율로 계산
  - F-값 = 집단간 분산의 평균 제곱 / 집단내 총분산의 평균 제곱

① ② ③

$$\sum_{i=1}^{r}\sum_{j=1}^{n}(Y_{ij} - \bar{Y})^2 = \sum_{i=1}^{r} n(\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^{r}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_i)^2$$

SST = SSTR + SSE

① SST : 총제곱합 (Total Sum of Squares)

② SSTR : 처리제곱합 (Treatment Sum of Squares)

③ SSE : 오차제곱합 (Error Sum of Squares)

평균차이 검정 방법론 개요

- 분산분석 사후검정
  - 다중비교(Multiple Comparison)
    - 셋 이상의 집단의 평균을 두 개씩 짝지어 세부적으로 값의 차이를 비교할 수 있는 분석방법임.
  - 분석방법의 종류

- 최소유의차(LSD)
- 본페로니(Bonferroni)
- 튜키(Tukey)
- 던컨(Duncan)
- 쉐페(Scheffe)
- 던넷(Dunnett)
- 기타

각 분석방법들은 저마다 민감한 정도의 차이는 있으나, 그 차이가 크지 않아 어떠한 방법을 사용하더라도 결과가 심하게 달라지는 경우는 드문 편임.

- TukeyHSD()
- 계절별 아파트 평당 거래가격의 다중비교 사후검정 결과

```
> group_aov <- aov(price_pyung ~ season, data=apt2)
> TukeyHSD(group_aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price_pyung ~ season, data = apt2)

$season
                 diff        lwr        upr     p adj
겨울-가을 -27.923640 -45.413924 -10.433355 0.0002408
봄-가을   -42.082121 -58.887937 -25.276305 0.0000000
여름-가을  -7.425984 -24.395575   9.543608 0.6744629
봄-겨울   -14.158481 -31.542404   3.225441 0.1554967
여름-겨울  20.497656   2.955354  38.039958 0.0142820
여름-봄    34.656137  17.796192  51.516083 0.0000008
```

- diff: 개별집단간 평균 차이
- lwr: 하한 신뢰계수 구간
- upr: 상한 신뢰계수 구간
- P adj: p-값

데이터 가공하기

```
## 데이터 준비 및 가공하기
### 사용하게 될 패키지와 작업 폴더 파일 확인
library(dplyr) # dplyr 패키지
library(ggplot2) # 데이터 시각화 ggplot2 패키지

setwd("K:\\기타\\2019년2학기\\수치해석\\실습데이터") # 실습데이터가 있는 폴더로 작업폴더 변경
getwd() # 현재 작업 중인 폴더 (변경) 확인
list.files() # 현재 폴더내의 파일들 이름 확인

### 데이터 불러오기와 데이터 마이닝
apt <- read.csv("데이터_아파트매매가격.csv") # csv 데이터 불러오기
str(apt) # 데이터 구조 확인하기
summary(apt) # 데이터 전체 요약 통계 확인하기
```

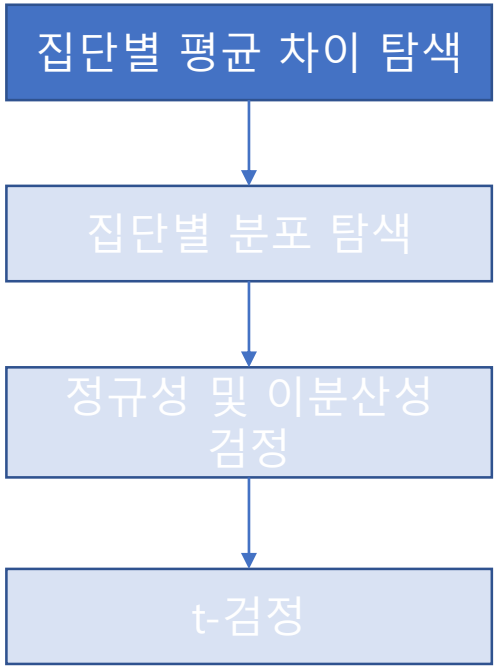
평균 차이 검정

데이터 가공하기

```
### 데이터 불러오기와 데이터 마이닝
apt <- read.csv("데이터_아파트매매가격.csv") # csv 데이터 불러오기
str(apt) # 데이터 구조 확인하기
summary(apt) # 데이터 전체 요약 통계 확인하기

attach(apt) # 데이터 객체 바로 접근하기
apt <- mutate(apt, price_pyung = apt_price / area_m2 * 3.3) # 아파트 거래 가격 평당 가격 변환하여 할당
x <- ifelse(year_built <1980, "1970s", # 아파트 건축년대 변수생성 작업: 1970년대 건축된 아파트
            ifelse(year_built >=1980 & year_built < 1990, "1980s" , # 1980년대 건축된 아파트
                   ifelse(year_built >=1990 & year_built < 2000, "1990s" , # 1990년대 건축된 아파트
                          ifelse(year_built >=2000 & year_built < 2010, "2000s" ,  # 2000년대 건축된 아파트
                                 "2010s")))) # 2010년대 건축된 아파트
apt <- mutate(apt, yr_built = x) # 아파트 건축년대 집단
table(apt$yr_built) # 일원빈도분포표

apt <- mutate(apt, yr_built2 = ifelse(year_built < 2000, "old",  "New")) # 아파트 건축년대 2000년 기준 요인 변수 생성
apt$yr_built2<- (as.factor(apt$yr_built2)) # 아파트 건축년대별 변수 yr_built : 문자 -> 요인 전환
table(apt$yr_built, apt$yr_built2) # 이원빈도분포표

apt$season <- ifelse(apt$ym_sale <=201811, "가을", # 계절(season) 요인 변수 생성
                     ifelse(apt$ym_sale >= 201812 & apt$ym_sale <= 201902, "겨울" ,
                            ifelse(apt$ym_sale >=201903 & apt$ym_sale <= 201905, "봄" , "여름")))

table(apt$ym_sale, apt$season)  # 이원빈도분포표

table(apt$urban) # 시군구별 빈도분포 확인
apt <- mutate(apt, urban2 = as.factor(ifelse(urban == "동", "도시", "농촌")))
table(apt$urban, apt$urban2)  # 이원빈도분포표

str(apt) # 데이터 구조 확인하기
apt2 <- select(apt, apt_price:urban, price_pyung:urban2) # 선택된 열만 추출하기

str(apt2) # 데이터 구조 확인하기
write.csv(apt2, "apt2.csv") # apt2  데이터 저장하기
file.exists("apt2.csv") # 저장된 파일 존재 확인하기
detach() # 데이터 객체 바로 접근하기 해제
```

두 집단 평균차이 검정: t-test

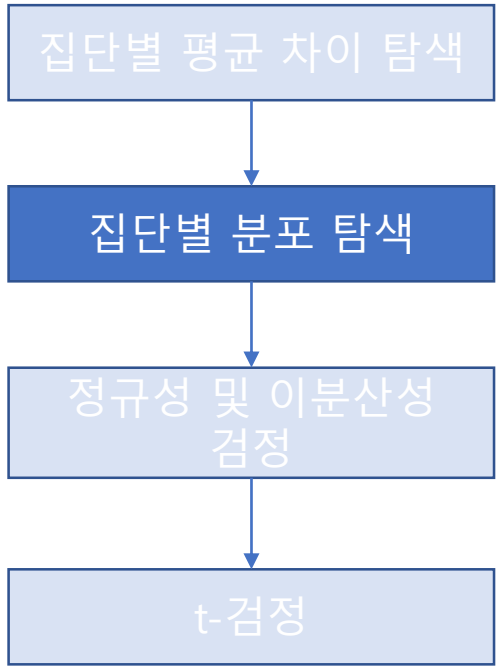
- 독립 두 표본 평균차이 검정: t-test
  - 도시와 농촌지역의 아파트의 평당 거래 가격(price_pyung)은 동일할까?



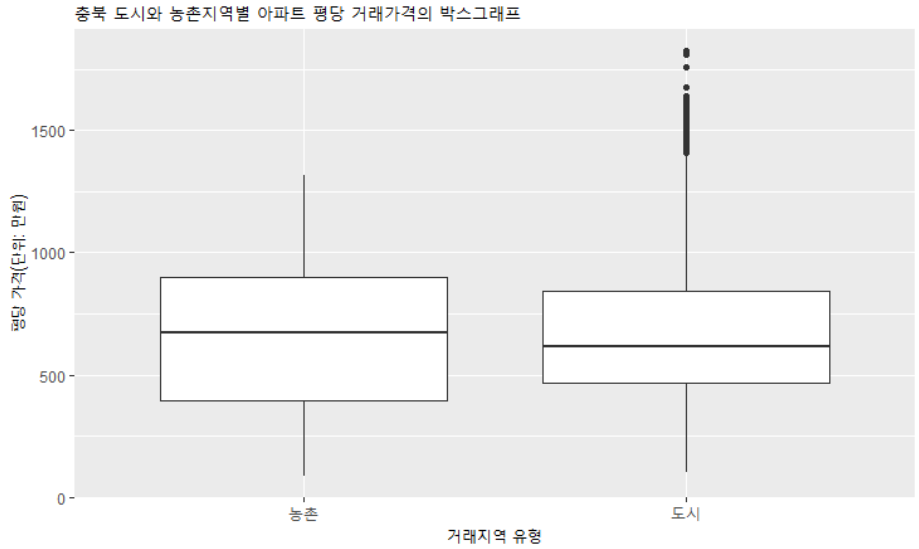

```
> apt2 <- read.csv("apt2.csv") # apt2.csv 데이터 불러오기
> str(apt2)
'data.frame':   13309 obs. of  13 variables:
 $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ apt_price  : int  4800 4500 4000 4000 4000 4000 4000 40
 $ area_m2    : num  39.8 39.8 39.8 39.8 39.8 39.8 39.8 39
 $ floor_no   : int  2 6 7 9 3 7 5 9 8 5 ...
 $ year_built : int  1998 1998 1998 1998 1998 1998 1998 19
 $ ym_sale    : int  201811 201811 201812 201812 201812 20
 $ day_sale   : int  6 13 10 10 10 10 10 10 10 10 ...
 $ urban      : Factor w/ 3 levels "동","면","읍": 3 3 3 3
 $ price_pyung: num  398 373 332 332 332 ...
 $ yr_built   : Factor w/ 5 levels "1970s","1980s",..: 3 3
 $ yr_built2  : Factor w/ 2 levels "New","Old": 2 2 2 2 2
 $ season     : Factor w/ 4 levels "가을","겨울",..: 1 1 2
 $ urban2     : Factor w/ 2 levels "농촌","도시": 1 1 1 1 1
> #* 1. 평균과 표준편차 확인하기
> apt2 %>% # 객체
+   group_by(urban2) %>% # 집단별 분류
+   summarise(price_m = mean(price_pyung),  # 평당 가격 집단별 평균
+             price_sd = sd(price_pyung)) # 표준편차 생성 및 확인
# A tibble: 2 x 3
  urban2 price_m price_sd
  <fct>    <dbl>    <dbl>
1 농촌      650.     266.
2 도시      674.     276.
```

두 집단 평균차이 검정: t-test

- 독립 두 표본 평균차이 검정: t-test
  - 도시와 농촌지역의 아파트의 평당 거래 가격(price_pyung)은 동일할까?

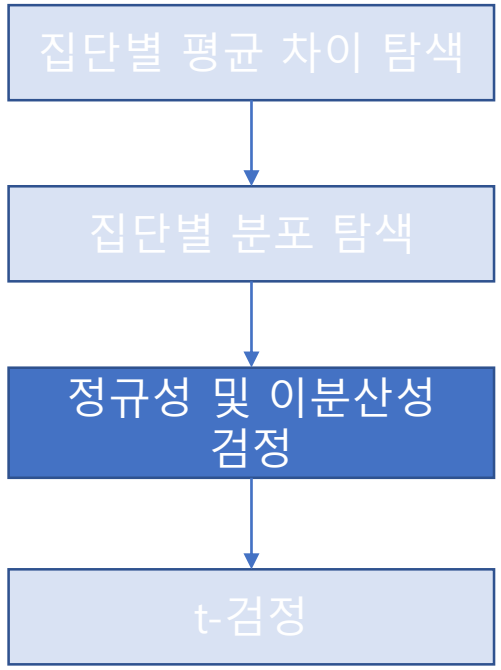



```
#* 2. 박스그래프로 분포 확인하기
ggplot(apt2, # 아파트 실거래 가격 자료
       aes(urban2, price_pyung)) + # 미학 그래프 셋팅(x = 읍면동, y= 평당 아파트 가격)
      geom_boxplot() + # 박스 그리프(도시 및 농촌지역 평당 거래 가격)
      labs(title = "충북 도시와 농촌지역별 아파트 평당 거래가격의 박스그래프", # 제목
           x = "거래지역 유형", # x축 제목
           y = "평당 가격(단위: 만원)")  # y축 제목
```



두 집단 평균차이 검정: t-test

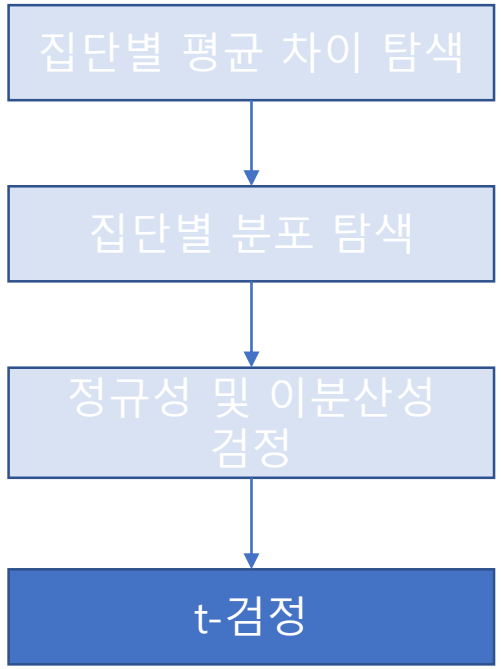
- 독립 두 표본 평균차이 검정: t-test
  - 도시와 농촌지역의 아파트의 평당 거래 가격(price_pyung)은 동일할까?




```
> ?var.test() # F Test to Compare Two Variances
> var.test( # H0: true ratio of variances is equal to 1
+   apt2[apt2$urban2=="도시", 9], # 도시이면서 9번째 열(평당 가격) 데이터
+   apt2[apt2$urban2=="농촌", 9]) # 농촌이면서 9번째 열(평당 가격) 데이터

        F test to compare two variances

data:  apt2[apt2$urban2 == "도시", 9] and apt2[apt2$urban2 == "농촌", 9]
F = 1.0775, num df = 8847, denom df = 4460, p-value = 0.004265
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.023766 1.133492
sample estimates:
ratio of variances
          1.077466
```

두 집단 평균차이 검정: t-test

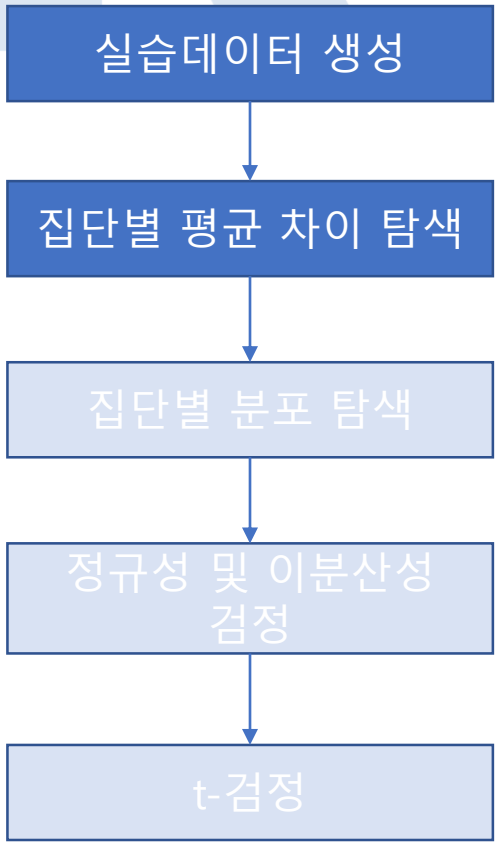
- 독립 두 표본 평균차이 검정: t-test
  - 도시와 농촌지역의 아파트의 평당 거래 가격(price_pyung)은 동일할까?




```
> t.test( # H0: true difference in means is equal to 0
+   price_pyung ~ urban2, # 평당 가격을 거래지역(urban2) 집단별로 비교
+   data = apt2, # 비교할 데이터 객체
+   alternative = "two.sided", # alternative = c("two.sided", "less", "greater")
+   paired = FALSE, # 대응 표본이 아닌 독립 표본
+   var.equal = FALSE,  # 이분산성 가정하여 진단
+   conf.level = 0.99) # 신뢰구간 범위: 99%

        Welch Two Sample t-test

data:  price_pyung by urban2
t = -4.9415, df = 9246.3, p-value = 7.888e-07
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
 -37.15406 -11.68880
sample estimates:
mean in group 농촌 mean in group 도시
         649.9099           674.3314
```

두 집단 평균차이 검정: t-test

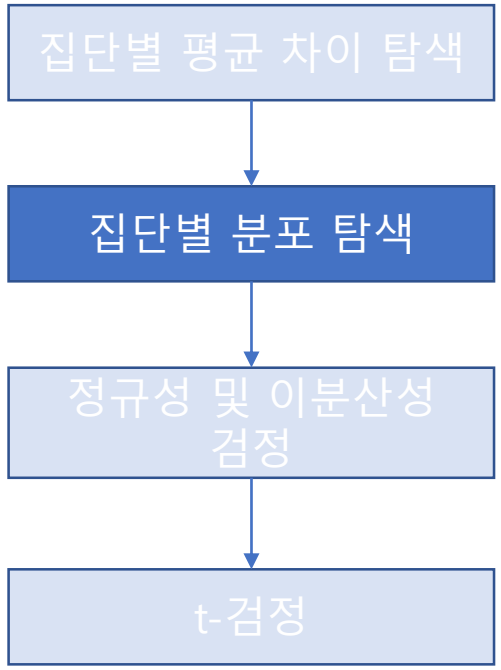
- 대응 두 표본 평균 차이 검정
  - 수치해석 수강생들은 중간고사보다 기말고사의 성적이 더 좋을까?



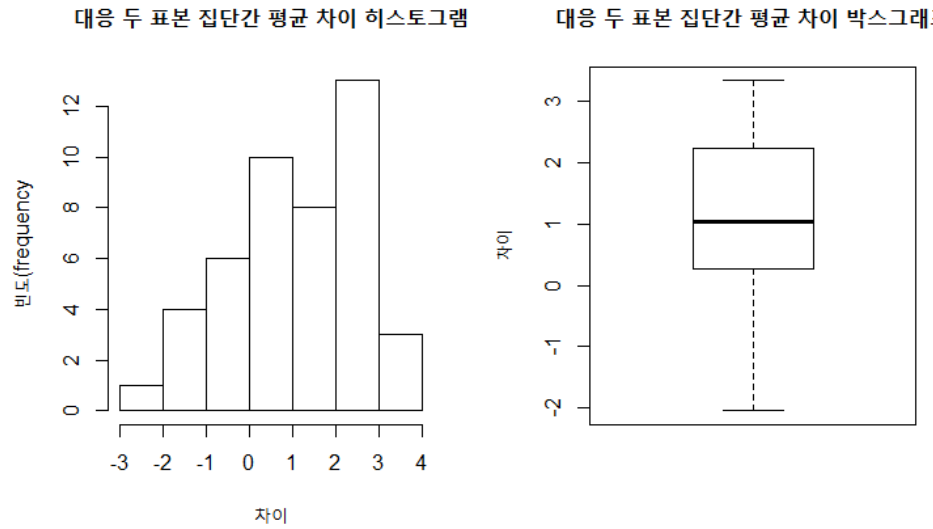

```
> mid_score <- rnorm(n=45, mean=23, sd=1) # 부삭위 성규분포 형태의
> final_score <- rnorm(n=45, mean=24, sd=1) # 무작위 정규분포 형태
> x <- as.data.frame(cbind(id= 1:45, mid = mid_score, final=fina
> class(x) # 속성 확인
[1] "data.frame"
> #* 1. 요약 통계
> summary(x) # x 객체 요약 통계
       id             mid            final
 Min.   : 1   Min.   :20.65   Min.   :21.30
 1st Qu.:12   1st Qu.:22.49   1st Qu.:23.45
 Median :23   Median :23.09   Median :24.09
 Mean   :23   Mean   :23.15   Mean   :24.14
 3rd Qu.:34   3rd Qu.:24.03   3rd Qu.:25.08
 Max.   :45   Max.   :25.21   Max.   :26.20


> #* 1. 요약 통계
> summary(x) # x 객체 요약 통계
       id             mid            final
 Min.   : 1   Min.   :20.65   Min.   :21.30
 1st Qu.:12   1st Qu.:22.49   1st Qu.:23.45
 Median :23   Median :23.09   Median :24.09
 Mean   :23   Mean   :23.15   Mean   :24.14
 3rd Qu.:34   3rd Qu.:24.03   3rd Qu.:25.08
 Max.   :45   Max.   :25.21   Max.   :26.20
> x <- mutate(x, diff=final-mid) # 대응 두 표본의 차이 변수 생성
> summary(x$diff) # 차이변수 요약 통계
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.0476  0.2685  1.0316  0.9925  2.2412  3.3435
```

두 집단 평균차이 검정: t-test

- 대응 두 표본 평균 차이 검정
  - 수치해석 수강생들은 중간고사보다 기말고사의 성적이 더 좋을까?

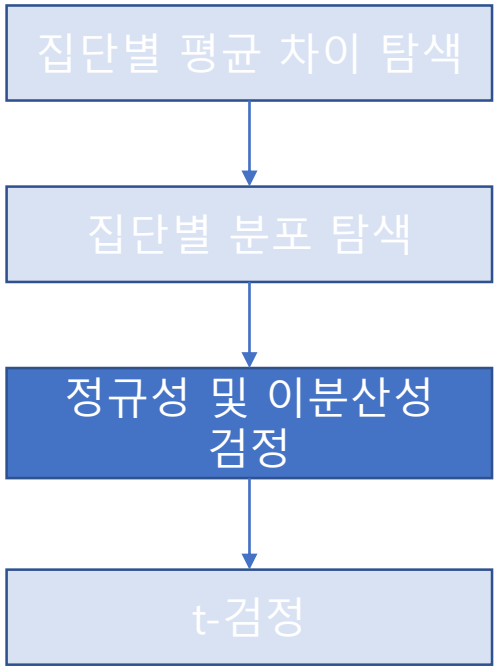



```
> #* 2. 집단별 분포 탐색: 히스토 그램과 박스그래프
> par(mfrow=c(1,2)) # 한 화면에 1*2 그래프 창 분할
> hist(x$diff,
+      main = "대응 두 표본 집단간 평균 차이 히스토그램",
+      xlab = "차이",
+      ylab = "빈도(frequency")
> boxplot(x$diff,
+         main = "대응 두 표본 집단간 평균 차이 박스그래프",
+         ylab = "차이")
```



두 집단 평균차이 검정: t-test

- 대응 두 표본 평균 차이 검정
  - 수치해석 수강생들은 중간고사보다 기말고사의 성적이 더 좋을까?

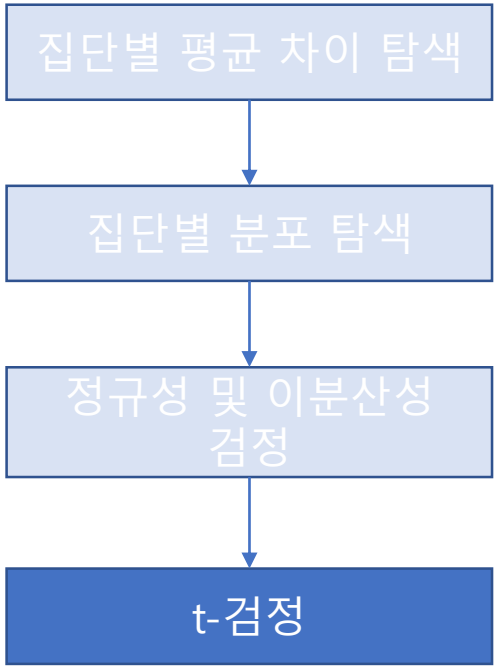



```
> #* 3-1. 정규성 검정
> ?shapiro.test()
> shapiro.test(x$mid) # 중간고사 성적 정규성 검정

        Shapiro-Wilk normality test

data:  x$mid
W = 0.97754, p-value = 0.5234

> shapiro.test(x$final) # 기말고사 성적 정규성 검정

        Shapiro-Wilk normality test

data:  x$final
W = 0.97024, p-value = 0.2955
```

```
> #* 3-2. 이분산성 검정
> ?var.test
> var.test(x$mid, x$final,
+          alternative = "greater") # 오른쪽 단측 검정

        F test to compare two variances

data:  x$mid and x$final
F = 1.0354, num df = 44, denom df = 44, p-value = 0.4544
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 0.6271428       Inf
sample estimates:
ratio of variances
          1.035372
```

두 집단 평균차이 검정: t-test

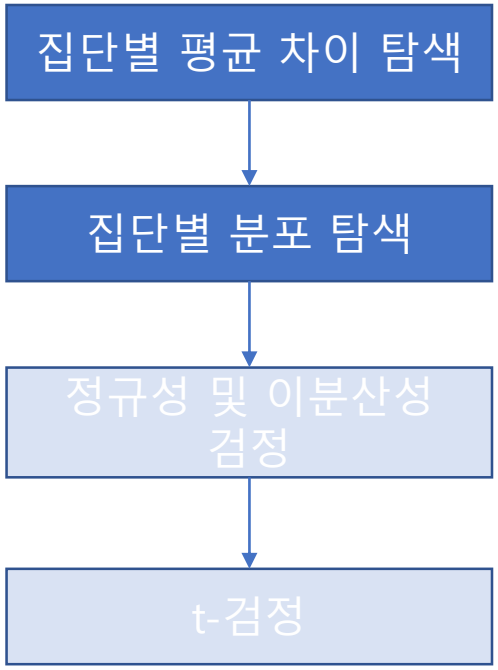
- 대응 두 표본 평균 차이 검정
  - 수치해석 수강생들은 중간고사보다 기말고사의 성적이 더 좋을까?
    - 오른쪽 단측검정: alternative = “greater”



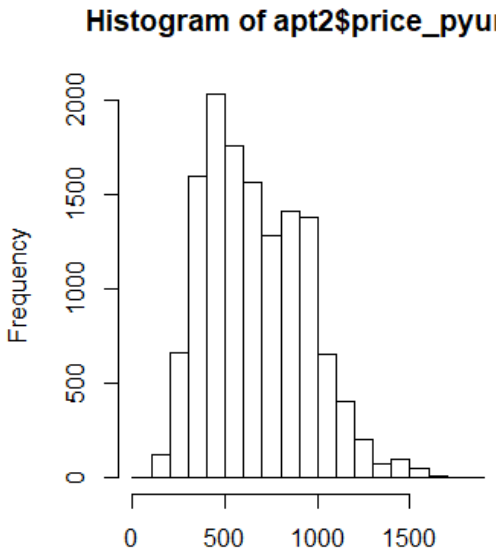

```
> #* 4. 대응 두 표본 평균검정(t-test)
> t.test(x$mid, x$final,
+        alternative = "greater", # 오른쪽 단측 검정
+        var.equal = TRUE, # 등분산성
+        paired=TRUE) # 대응표본

        Paired t-test

data:  x$mid and x$final
t = -4.417, df = 44, p-value = 1
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -1.370083       Inf
sample estimates:
mean of the differences
             -0.9925254
```

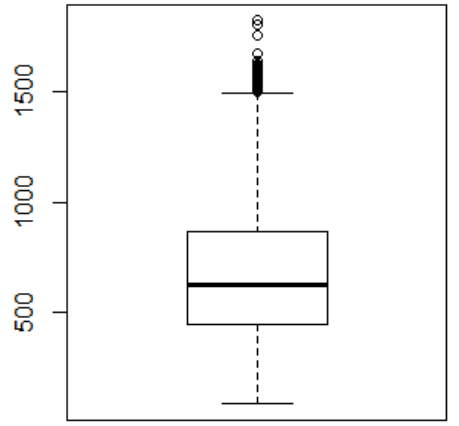
두 집단 평균차이 검정: t-test

- 단일표본 t-test (One sample t-test)
  - 지난 1년간 충청북도 지역의 아파트 평당 거래가격이 평균 670만원보다 작게 거래되었다고 할 수 있는가?



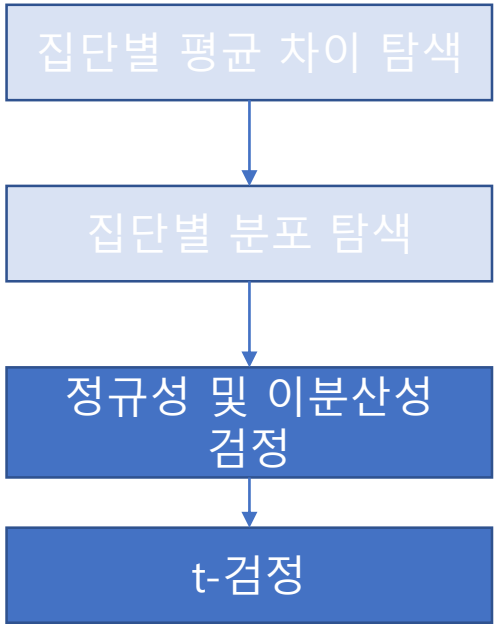

```
> #* 1. 요약통계
> mean(apt2$price_pyung)
[1] 666.1456
> #* 2. histgram 과 박스 그래프
> hist(apt2$price_pyung)
> boxplot(apt2$price_pyung)
```





두 집단 평균차이 검정: t-test

- 단일표본 t-test (One sample t-test)
  - 지난 1년간 충청북도 지역의 아파트 평당 거래가격이 평균 670만원보다 작게 거래되었다고 할 수 있는가?
    - 왼쪽 단측검정: alternative = "less"




```
> #* 3. 정규성 검정
> shapiro.test(apt2$price_pyung) # n>30이므로 생략 가능
Error in shapiro.test(apt2$price_pyung) :
  샘플의 크기는 반드시 3 부터 5000 이내에 있어야 합니다
> #* 4. 단일 표본 t-검정
> t.test(apt2$price_pyung, # Ho: true mean is not less than 650
+        mu = 670, # 비교하고자 하는 평균 값 설정
+        alternative = "less", # 오른쪽 단측검정 선택
+        conf.level = 0.95) # 신뢰구간 범위

        One Sample t-test

data:  apt2$price_pyung
t = -1.6306, df = 13308, p-value = 0.0515
alternative hypothesis: true mean is less than 670
95 percent confidence interval:
    -Inf 670.034
sample estimates:
mean of x
 666.1456
```

연습문제 01

- 다음 중 F-분포를 활용하여야만 집단간 평균의 차이를 검정할 수 있는 경우를 모두 적으시오.
  - ① 남자와 여자 간 소득의 차이 비교
  - ② 도시와 농촌지역의 주택가격은 차이가 있을까?
  - ③ 계절(봄, 여름, 가을, 겨울)별 아파트 거래가격의 평균은 동일한가?
  - ④ 중간고사에 비하여 기말고사의 성적은 올랐을까?
  - ⑤ 1, 2, 3, 4학년의 학생 성적의 평균의 차이는 없는가?
  - ⑥ 과외를 하기 전과 후의 반 학생들의 성적 변화
  - ⑦ 서울시의 집값 평균은 평당 1,000만원 보다 높을까?
  - ⑧ 남자와 여자의 성별, 청년기, 중년층, 장년층, 노년층의 연령대별 소득수준의 차이 비교

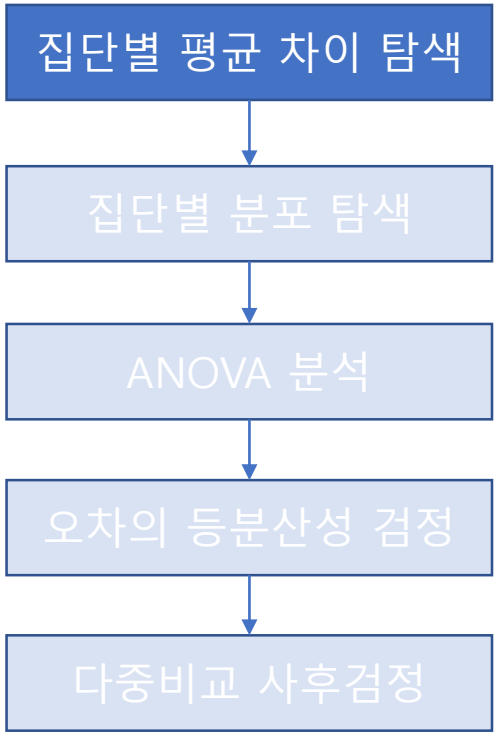
연습문제 02

- 2000년 이전(old)에 건축된 아파트와 이후(new)에 지어진 아파트의 평당 거래가격의 평균은 같을까?에 대하여 독립 두 표본 평균차이 검정으로 t-test를 실행하고자 한다.
  - 두 집단의 표본의 평균은 각각 얼마인가요? 적으시오.
  - 이분산성 검정에서 F-통계량(값)과 p-값이 얼마인 지 적으시오.
    - 이 때 이들 두 집단의 분산이 동일한가에 대한 귀무가설을 기각하여야 하나요?
  - T-test 실행 결과에서 t-값과 p-값을 각각 적으시오.
    - T-test 실행결과로 볼 때, 두 표본의 평균은 동일한지에 대한 귀무가설을 채택하여야 하나요?

둘 이상의 평균차이 검정: ANOVA

- 일원분산분석
  - 봄, 여름, 가을, 겨울의 계절별 아파트 거래 가격의 차이가 있는가?

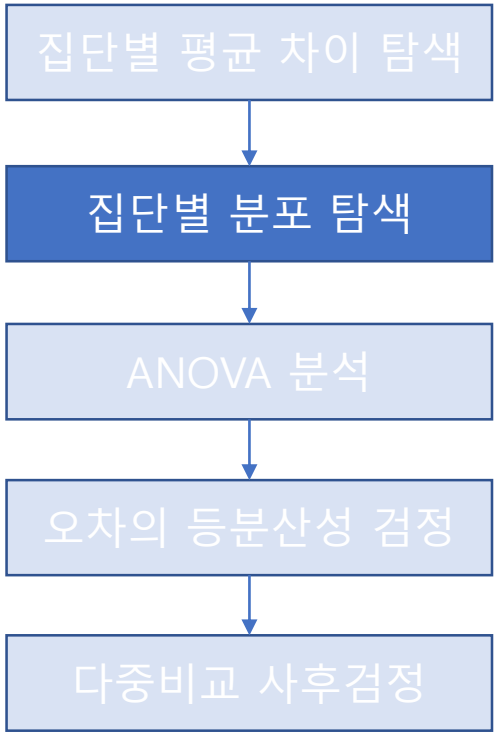
```
> ?aov() # Fit an Analysis of Variance Model
```



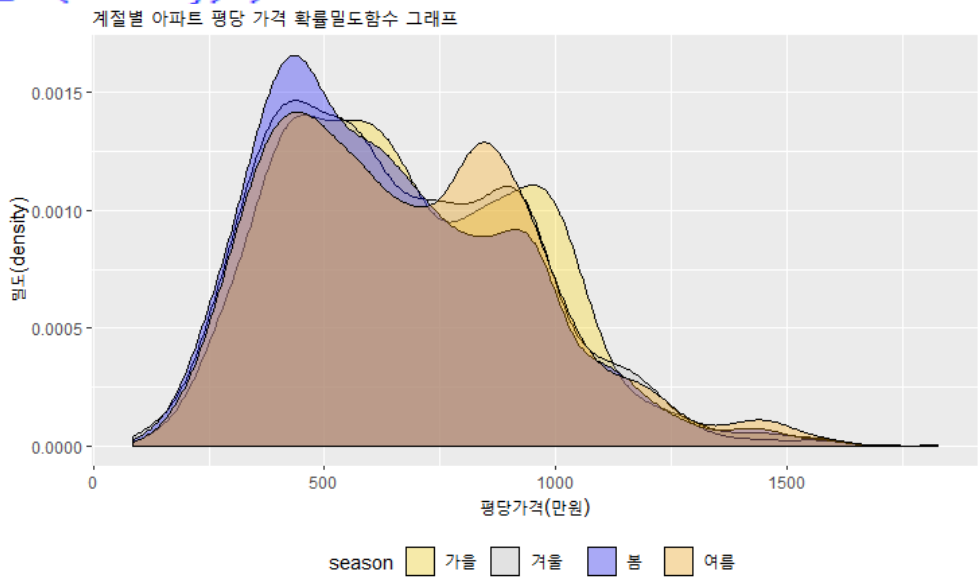

```
> #* 1. 요약 통계확인하기
> apt2 %>%
+   group_by(season) %>% # 계절별 집단 구분
+   summarise(n = n(), # 갯수
+             mean = mean(price_pyung), # 평균
+             var = var(price_pyung)) # 분산 요약통계
# A tibble: 4 x 4
  season     n  mean    var
  <fct>  <int> <dbl>  <dbl>
1 가을    3419  685. 72004.
2 겨울    3004  658. 72160.
3 봄       3510  643. 74187.
4 여름    3376  678. 77850.
```

둘 이상의 평균차이 검정: ANOVA

- 일원분산분석
  - 봄, 여름, 가을, 겨울의 계절별 아파트 거래 가격의 차이가 있는가?

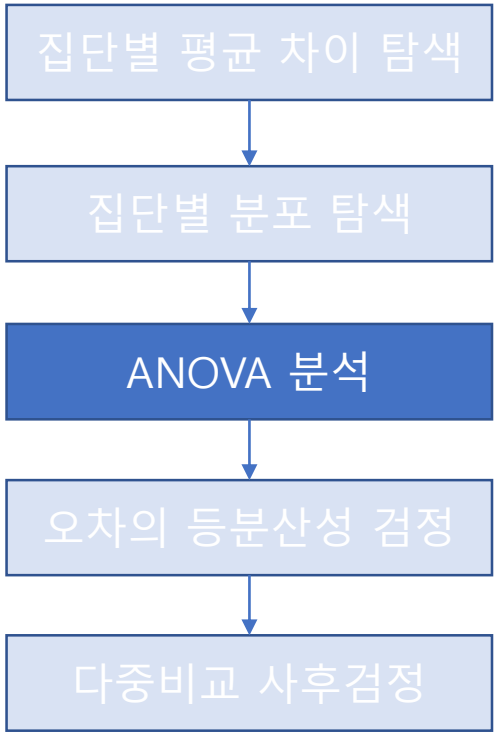



```
> u_d <- ggplot(apt2, aes(price_pyung, fill=season)) +
+   geom_density(alpha=0.3) +
+   scale_fill_manual(values = c('gold','gray', 'blue', 'orange')) +
+   theme(legend.position="bottom") + # 범례 위치: bottom, left, right
+   labs(title = "계절별 아파트 평당 가격 확률밀도함수 그래프",
+        x= "평당가격(만원)",
+        y = "밀도(density)")
> u_d
```



둘 이상의 평균차이 검정: ANOVA

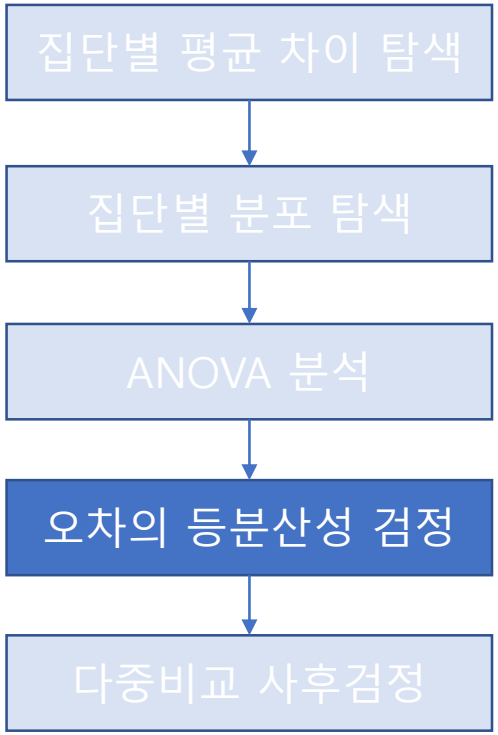
- 일원분산분석
  - 봄, 여름, 가을, 겨울의 계절별 아파트 거래 가격의 차이가 있는가?




```
> #* 3. 분산분석
> a.t <- aov(price_pyung ~ season, # 계절별 평당가격 분산분석
+            data = apt2) # 데이터 객체
> a.t # 할당 객체 결과
Call:
   aov(formula = price_pyung ~ season, data = apt2)

Terms:
                   season Residuals
Sum of Squares    3794707 985873109
Deg. of Freedom         3     13305

Residual standard error: 272.2094
Estimated effects may be unbalanced
> summary(a.t) # 분산분석 결과의 요약 통계
              Df    Sum Sq Mean Sq F value   Pr(>F)
season         3   3794707 1264902   17.07 4.62e-11 ***
Residuals  13305 985873109   74098
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

둘 이상의 평균차이 검정: ANOVA

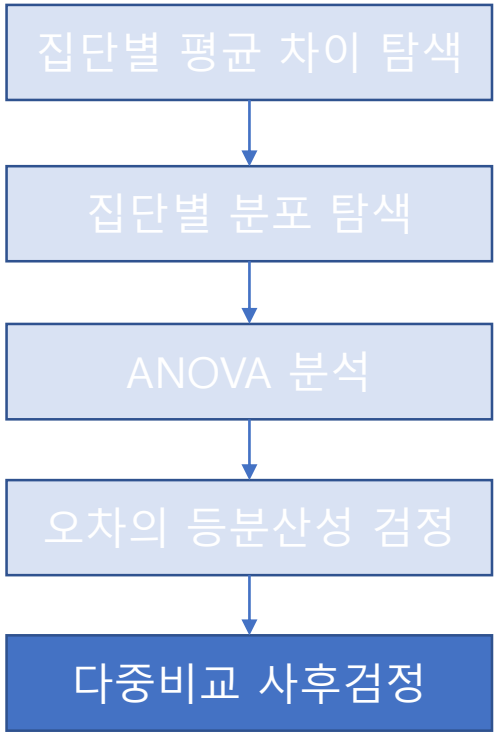
- 일원분산분석
  - 봄, 여름, 가을, 겨울의 계절별 아파트 거래 가격의 차이가 있는가?




```
> library(lawstat)
> ?bartlett.test() # Bartlett Test of Homogeneity of Variances
> b.t <- bartlett.test(price_pyung ~ season, # 바렛 이분산성 검정
+                      data = apt2) # 데이터 객체
> b.t

        Bartlett test of homogeneity of variances

data:  price_pyung by season
Bartlett's K-squared = 6.6246, df = 3, p-value = 0.08488
```

둘 이상의 평균차이 검정: ANOVA

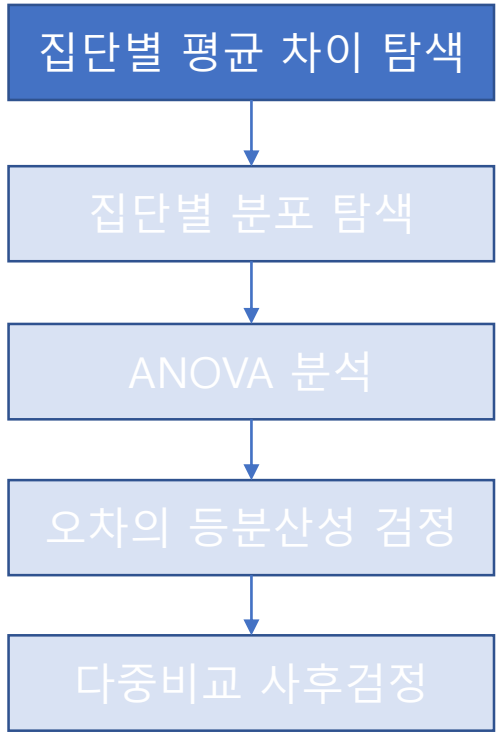
- 일원분산분석
  - 봄, 여름, 가을, 겨울의 계절별 아파트 거래 가격의 차이가 있는가?




```
> #* 5. 사후검정: 다중 비교
> group_aov <- aov(price_pyung ~ season, data=apt2) # 절차 1: aov() 함수 실행
> ?TukeyHSD() # Compute Tukey Honest Significant Differences
> TukeyHSD(group_aov) #  절차 2: 터키 사후 검정
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price_pyung ~ season, data = apt2)

$season
                 diff        lwr        upr     p adj
겨울-가을 -27.923640 -45.413924 -10.433355 0.0002408
봄-가을   -42.082121 -58.887937 -25.276305 0.0000000
여름-가을  -7.425984 -24.395575   9.543608 0.6744629
봄-겨울   -14.158481 -31.542404   3.225441 0.1554967
여름-겨울  20.497656   2.955354  38.039958 0.0142820
여름-봄    34.656137  17.796192  51.516083 0.0000008
```

둘 이상의 평균차이 검정: ANOVA

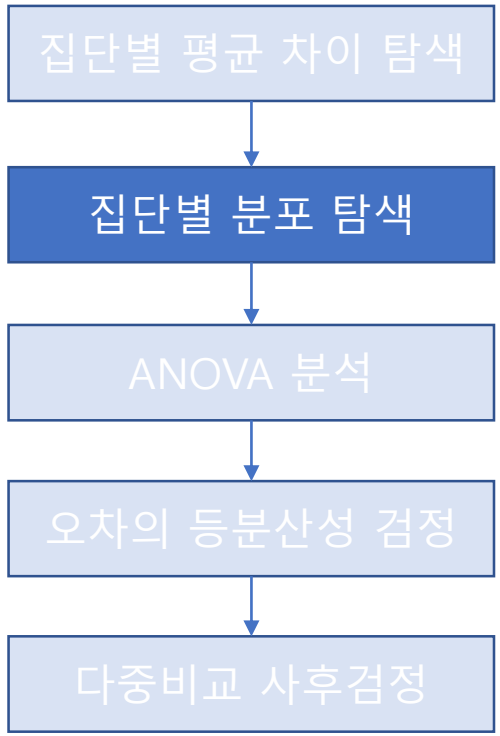
- 이원분산분석
  - 계절별 읍면동별 아파트 거래 가격의 차이가 있는가?



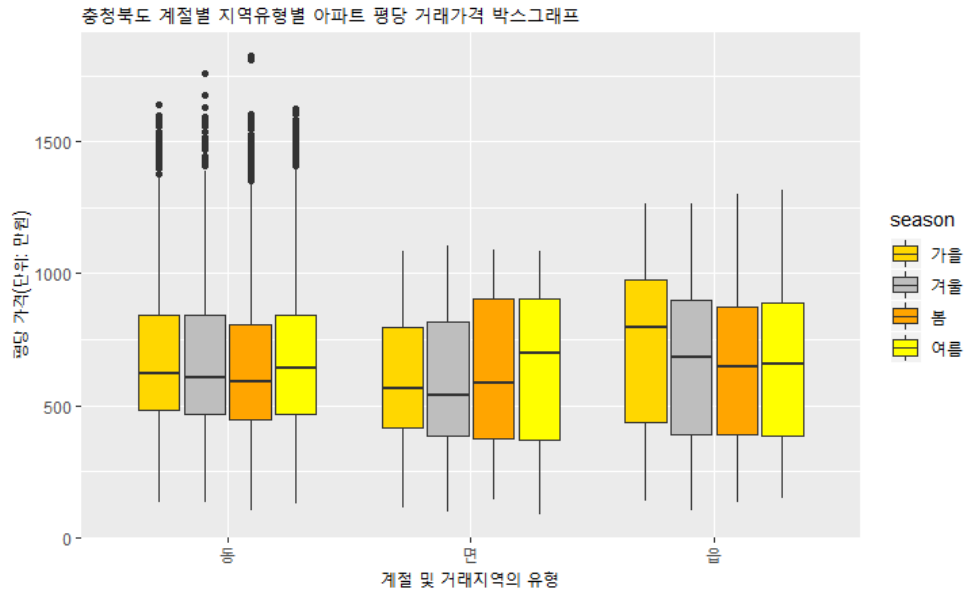

```
> #* 1. 요약 통계확인하기
> apt2 %>%
+   group_by(season, urban) %>% # 계절별, 읍면동별  집단 구분
+   summarise(n = n(), # 갯수
+             mean = mean(price_pyung), # 평균
+             var = var(price_pyung)) # 분산 요약통계
# A tibble: 12 x 5
# Groups:   season [4]
   season urban     n  mean    var
   <fct>  <fct> <int> <dbl>  <dbl>
 1 가을   동      2088  681. 71024.
 2 가을   면       341  597. 52876.
 3 가을   읍       990  725. 76486.
 4 겨울   동      1934  671. 74137.
 5 겨울   면       265  593. 59212.
 6 겨울   읍       805  645. 69856.
 7 봄     동      2348  652. 76948.
 8 봄     면       384  612. 71752.
 9 봄     읍       778  633. 66416.
10 여름   동      2478  692. 80268.
11 여름   면       342  645. 71396.
12 여름   읍       556  636. 68022.
```

둘 이상의 평균차이 검정: ANOVA

- 이원분산분석
  - 계절별 읍면동별 아파트 거래 가격의 차이가 있는가?

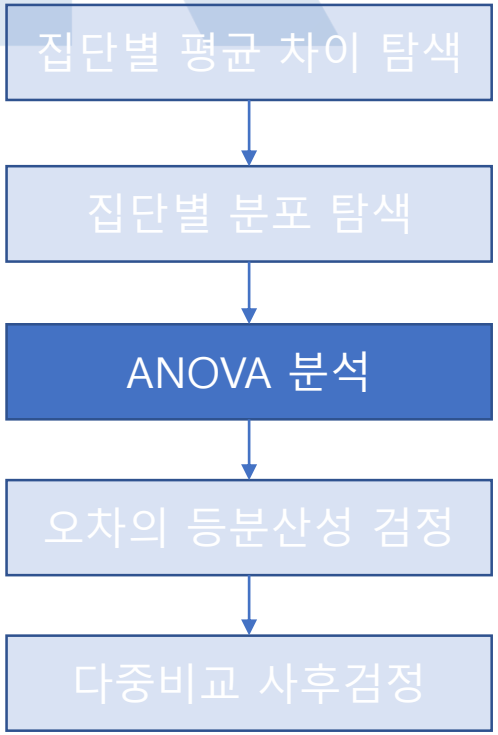



```
> #* 2. 집단별 분포 확인하기
> ggplot(apt2, # 아파트 실거래 가격 자료
+        aes(urban, price_pyung, fill=season)) + # 미학 그래프 셋팅(x = 읍면동 요인
+   geom_boxplot() + # 박스 그래프(읍면동 평당 거래 가격)
+   scale_fill_manual(values = c('gold','gray', 'orange', 'yellow')) +
+   labs(title = "충청북도 계절별 지역유형별 아파트 평당 거래가격 박스그래프", # 제목
+        x = "계절 및 거래지역의 유형", # x축 제목
+        y = "평당 가격(단위: 만원)") # y축 제목
```



둘 이상의 평균차이 검정: ANOVA

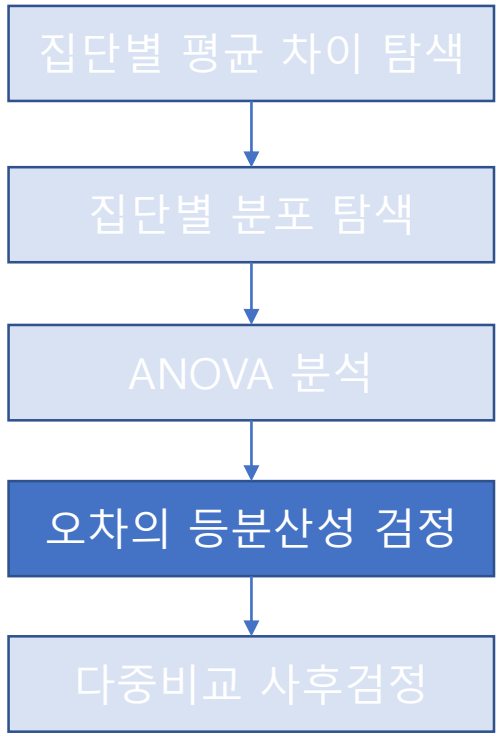
- 이원분산분석
  - 계절별 읍면동별 아파트 거래 가격의 차이가 있는가?




```
> #* 3. 분산분석
> a.t <- aov(price_pyung ~ season + urban, data = apt2) # 이원분산분석: 상호작용항 미추가
> summary(a.t)
               Df    Sum Sq Mean Sq F value   Pr(>F)
season          3   3794707 1264902   17.14 4.15e-11 ***
urban           2   4345301 2172650   29.45 1.74e-13 ***
Residuals   13303 981527809   73782
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> a.t2 <- aov(price_pyung ~ season * urban, data = apt2) # 이원 분산분석: 상호작용항 추가
> summary(a.t2)
               Df    Sum Sq Mean Sq F value   Pr(>F)
season          3   3794707 1264902  17.206 3.79e-11 ***
urban           2   4345301 2172650  29.553 1.56e-13 ***
season:urban    6   3968309  661385   8.996 7.81e-10 ***
Residuals   13297 977559499   73517
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

둘 이상의 평균차이 검정: ANOVA

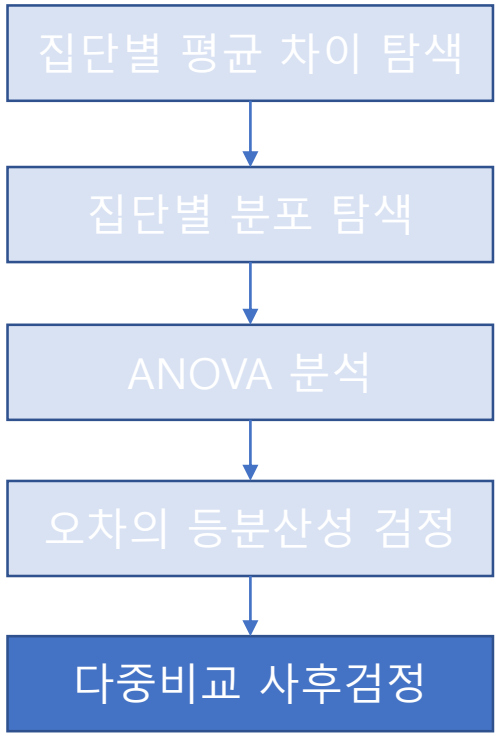
- 이원분산분석
  - 계절별 읍면동별 아파트 거래 가격의 차이가 있는가?




```
> #* 4. 오자의 등분산성 검성
> ?bartlett.test() # Bartlett Test of Homogeneity of Variances
> b.t <- bartlett.test(price_pyung ~ interaction(season, urban), data = apt2)
> b.t

        Bartlett test of homogeneity of variances

data:  price_pyung by interaction(season, urban)
Bartlett's K-squared = 43.529, df = 11, p-value = 8.782e-06
```

둘 이상의 평균차이 검정: ANOVA

- 이원분산분석
  - 계절별 읍면동별 아파트 거래 가격의 차이가 있는가?




```
> #* 5. 사후검정: 다중 비교
> group_aov <- aov(price_pyung ~ season + urban, data=apt2) # 절
> ?TukeyHSD() # Compute Tukey Honest Significant Differences
> TukeyHSD(group_aov) #  절차 2: 터키 사후 검정
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price_pyung ~ season + urban, data = apt2)

$season
                 diff       lwr        upr     p adj
겨울-가을 -27.923640 -45.37665 -10.470631 0.0002320
봄-가을   -42.082121 -58.85212 -25.312122 0.0000000
여름-가을  -7.425984 -24.35941   9.507442 0.6729688
봄-겨울   -14.158481 -31.50536   3.188393 0.1540408
여름-겨울  20.497656   2.99274  38.002572 0.0139977
여름-봄    34.656137  17.83212  51.480151 0.0000007

$urban
            diff       lwr        upr     p adj
면-동 -61.225014 -79.93729 -42.512740 0.0000000
읍-동  -9.034633 -22.27732   4.208052 0.2460165
읍-면  52.190381  31.36041  73.020348 0.0000000
```

상호작용항 추가 경우

```
> group_aov <- aov(price_pyung ~ season * urban, data=apt2) # 절차 1
> TukeyHSD(group_aov) #  절차 2: 터키 사후 검정
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = price_pyung ~ season * urban, data = apt2)

$season
                 diff       lwr        upr     p adj
겨울-가을 -27.923640 -45.34526 -10.502017 0.0002247
봄-가을   -42.082121 -58.82196 -25.342280 0.0000000
여름-가을  -7.425984 -24.32896   9.476990 0.6717046
봄-겨울   -14.158481 -31.47416   3.157197 0.1528179
여름-겨울  20.497656   3.02422  37.971092 0.0137613
여름-봄    34.656137  17.86238  51.449895 0.0000007

$urban
            diff       lwr        upr     p adj
면-동 -61.225014 -79.90364 -42.546392 0.0000000
읍-동  -9.034633 -22.25350   4.184236 0.2447818
읍-면  52.190381  31.39787  72.982888 0.0000000

$`season:urban`
                          diff         lwr         upr     p adj
겨울:동-가을:동   -9.9142004  -37.883589  18.0551880 0.9918368
봄:동-가을:동    -29.2827161  -55.941289  -2.6241434 0.0172831
여름:동-가을:동   10.6245985  -15.702771  36.9519684 0.9768522
```

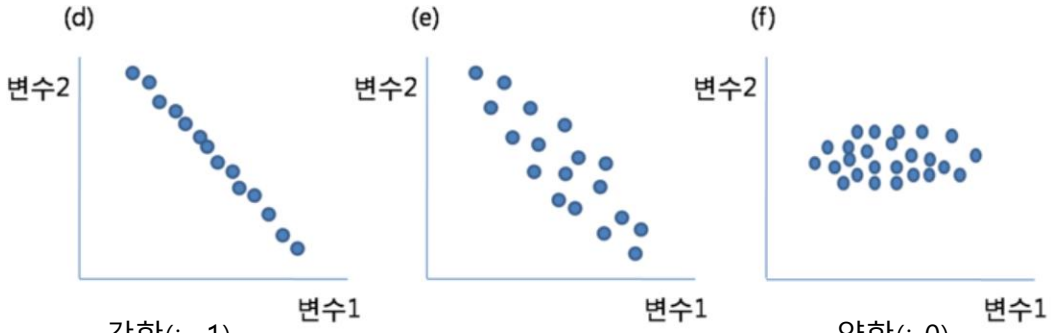
연습문제 03

- 아파트 평당 거래 가격이 2개 수준의 읍면동(urban)별 5개 수준의 건축년대(yr_built) 별 차이가 있는 지 여부를 진단하고, 사후 검정으로 다중 검정을 수행하고자 한다. 다음의 질문에 답하시오.
    - 1990년대 건축된 읍지역의 아파트 평당 거래가격의 평균은 얼마인가요?
    - 분산분석 실행 결과에서 urban와 평균제곱합과 F-값은 얼마인가요?
    - 사후 검정인 다중비교 검정에서 면과 동 지역의 평당 거래가격의 차이는 평균 얼마인지 적으시오.

상관관계 분석

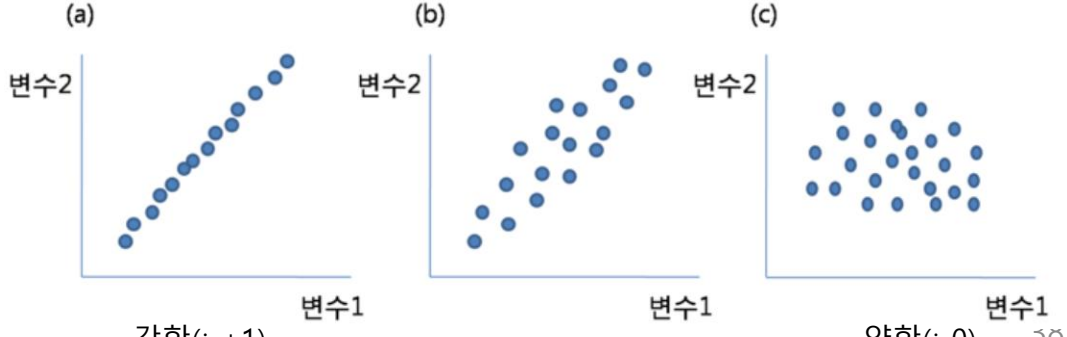
상관관계의 개념

- 관계의 정의
  - 공분산:
    - 2개의 확률변수의 상관정도를 나타내는 값
  - 상관성(연관성): correlation or association
    - 어떤 한 변수가 다른 변수들과 같이 공변(corvariance)하는 것
    - 양의 상관성(-1<r<0): 동일한 방향으로 공변
    - 음의 상관성(0<r<1): 반대의 방향으로 공변
    - 상관성 없음(r=0): 관계 없음
  - 인과성: causality
    - 선행하는 한 변수(X)가 후행하는 다른 변수의 원인(Y)이 되고 있다고 믿어지는 관계
      - X = 설명변수, 독립변수
      - Y = 결과변수, 종속변수



<음(-)의 상관성>

- 상관성과 인과성의 차이
  - 상관성이 있다는 것은 잠재적 인과성을 내포
    - 상관관계는 인과관계의 필요조건(necessary condition)이다
  - 5가지 유형
    - 단순한 우연의 일치
      - **"까마귀 날자 배 떨어진다"**
    - 반영되지 않았던 제3의 변인 z 가 x 와 y 두 변인에게 영향을 끼칠 수 있음(조절 및 매개 효과)
    - 서로가 서로에게 원인인 동시에 결과가 됨(동시성)
    - x 가 원인이고 y 가 결과 (x → y)
    - y 가 원인이고 x 가 결과 (y → x)

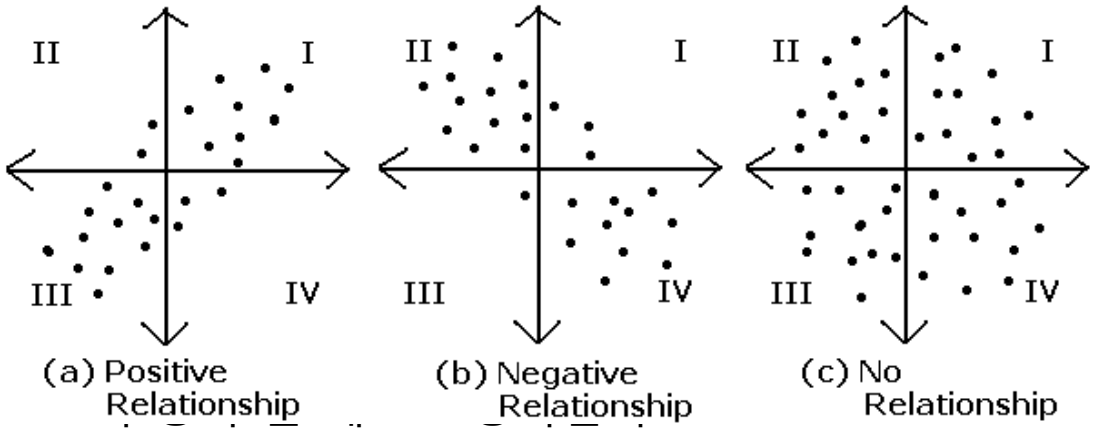


<양(+)의 상관성>

상관관계의 개념

- 공분산과 상관계수(correlation coefficient)
  - 공분산: x의 편차와 y의 편차를 곱한 값들의 평균

$$Cov(X,\ Y)\ =\ E((X\ -\ \mu)(Y\ -\ \nu))$$



- Cov(X, Y) > 0
- Cov(X, Y) < 0 X가 증가 할 때 Y는 감소한다.
- Cov(X, Y) = 0 공분산이 0이라면 두 변수간에는 아무런 선형관계가 없으며 두 변수는 서로 독립적인 관계에 있음
- x나 y의 변수의 단위의 크기에 영향을 받음
  - 10만점 x와 100점 만점 x의 공분산 크기는 다름

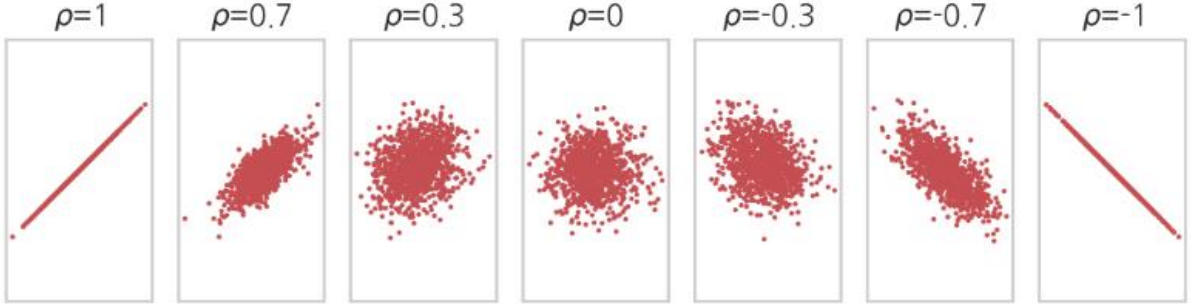
상관관계의 개념

- 공분산과 상관계수
  - 상관계수(ρ)
    - 공분산에서 확률변수의 절대적 크기에 영향을 받지 않도록 분산의 크기로 표준화한 값
    - 특성
      ① 상관계수의 절대값은 1을 넘을 수 없다.
      ② 확률변수 X, Y가 독립이라면 상관계수는 0이다.
      ③ X와 Y가 선형적 관계라면 상관계수는 1 혹은 -1이
      ④ 상관계수는 직선식의 기울기(β)와는 아무런 상관이 없다

$$Cov(X,\ Y) = E((X - \mu)(Y - \nu))$$

$$\rho = \frac{Cov(X,\ Y)}{\sqrt{Var(X)\ Var(Y)}}, \quad -1 \leq \rho \leq 1$$

  - 상관계수의 크기와 산점도(분포) 형태



상관관계의 개념

- 상관계수의 종류
  - 자료 형태에 따른 분석 방법 선택

두 변수의 자료형태	분석방법	예시
연속(양) + 연속(양)	Pearson's correlation	▪ 아버지의 혈압과 아들의 혈압의 관계 ▪ 입원기간과 수술시간의 관계 ▪ 혈압과 연령의 관계
연속(양) + 순위(서열)	Spearman's correlation Kendall's correlation	▪ 경제적 수입과 삶의 질의 정도(리커트 척도)와의 관계 ▪ 통증(5점 scale)과 암의 stage(5점 scale)와의 관계
순위(서열) + 순위(서열)		

- 분석절차
  - 자료의 속성 파악 → 산점도 그래프로 경향성 파악 → 적합한 분석방법 채택 → 상관계수 계산 → 가설검정 및 통계적 유의성 파악

상관관계의 개념

- 공분산과 상관계수

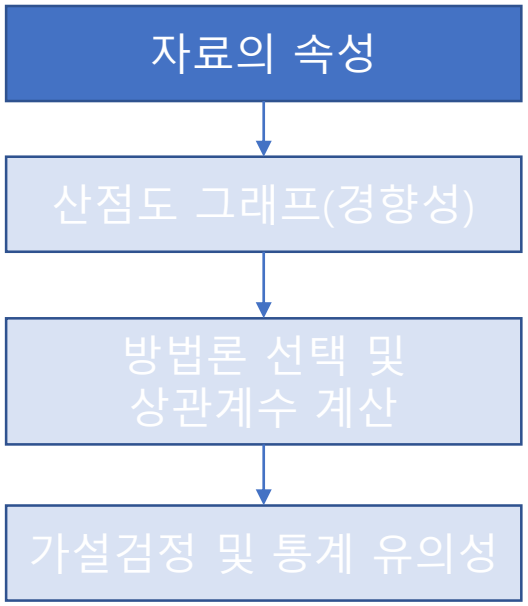
$$Cov(X,\ Y) = E((X - \mu)(Y - \nu))$$

$$\rho = \frac{Cov(X,\ Y)}{\sqrt{Var(X)\ Var(Y)}}, \quad -1 \leq \rho \leq 1$$

```
> ### 공분산과 상관관계
> ?cov # Covariance
> cov(price_pyung, floor_no) # 공분산: 평당 가격과 거래 아파트 층수
[1] 851.773
> cov(price_pyung, built_age) # 공분산: 평당 가격과 거래 아파트 연령
[1] -2142.107
> ?cor # Correlation
> cor(price_pyung, floor_no) # 상관계수: 평당 가격과 거래 아파트 층수
[1] 0.4785796
> cor(price_pyung, built_age) # 상관계수: 평당 가격과 거래 아파트 연령
[1] -0.7953403
> ?var()
> cor.1 <- cov(price_pyung, floor_no) / sqrt(var(price_pyung)*var(floor_no))
> cor.1 # 공분산과 분산을 이용하여 상관계수 구하기
[1] 0.4785796
> cor.2 <- cov(price_pyung, built_age) / sqrt(var(price_pyung)*var(built_age))
> cor.2 # 공분산과 분산을 이용하여 상관계수 구하기
[1] -0.7953403
```

상관관계 검정절차

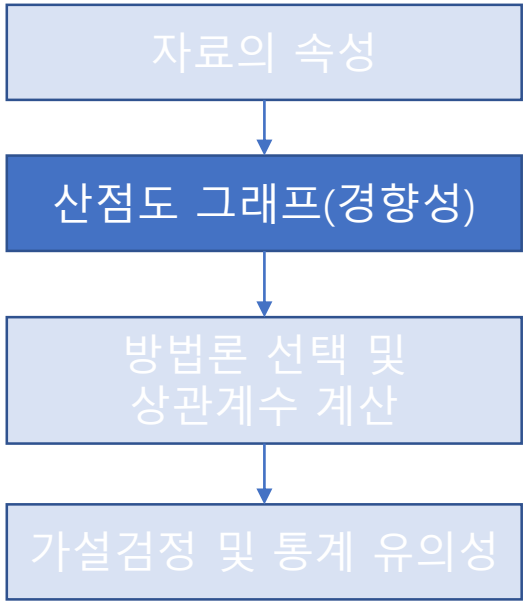
- 상관계수 분석절차
  - 아파트 평당 거래가격(price_pyung)과 아파트 건축년도(year_built)는 통계적으로 유의한 상관관계가 있는가? 그리고 그 관계는 어떠한가?



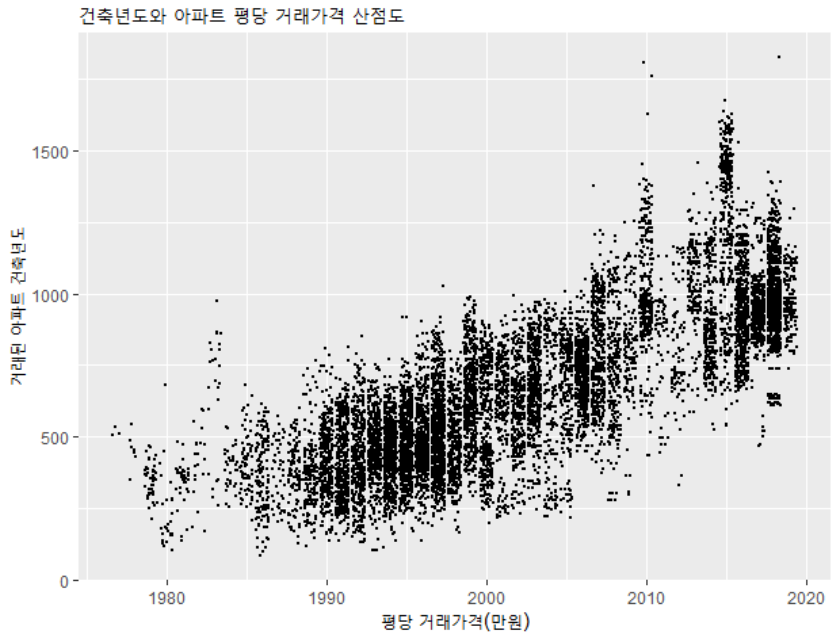

```
> #* 1. 자료의 속성 확인
> str(apt3) # 데이터 구조 확인하기
'data.frame':   13309 obs. of  14 variables:
 $ X          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ apt_price  : int  4800 4500 4000 4000 4000 4
 $ area_m2    : num  39.8 39.8 39.8 39.8 39.8 3
 $ floor_no   : int  2 6 7 9 3 7 5 9 8 5 ...
 $ year_built : int  1998 1998 1998 1998 1998 1
 $ ym_sale    : int  201811 201811 201812 20181
 $ day_sale   : int  6 13 10 10 10 10 10 10 10
 $ urban      : Factor w/ 3 levels "동","면","읍
 $ price_pyung: num  398 373 332 332 332 ...

> summary(apt3) # 결측치와 이상치 확인을 위한 요약통계
       X           apt_price       year_built      ym_sale          day_sale      urban       price_pyung
 Min.   :    1   Min.   :  1250   Min.   :1977   Min.   :201809   Min.   : 1.00   동:8848   Min.   :  85.4
 1st Qu.: 3328   1st Qu.:  7700   1st Qu.:1995   1st Qu.:201811   1st Qu.: 8.00   면:1332   1st Qu.: 444.4
 Median : 6655   Median : 13500   Median :2001   Median :201903   Median :15.00   읍:3129   Median : 627.3
 Mean   : 6655   Mean   : 14553   Mean   :2003   Mean   :201874   Mean   :15.63             Mean   : 666.
 3rd Qu.: 9982   3rd Qu.: 19500   3rd Qu.:2012   3rd Qu.:201906   3rd Qu.:23.00             3rd Qu.: 866.
 Max.   :13309   Max.   :108000   Max.   :2019   Max.   :201908   Max.   :31.00             Max.   :1825.
```

상관관계 검정절차

- 상관계수 분석절차
  - 아파트 평당 거래가격(price_pyung)과 아파트 건축년도(year_built)는 통계적으로 유의한 상관관계가 있는가? 그리고 그 관계는 어떠한가?

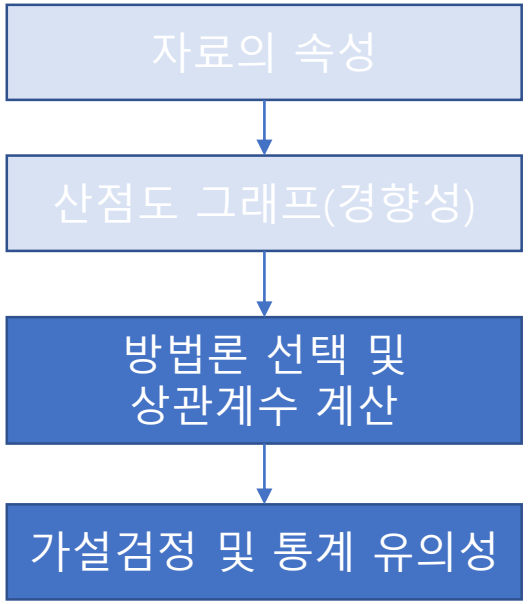



```
> #* 2.산점도 그래프로 경향성 확인하기
> ggplot(apt3, aes(year_built, price_pyung)) + # 그래픽 x, y 틀 설정
+   geom_jitter(cex=0.3) + # jitter 형태의 산점도 그래프: 점의 크기는 0.3배
+   labs(title = "건축년도와 아파트 평당 거래가격 산점도", # 제목
+        x = "평당 거래가격(만원)", # 축제목
+        y = "거래된 아파트 건축년도")
```



상관관계 검정절차

- 상관계수 분석절차
  - 아파트 평당 거래가격(price_pyung)과 아파트 건축년도(year_built)는 통계적으로 유의한 상관관계가 있는가? 그리고 그 관계는 어떠한가?




```
> #* 3. 분석방법론 선택과 상관계수 구하기
> cor(year_built, price_pyung, # 검정할 두 변수 선택
+           method="pearson") # 방법론 선택: 두 변수 모두 양적 변수
[1] 0.7953403
> #* 4. 가설검정 및 통계적 유의성 진단
> ?cor.test
> cor.test(year_built, price_pyung, # 검정할 두 변수 선택
+                     method="pearson") # 방법론 선택: 두 변수 모두 양적 변수

        Pearson's product-moment correlation

data:  year_built and price_pyung
t = 151.36, df = 13307, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7890122 0.8014996
sample estimates:
      cor
0.7953403
```

상관관계 검정방법

- 상관계수 검정방법
  - 양측 및 단측검정
    - 양측 검정: H0= no association(==0)
    - 왼쪽 단측 검정: H0= negative association(<0)
    - 오른쪽 단측 검정: H0=positive association(>0)
  - 상관계수 산출방법
    - 피어슨 상관계수
    - 스피어만 상관계수
    - 켄달 상관계수
  - 실습

```r
cor.test(built_age, price_pyung, # 두 변수 상관관계 분석
         method="pearson", # method="pearson"
         alternative="two.sided") # no association(==0)

cor.test(built_age, price_pyung, # 두 변수 상관관계 분석
         method="pearson", # method="pearson"
         alternative="less") # negative association(<0)

cor.test(built_age, price_pyung, # 두 변수 상관관계 분석
         method="pearson", # method="pearson"
         alternative="greater" ) # corresponds to positive
```

```r
cor.test(built_age, price_pyung,
         method="spearman", # method="spearman"
         exact=FALSE) # 동일순위(ties)가 있을 경우 정확한 p-값 산출하지 않도록 설정

cor.test(built_age, price_pyung,
         method="kendall") # 상관관계 분석method="kendall"
```

다중 상관관계와 산점도 매트릭스

- 3개 이상의 변수들간의 상관계수 구하기

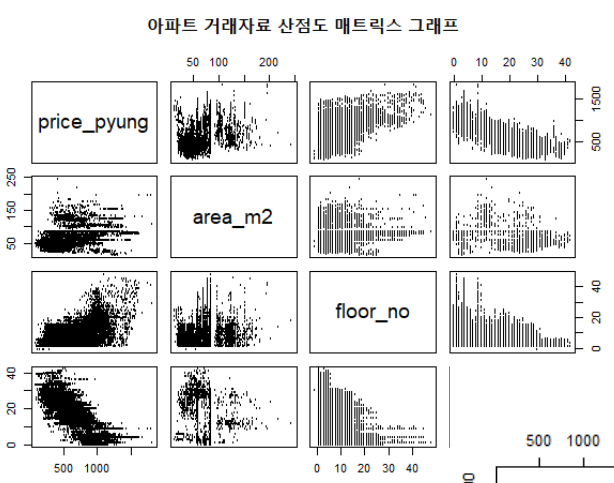
```
> #### 3개 이상 다중 변수 상관관계 계수 매트릭스
> v.cor <- apt3 %>% select(price_pyung, area_m2, floor_no, built_age
> cor(v.cor, # 해당 객체로 다중 상관관계 매트릭스 구하기
+                 method = "pearson") # method = c("pearson", "kendal
             price_pyung    area_m2   floor_no  built_age
price_pyung   1.0000000  0.3320946  0.4785796 -0.7953403
area_m2       0.3320946  1.0000000  0.1871514 -0.2920381
floor_no      0.4785796  0.1871514  1.0000000 -0.3974726
built_age    -0.7953403 -0.2920381 -0.3974726  1.0000000
> round(cor(v.cor), 3) # 소수점 세자리까지 반올림하여 결과 반환
             price_pyung area_m2 floor_no built_age
price_pyung        1.000   0.332    0.479    -0.795
area_m2            0.332   1.000    0.187    -0.292
floor_no           0.479   0.187    1.000    -0.397
built_age         -0.795  -0.292   -0.397     1.000
```

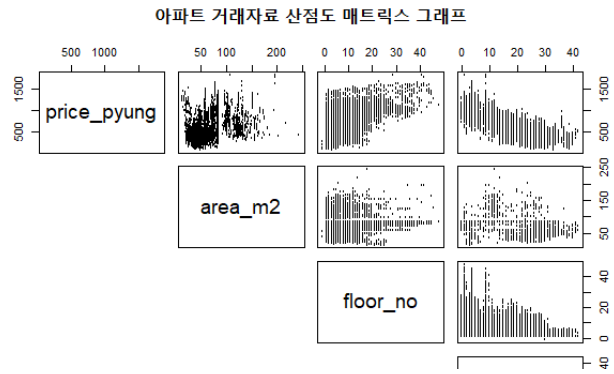
상관관계 분석

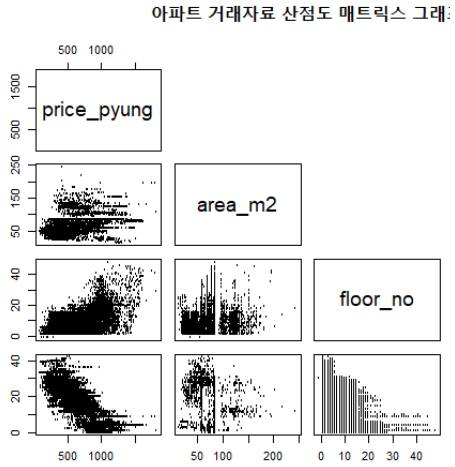
다중 상관관계와 산점도 매트릭스

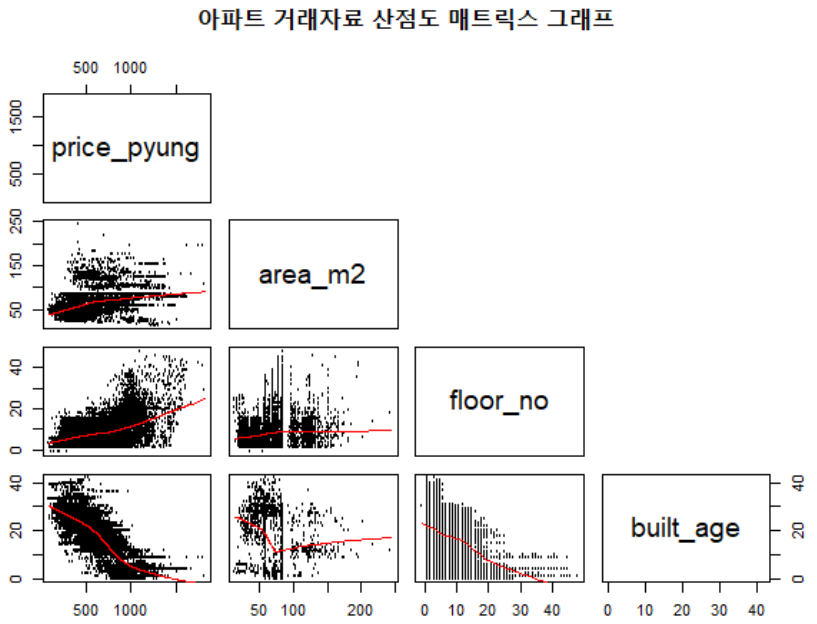
- 기본 산점도 매트릭스 작성

```
> ##### 다중 기본 산점도 매트릭스
> ?pairs() # Scatterplot Matrices
> pairs(formula =  ~ price_pyung + area_m2 + floor_no + built_age, # 산점도 적용 변수들
+       data=apt3, # 데이터 객체
+       cex=0.2, # 점의 크기는 0.2배
+       main="아파트 거래자료 산점도 매트릭스 그래프") # 제목
> pairs(formula =  ~ price_pyung + area_m2 + floor_no + built_age, # 산점도 적용 변수들
+       data=apt3, # 데이터 객체
+       lower.panel = NULL, # 대각선 아래 매트릭스 생략
+       cex=0.2, # 점의 크기는 0.2배
+       main="아파트 거래자료 산점도 매트릭스 그래프") # 제목
> pairs(formula =  ~ price_pyung + area_m2 + floor_no + built_age, # 산점도 적용 변수들
+       data=apt3, # 데이터 객체
+       upper.panel = NULL, # 대각선 위 매트릭스 생략
+       cex=0.2, # 점의 크기는 0.2배
+       main="아파트 거래자료 산점도 매트릭스 그래프") # 제목
> pairs(formula =  ~ price_pyung + area_m2 + floor_no + built_age, # 산점도 적용 변수들
+       data=apt3, # 데이터 객체
+       upper.panel = NULL, # 대각선 위 매트릭스 생략
+       cex=0.2, # 점의 크기는 0.2배
+       main="아파트 거래자료 산점도 매트릭스 그래프", # 제목
+       panel=panel.smooth) # 패널에 추세선을 그릴
```







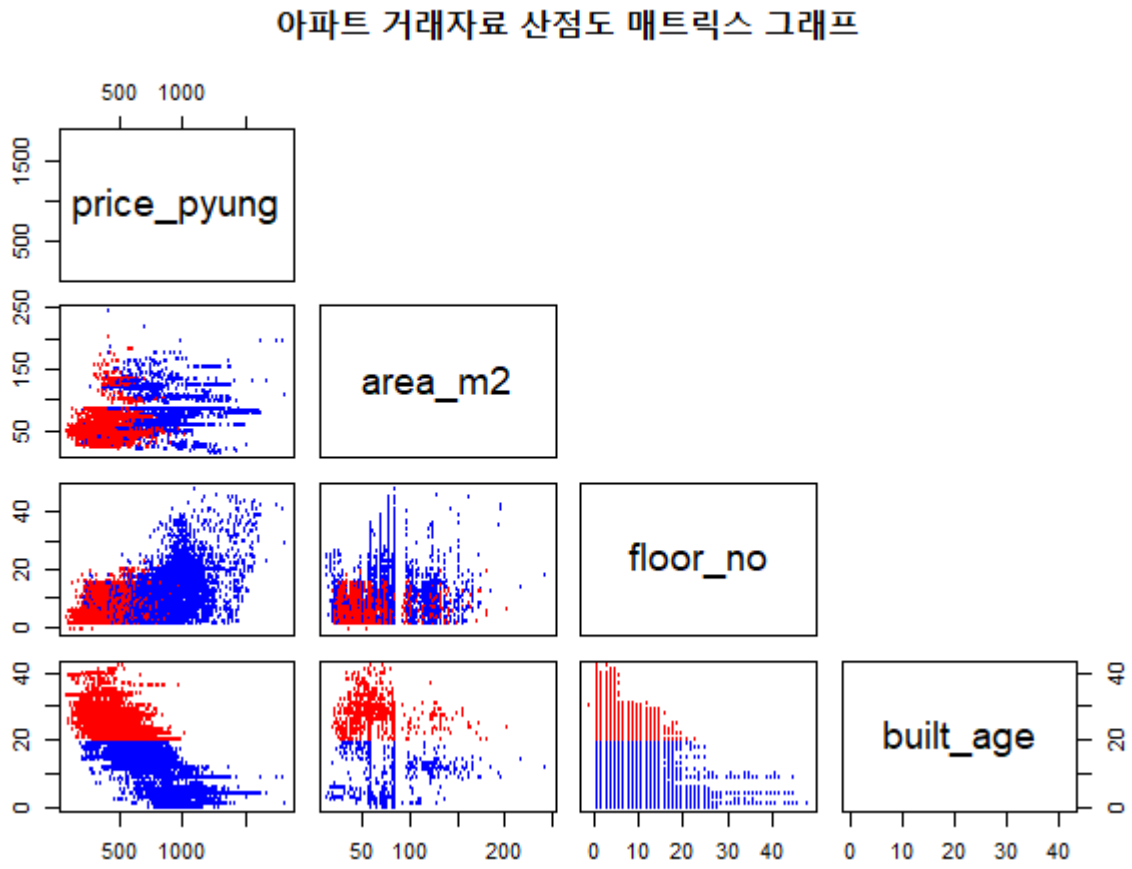


상관관계 분석

다중 상관관계와 산점도 매트릭스

- 기본 산점도 매트릭스 작성하기

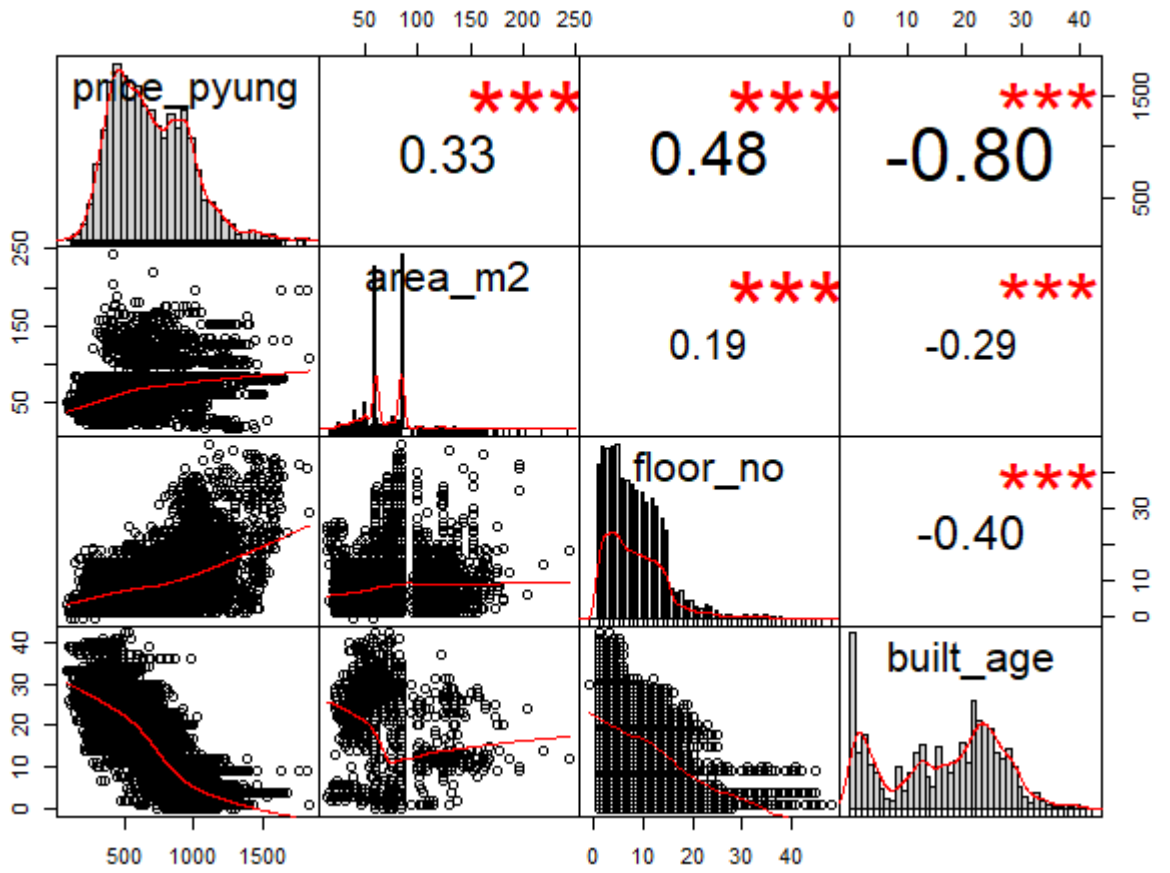
```
> my_cols <- c("blue", "red")  # 2가지 색상 지정
> pairs(formula =  ~ price_pyung + area_m2 + floor_no + built_age, # 산점도 적용 변수들
+       data=apt3, # 데이터 객체
+       col = my_cols[apt3$yr_built2],  # 건축년도 2000이전과 이후 점색 다르게 설정
+       upper.panel = NULL, # 대각선 위 매트릭스 생략
+       cex=0.2, # 점의 크기는 0.2배
+       main="아파트 거래자료 산점도 매트릭스 그래프") # 제목
```



다중 상관관계와 산점도 매트릭스

- 기타 산점도 매트릭스 작성하기
  - "PerformanceAnalytics"패키지 활용

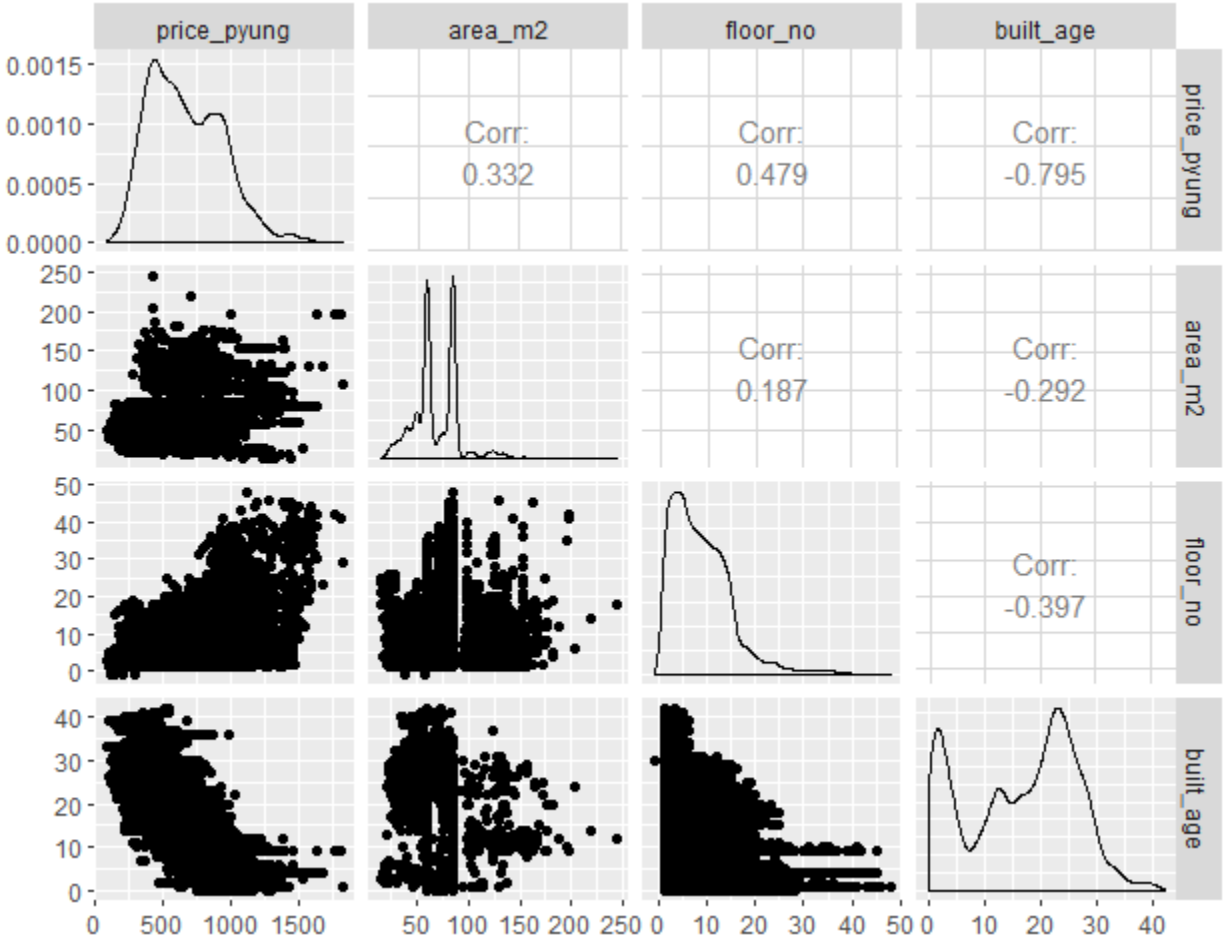
```
#*** 기타 다중 산점도 매트릭스 그래프
install.packages("PerformanceAnalytics")
library(PerformanceAnalytics)

?chart.Correlation() # On top the (absolute) value of the correlatio
chart.Correlation(v.cor, # 산점도 매트릭스를 사용할 객체(또는 함수)
                  method = "pearson", #  method = c("pearson", "kend
                  histogram=TRUE, # 우상단 부에 히스토 그램
                  cex = 0.2) # 점의 크기는 0.2배
```



다중 상관관계와 산점도 매트릭스

- 기타 산점도 매트릭스 작성하기
  - "GGally" 패키지 활용

```
> library(GGally)
> ?ggpairs() # A ggplot2 generalized pairs plot
> ggpairs(v.cor)
```



연습문제 04

- 다음 중 피어슨 상관계수로 상관관계 분석을 하는 것이 옳지 않은 것을 구하시오.
  - ① 아버지의 혈압과 아들의 혈압의 관계
  - ② 경제적 수입과 삶의 질의 정도(리커트 척도)와의 관계
  - ③ 키와 체질량 지수와의 관계
  - ④ 통증(5점 scale)과 암의 stage(5점 scale)와의 관계
  - ⑤ 입원기간과 수술시간의 관계
  - ⑥ 혈압과 연령의 관계
  - ⑦ 중간고사 순위와 기말고사의 성적 순위와의 관계
  - ⑧ 아파트 평당 거래가격과 거래된 아파트 단지의 건축년도와의 관계
  - ⑨ 지하철역까지의 도보거리와 아파트 가격과의 관계

연습문제 05

- 최근 거래되어진 아파트의 건축연령(built_age)과 평당 거래가격(price_pyung)과의 상관분석을 실시하고, 다음의 질문에 답하시오.

  - 이 상관관계 검정에서 아파트 평당 거래가격은 연령이 오래될수록 감소한다라고 귀무가설을 설정할 때, 인수 alternative는 무엇으로 설정하여야 하는가?

  - 음의 상관성을 귀무가설로 설정하여 상관관계 검정을 하였을 때의 t-값과 상관계수 값을 구하시오?

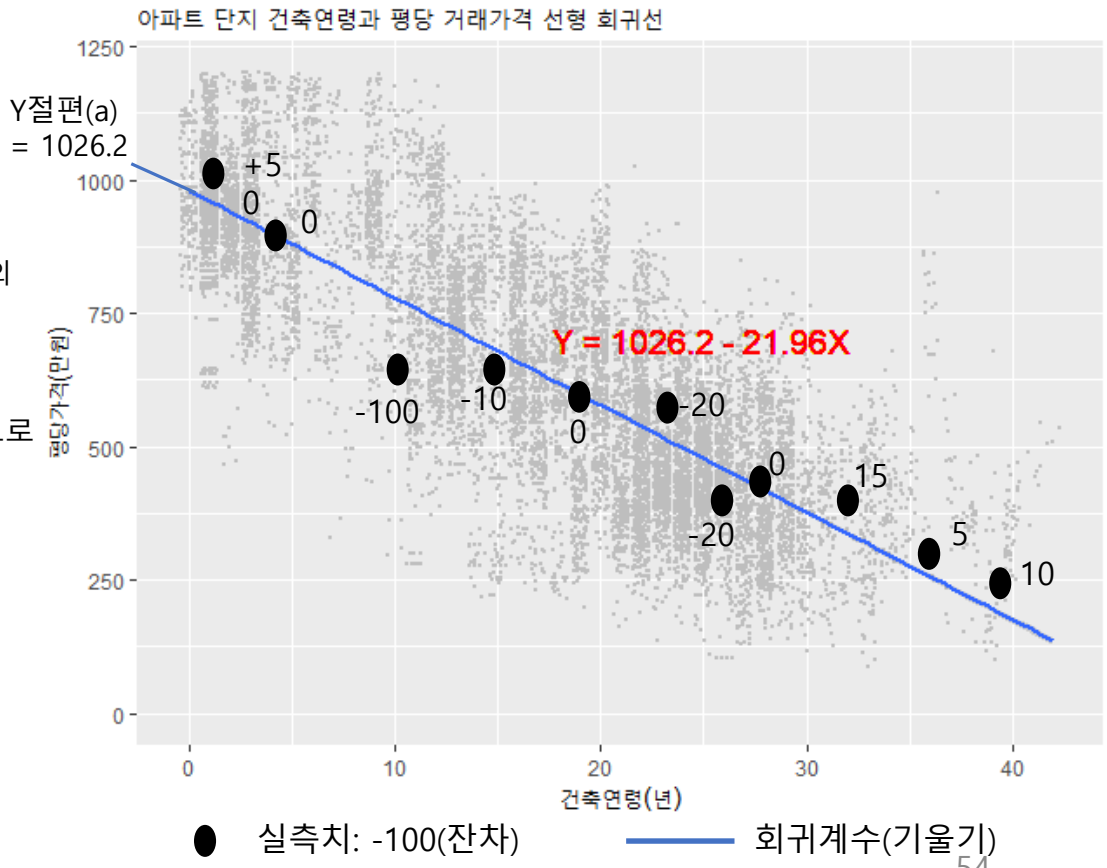
회귀와 예측의 개요

- 회귀분석(regression analysis)
  - 원인과 결과가 되는 두 변수의 선형 상관성을 기본으로 하여 1차 선형 방정식으로 관계를 일반화하는 방법
- 회귀(regression)
  - 평균으로의 회귀현상
  - 두 변수의 관계가 선형관계의 평균으로 돌아간다는 의미
- 선형회귀모형(linear regression model)
  - 최소자승법(최소제곱법)과 기울기(회귀계수)
    - 두 변수의 회귀직선의 예측치(회귀식에 의한 추정치)와 관측치(실측치)의 차이의 제곱의 합이 최소가 되는 직선을 구하는 방법
- 결과 활용
  - (통계적 유의성) 원인이 되는 설명변수가 결과로 나타나는 종속변수에 통계적으로 유의한 영향력을 미치는가?
  - (방향성) 설명변수와 종속변수의 기울기의 방향
    - 아파트 평당 거래가격은 오래된 아파트 일수록 낮다
  - (중요도, 영향력) 어느 정도의 크기로 영향을 주는가?(회귀계수의 크기)
    - 1년 더 아파트가 오래되면 평당 21.96만원 거래가격이 감소한다.
  - (상대적 중요도) 설명변수들 중 어떤 변수들이 보다 더 상대적으로 중요할 까?
    - 표준화 회귀계수의 크기 비교
  - (예측성) 설명변수가 한 단위 증가할 때 종속변수는 얼마나 증가할 것인가?
    - 나의 집(아파트)가 10년 노후화 되면 219.6백만원 평당 가격이 낮아질 것이다.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

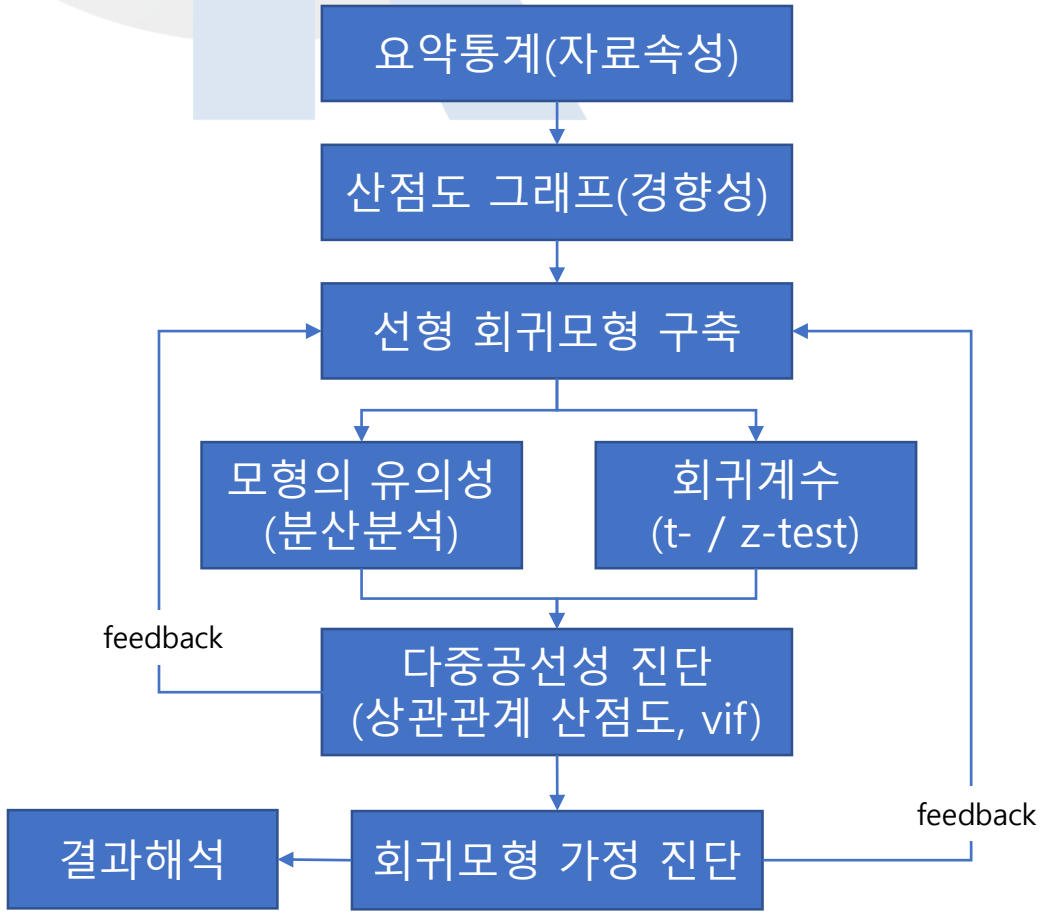
$$\min \sum e_i^2 = \min (Y_i - \hat{Y}_i)^2$$

잔차의 제곱합 최소직선



회귀와 예측의 개요

- 회귀분석 절차



① 요약통계량 확인
- 자료형태 확인: 변수속성
- 이상치 및 입력오류 확인: 평균, 표준편차, 최솟값, 최댓값 등
- 결측치(na) 확인

② 산점도 그래프 확인
- 선형, 비선형 형태와 관계의 방향성 확인

③ 선형회귀모형 구축

④ 모형 및 설명변수 회귀계수 유의성 진단
- 모형진단: 분산분석
- 회귀계수 진단: t 또는 z-test
- 모형간 비교(최종모형 선택): 수정결정계수, AIC, BIC, 분산분석 등

⑤ 다중공선성 진단
- 상관관계 매트릭스, 분산팽창계수 등

⑥ 회귀모형 가정 진단
- 선형성: 설명변수와 종속변수의 선형 관계 충족
- 독립성: 다중 회귀분석시 진단. 설명변수간 다중공선성 진단 활용
- 등분산성: 잔차의 정규성
- 정규성: 잔차의 정규성

⑦ 결과해석
- 모형 결정계수(adj. R-squared), 회귀계수(유의성, 방향성, 크기), 표준화 회귀계수 등 활용

회귀와 예측의 개요

- 회귀분석 함수
  - ?lm # Fitting Linear Models
    - lm(formula, data, subset, ……)

- 선형회귀모형 함수 기본 형식

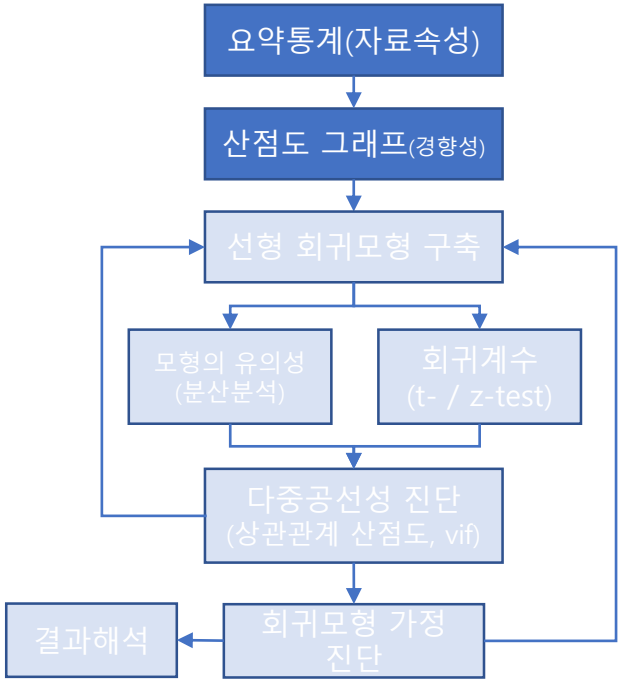
기호	내용
Y ~ X	response variable~predictor variables
lm(Y ~ X)	Yi = β0 + β1Xi + ei
lm(Y ~ X + Z)	Yi = β0 + Xiβ1 + Ziβ2 + ei

기호	예	설명
+	+ X	include this variable
-	- X	delete this variable
:	X : Z	include the interaction between these variables
*	X * Y	include these variables and the interactions between them
	X	Z
^	(X + Z + W)^3	include these variables and all interactions up to three way
I	I(X*Z)	as is: include a new variable consisting of these variables multiplied
1	X - 1	intercept: delete the intercept (regress through the origin

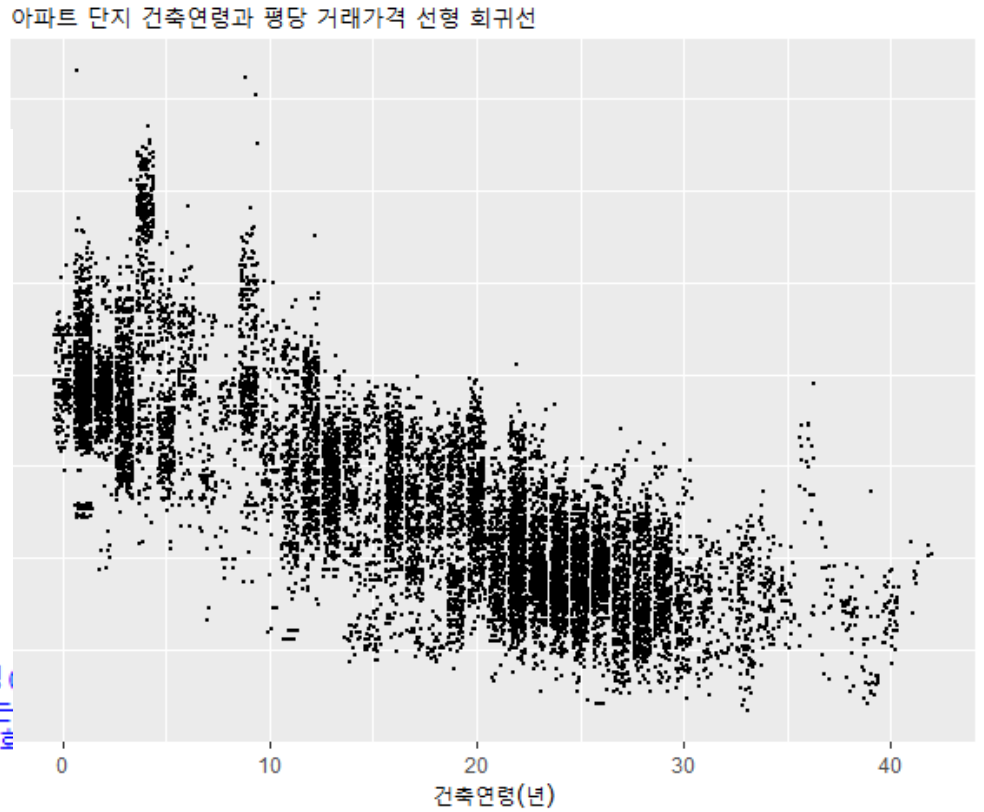
단순 선형회귀모형

- 단순 선형회귀모형: 한 개의 종속변수와 한 개의 설명변수
  - 아파트의 건축연령(built_age)은 평당 아파트 거래가격(price_pyung)에 어떠한 영향을 주는가?

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



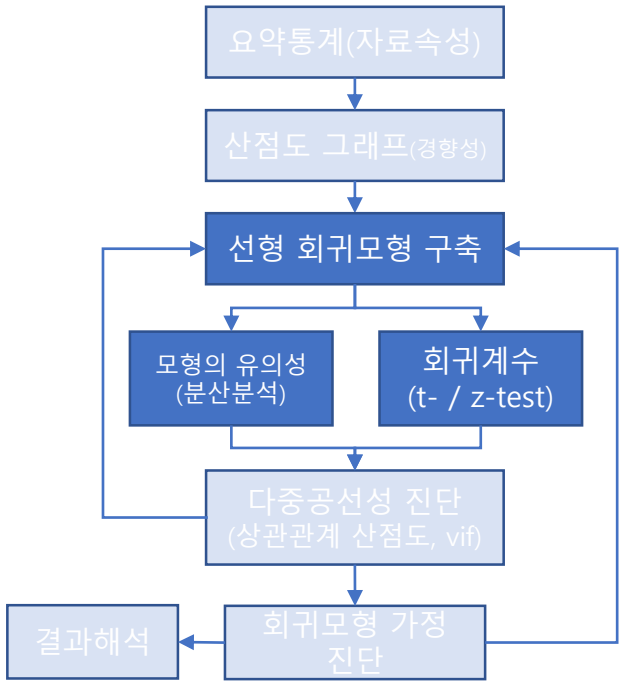

```
> ### 단순 선형회귀분석: 분석절차
> #* 1. 요약 통계
> x <- select(apt3, price_pyung, built_age)
> str(x)
'data.frame':   13309 obs. of  2 variables:
 $ price_pyung: num  398 373 332 332 332 ...
 $ built_age  : int  21 21 21 21 21 21 21 21 21 21 ...
> summary(x)
  price_pyung        built_age
 Min.   :  85.48   Min.   : 0.00
 1st Qu.: 444.44   1st Qu.: 7.00
 Median : 627.35   Median :18.00
 Mean   : 666.15   Mean   :16.39
 3rd Qu.: 866.05   3rd Qu.:24.00
 Max.   :1825.47   Max.   :42.00
> #* 2. 산점도 그래프
> c <- ggplot(apt3, # 아파트 거래가격 데이터
+             aes(built_age, price_pyung)) # 그래프 셋팅
> c + geom_jitter(cex=0.5, col="black") + # 중첩되지 않는
+   labs(title = "아파트 단지 건축연령과 평당 거래가격 선형 회
+        x = "건축연령(년)", y="평당가격(만원)")
```



단순 선형회귀모형

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- 단순 선형회귀모형: 한 개의 종속변수와 한 개의 설명변수
  - 아파트의 건축연령(built_age)은 평당 아파트 거래가격(price_pyung)에 어떠한 영향을 주는가?




```
> #* 3. 모형구축
> lm <- lm(price_pyung ~ built_age # 종속변수 ~ 설명변수
+          , data=apt3) # 데이터 객체
> #* 4. 모형 및 회귀계수 진단
> summary(lm) # lm객체에 저장된 회귀분석의 결과 호출

Call:
lm(formula = price_pyung ~ built_age, data = apt3)

Residuals:
    Min      1Q  Median      3Q     Max
-542.53 -107.28  -15.39   97.23  980.84

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1026.1592     2.7768   369.5   <2e-16 ***
built_age    -21.9605     0.1451  -151.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 165.3 on 13307 degrees of freedom
Multiple R-squared:  0.6326,    Adjusted R-squared:  0.6325
F-statistic: 2.291e+04 on 1 and 13307 DF,  p-value: < 2.2e-16
```

- 절편(상수, intercept)
  - 어떠한 영향요인이 없을 경우에서의 평균값
  - 독립변수 x들이 모두 0일 때의 값
- 회귀계수
  - 회귀선의 기울기
  - 독립변수가 한 단위 변화함에 따라 종속변수가 미치는 영향력 크기
- 절편 및 회귀계수 진단
  - t-test 또는 z-test
  - H0: 값은 0이다.

- **결정계수**(決定係數, coefficient of determination, R^2)
  - 추정한 선형 모형이 주어진 자료에 적합한 정도를 재는 척도
  - 반응 변수의 변동량 중에서 설명가능한 부분의 비율
  - 단순회귀분석을 하는 경우에는 일반 결정계수를 사용

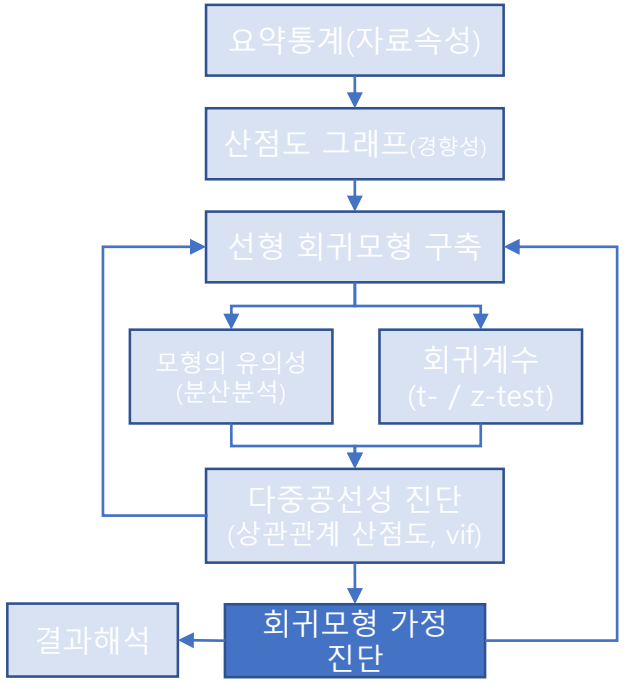
$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})}{\sum(y_i - \bar{y})^2} = \frac{\text{회귀선에 의해 설명되는 변동}}{\text{전체 변동}}$$

- **조정(수정)된 결정계수**
  - 다중회귀모형에서 사용
  - 추가되는 독립변수의 과대영향을 조정함
- 모형의 결정계수 진단
  - F-분포와 통계량

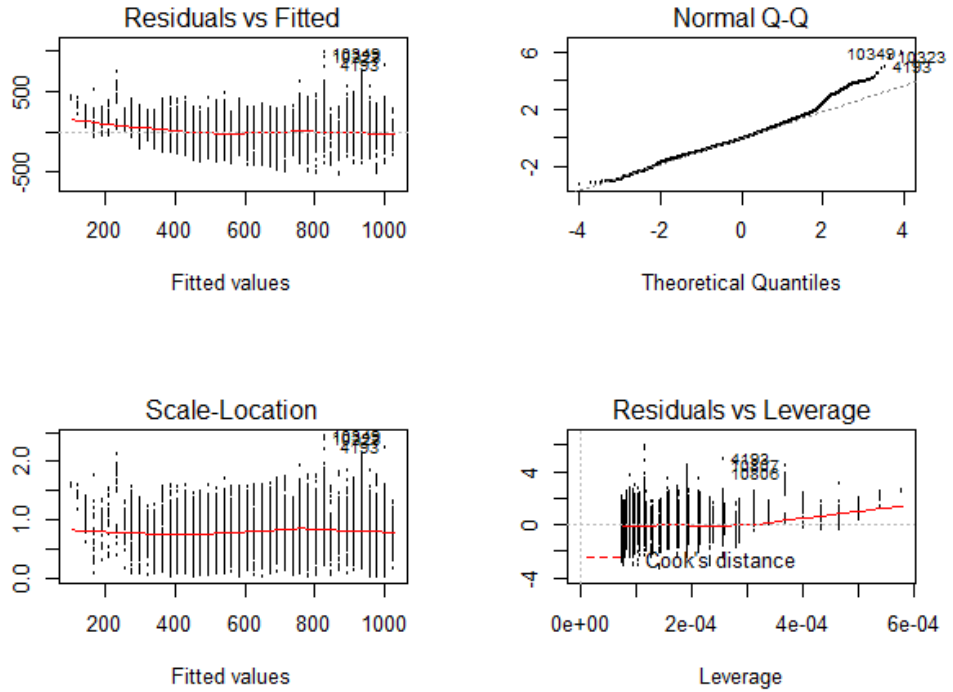
단순 선형회귀모형

- 단순 선형회귀모형: 한 개의 종속변수와 한 개의 설명변수
  - 아파트의 건축연령(built_age)은 평당 아파트 거래가격(price_pyung)에 어떠한 영향을 주는가?

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



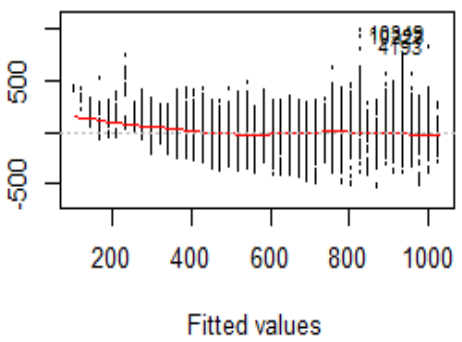

```
> #* 5. 다중공선성 진단: 단순 선형회귀모형에는 적용 안됨
> #* 6. 회귀모형 가정 진단
> par(mfrow = c(2,2)) # 한 화면에 2*2 그래프 창 설정
> plot(lm, cex=0.3) # 선형회귀 진단 그래프
```



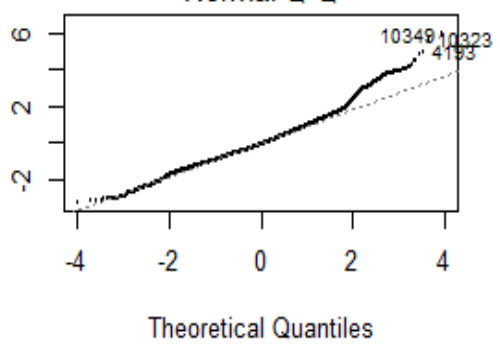
단순 선형회귀모형

- 기본가정 진단

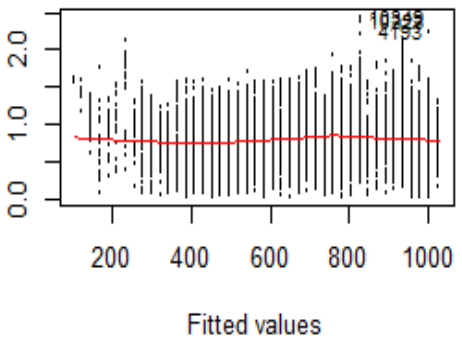
① Residuals vs Fitted



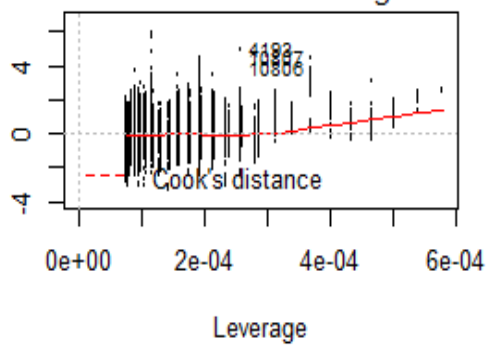
② Normal Q-Q



③ Scale-Location



④ Residuals vs Leverage

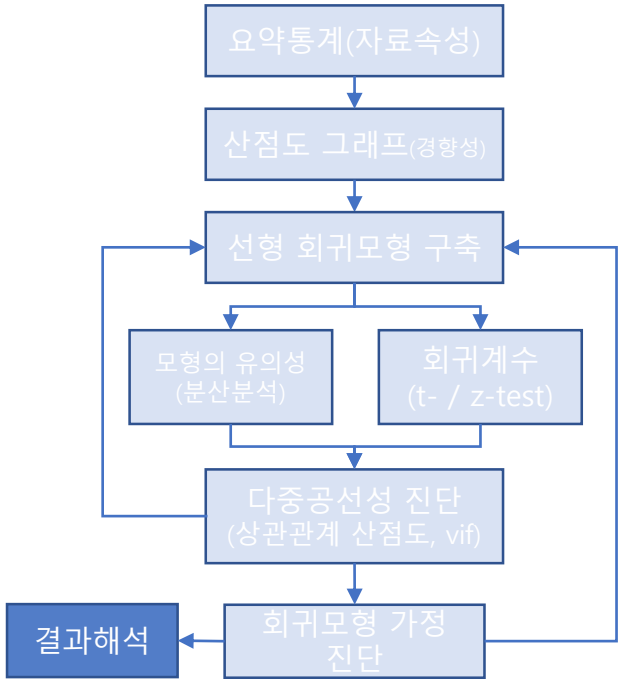


- 정규성(Normality): ②
  - 잔차의 평균은 0, normal Q-Q plot: 표준화된 잔차의 probability plot
- 독립성(Independence)
  - 다중공선성: 단순회귀분석에서는 적용 안 됨
- 선형성(Linearity): ①
  - 잔차와 예측값의 체계적인 연관성이 없어야 됨
- 등분산성(Homoscedasticity): ③
  - 분산이 일정. 무작위 잡음(random noise). 수평선 주의 random band 형태
- 기타: ④
  - 관측치의 영향력(leverage)
    - 큰지레점(high leverage point)
  - 이상치(outlier): 아주 큰 잔차
    - 영향관측치(influential observation)

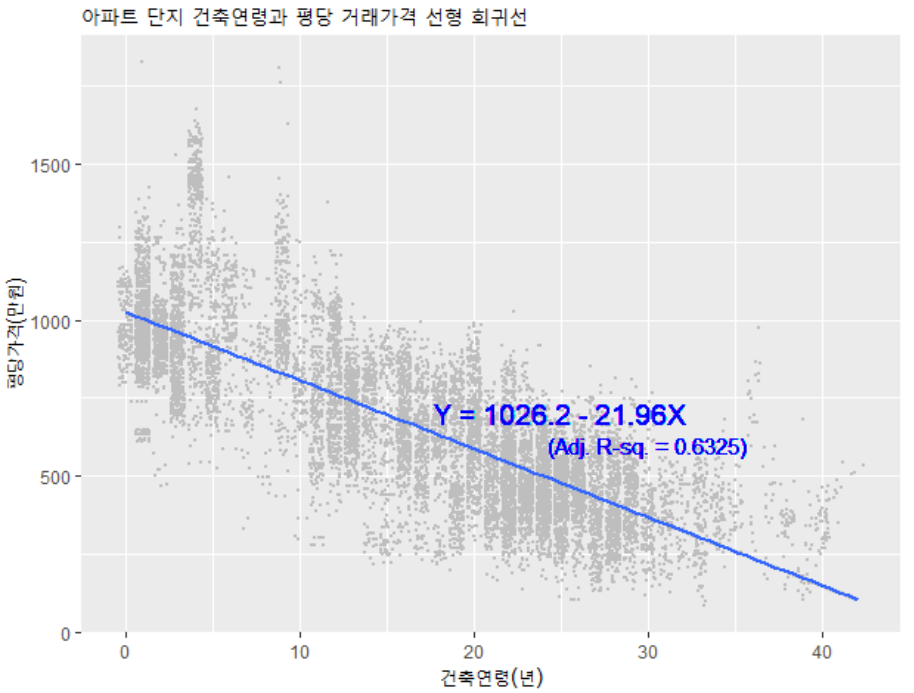
단순 선형회귀모형

- 단순 선형회귀모형: 한 개의 종속변수와 한 개의 설명변수
  - 아파트의 건축연령(built_age)은 평당 아파트 거래가격(price_pyung)에 어떠한 영향을 주는가?

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



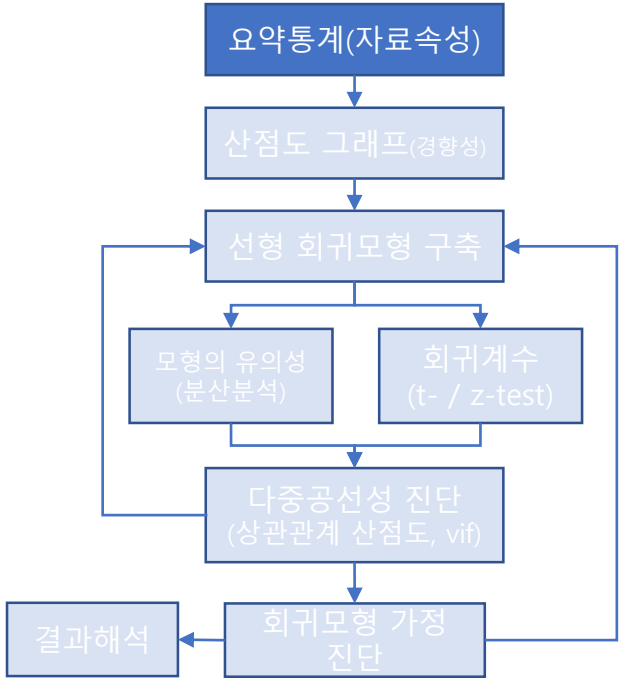

```
> #* 7. 결과해석
> c <- ggplot(apt3, # 아파트 거래가격 데이터
+             aes(built_age, price_pyung)) # 그래프 셋팅(x = 아파트
> c + geom_jitter(cex=0.2, col="gray") + # 중첩되지 않는 점 그래프
+   geom_smooth(method = "lm", level =0.95) + # lm 예측 선형회귀식
+   geom_text(aes(x=25,y=700,label = "Y = 1026.2 - 21.96X"),
+             color="blue",
+             size=5) +
+   geom_text(aes(x=30,y=600,label = "(Adj. R-sq. = 0.6325)"),
+             color="blue",
+             size=4) +
+   labs(title = "아파트 단지 건축연령과 평당 거래가격 선형 회귀선",
+        x = "건축연령(년)", y="평당가격(만원)")
```



다중 선형회귀 모형

- 범주형 변수
  - 요인(factor)으로 되어 있으면 R에서는 자동으로 더미변수를 생성함
  - 문자로 되어 있는 경우, 요인으로 변환(as.factor())하면 됨
  - 분석결과는 준거(기준)이 되는 변수에 비하여 회귀계수만큼 얼마의 차별적인 영향이 있다고 표현

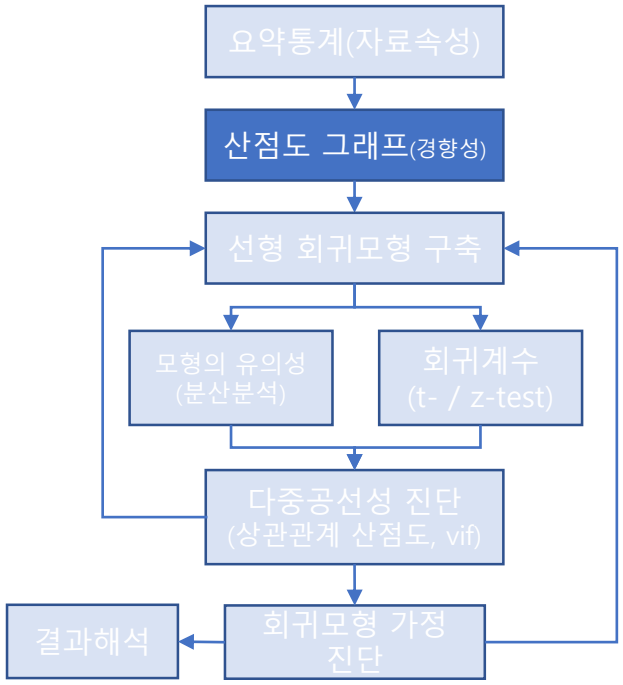
- 아파트의 전용면적(area_m2), 층수(floor_no), 건축년령(built_age), 그리고 거래지역 유형(urban2)은 평당 실거래 가격에 얼마만큼의 가격 결정요인이 될까?



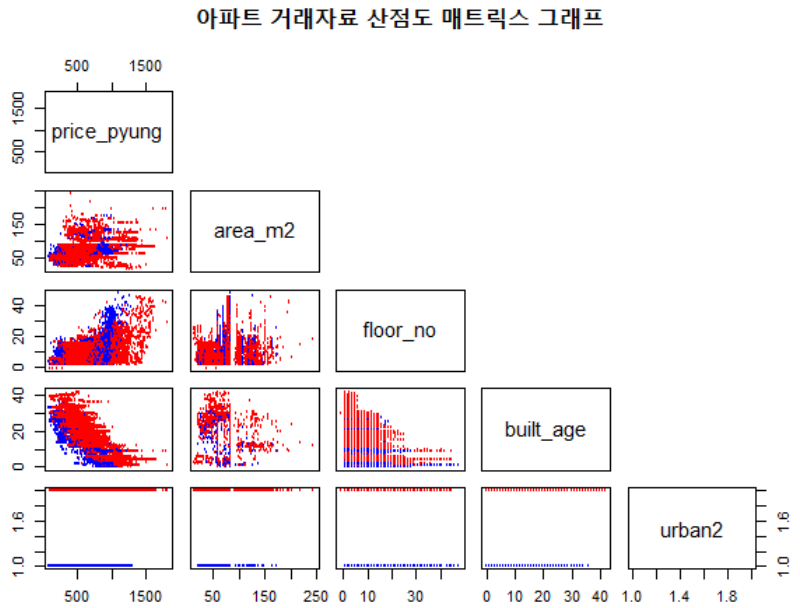

```
> #* 1. 요약 통계
> x <- select(apt3, price_pyung, area_m2, floor_no, built_age, urban2) # 다중선형회귀모형 선택 변수들 선정
> str(x)
'data.frame':   13309 obs. of  5 variables:
 $ price_pyung: num  398 373 332 332 332 ...
 $ area_m2    : num  39.8 39.8 39.8 39.8 39.8 39.8 39.8 39.8 39.8 39.8 ...
 $ floor_no   : int  2 6 7 9 3 7 5 9 8 5 ...
 $ built_age  : int  21 21 21 21 21 21 21 21 21 21 ...
 $ urban2     : Factor w/ 2 levels "농촌","도시": 1 1 1 1 1 1 1 1 1 1 ...
> summary(x)
  price_pyung        area_m2         floor_no        built_age      urban2
 Min.   :  85.48   Min.   : 14.98   Min.   :-1.000   Min.   : 0.00   농촌:4461
 1st Qu.: 444.44   1st Qu.: 58.93   1st Qu.: 4.000   1st Qu.: 7.00   도시:8848
 Median : 627.35   Median : 60.00   Median : 8.000   Median :18.00
 Mean   : 666.15   Mean   : 68.92   Mean   : 8.835   Mean   :16.39
 3rd Qu.: 866.05   3rd Qu.: 84.89   3rd Qu.:12.000   3rd Qu.:24.00
 Max.   :1825.47   Max.   :244.07   Max.   :48.000   Max.   :42.00
```

다중 선형회귀 모형

- 아파트의 전용면적(area_m2), 층수(floor_no), 건축년령(built_age), 그리고 거래지역 유형(urban2)은 평당 실거래 가격에 얼마만큼의 가격 결정요인이 될까?

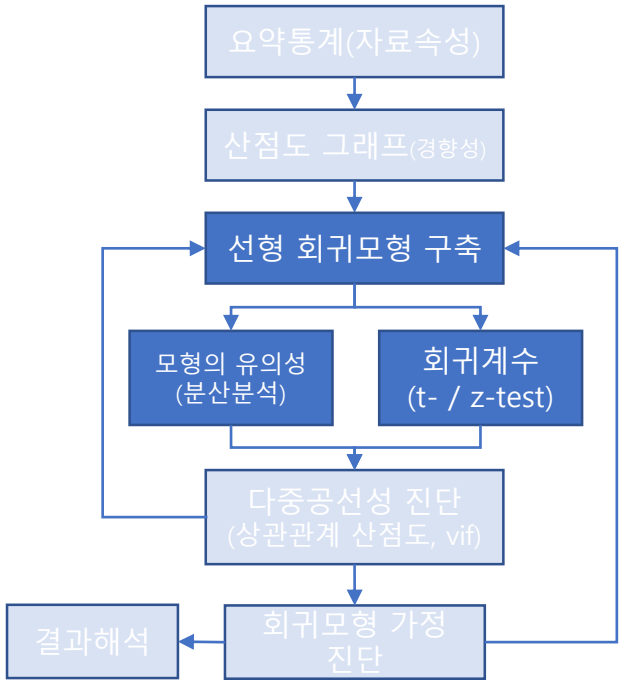



```
> #* 2. 산점도 그래프
> my_cols <- c("blue", "red")  # 2가지 색상 지정
> pairs(x, # 산점도 적용 변수들
+       data=apt3, # 데이터 객체
+       col = my_cols[apt3$urban2],  # 거래지역 유형별 점색 다르게 설정
+       upper.panel = NULL, # 대각선 위 매트릭스 생략
+       cex=0.2, # 점의 크기는 0.2배
+       main="아파트 거래자료 산점도 매트릭스 그래프") # 제목
```



다중 선형회귀 모형

- 아파트의 전용면적(area_m2), 층수(floor_no), 건축년령(built_age), 그리고 거래지역 유형(urban2)은 평당 실거래 가격에 얼마만큼의 가격 결정요인이 될까?

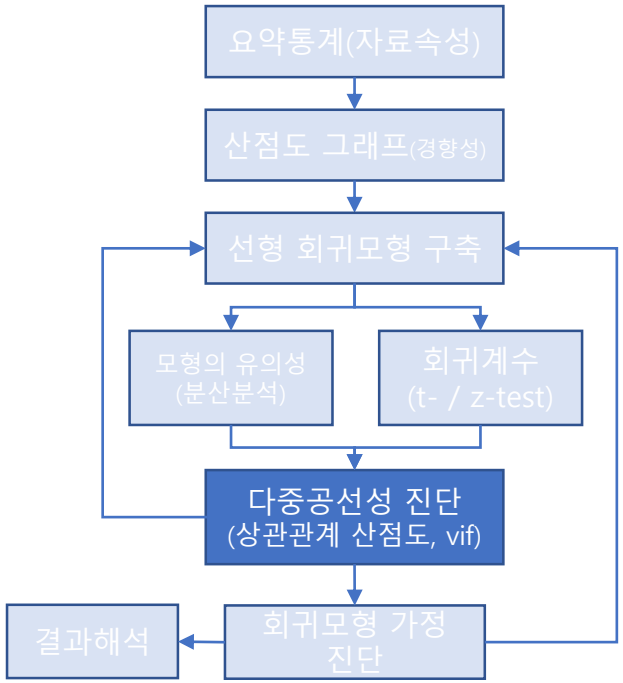



```
> #* 3. 모형구축
> lm_2 <- lm(price_pyung ~
+              area_m2 + floor_no + built_age + urban2)
> #* 4. 모형진단
> summary(lm_2)

Call:
lm(formula = price_pyung ~ area_m2 + floor_no + built_age + urban2)

Residuals:
    Min      1Q  Median      3Q     Max
-592.05  -95.65   -3.12   85.25  736.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 803.1352     5.4373  147.71   <2e-16 ***
area_m2       0.7448     0.0543   13.72   <2e-16 ***
floor_no      7.0423     0.2013   34.98   <2e-16 ***
built_age   -21.8320     0.1434 -152.25   <2e-16 ***
urban2도시  161.5030     2.6773   60.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.4 on 13304 degrees of freedom
Multiple R-squared:  0.7425,    Adjusted R-squared:  0.7424
F-statistic:  9590 on 4 and 13304 DF,  p-value: < 2.2e-16
```

- 결과해석
  - 모형
    - 결정계수와 F-통계량
  - 회귀계수
    - 절편(intercept)
    - 연속회귀계수와 t-통계량
    - (명목)더미변수와 t-통계량

다중 선형회귀 모형

- 아파트의 전용면적(area_m2), 층수(floor_no), 건축년령(built_age), 그리고 거래지역 유형(urban2)은 평당 실거래 가격에 얼마만큼의 가격 결정요인이 될까?

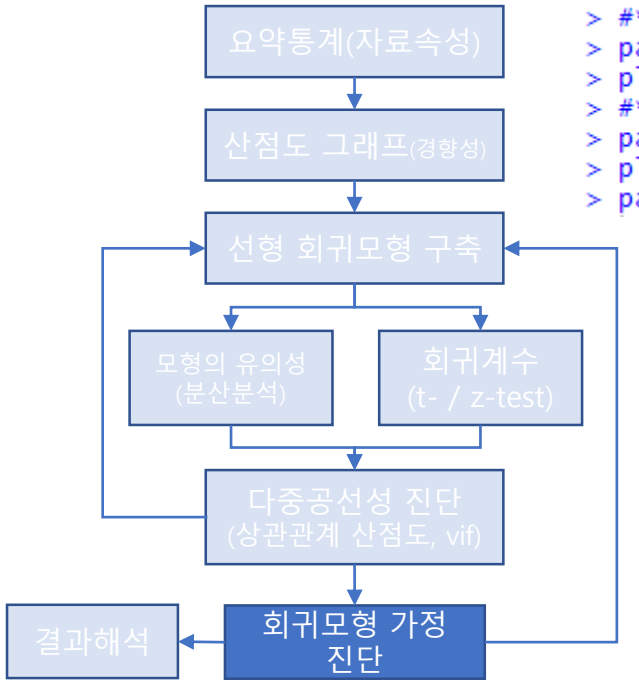


- 다중공선성(Multicollinearity)
  - 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제
  - 독립성 가정 위배
  - VIF 즉, 분산팽창계수가 10 이상(또는 5이상) 일때 다중공선성이 존재한다고 판단

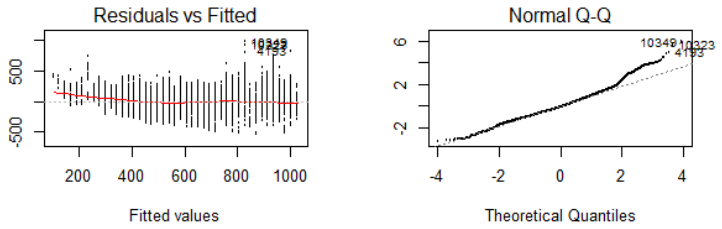
```
> library(car)
필요한 패키지를 로딩중입니다: carData

다음의 패키지를 부착합니다: 'car'

The following object is masked from 'package:laws

    levene.test

The following object is masked from 'package:dply

    recode

> ?vif() # Variance Inflation Factors
> vif(lm_2)
  area_m2  floor_no built_age    urban2
 1.114160  1.199283  1.393454  1.109796
> round(vif(lm_2), 3)
  area_m2  floor_no built_age    urban2
    1.114     1.199     1.393     1.110
```

다중 선형회귀 모형

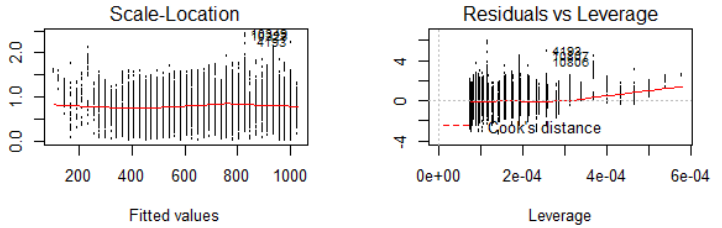
- 아파트의 전용면적(area_m2), 층수(floor_no), 건축년령(built_age), 그리고 거래지역 유형(urban2)은 평당 실거래 가격에 얼마만큼의 가격 결정요인이 될까?

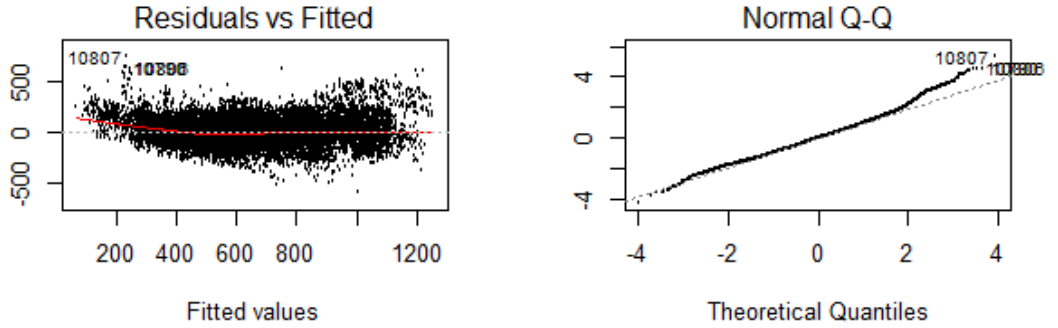



```
> #* 회귀모형 가정 진단
> par(mfrow = c(2,2)) # 한 화면에 2*2 그래프 창 설정
> plot(lm_2, cex=0.3) # 선형회귀 진단 그래프
> #* 6. 회귀모형 가정 진단
> par(mfrow = c(2,2)) # 한 화면에 2*2 그래프 창 설정
> plot(lm_2, cex=0.3) # 선형회귀 진단 그래프
> par(mfrow = c(1,1))
```

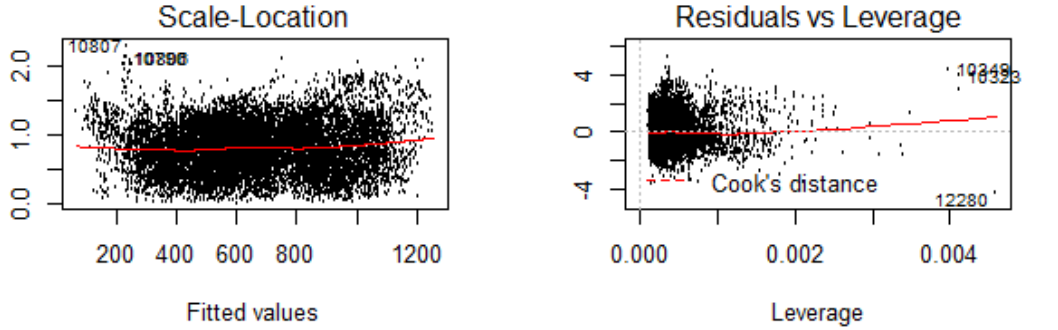


단순선형회귀모형 진단결과와 비교





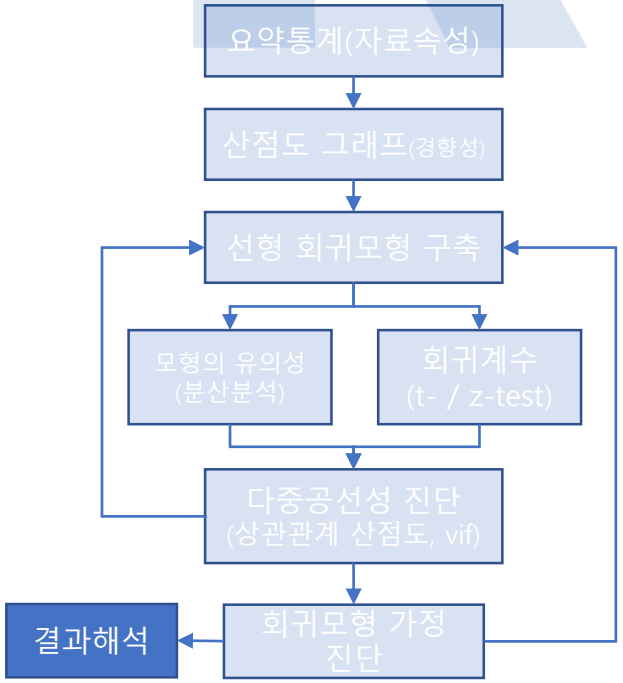
다중선형회귀모형 진단결과



다중 선형회귀 모형

- 아파트의 전용면적(area_m2), 층수(floor_no), 건축년령(built_age), 그리고 거래지역 유형(urban2)은 평당 실거래 가격에 얼마만큼의 가격 결정요인이 될까?
  - 표준화 회귀계수(standardized coefficient, beta)

$$\beta_i^* = \beta_i \cdot \sigma_{x_i} / \sigma_y$$




```
> #* 7. 표준화 회귀계수: 변수의 상대적 중요성(영향력)의 크기 비교
> install.packages("lm.beta")
Error in install.packages : Updating loaded packages
> library(lm.beta)
> ?lm.beta()
> # lm.beta(회귀분석결과)
> lm_2 <- lm(price_pyung ~
+              area_m2 + floor_no + built_age + urban2, data=apt3)
> summary(lm_2) # 다중선형회귀모형 결과 반환

Call:
lm(formula = price_pyung ~ area_m2 + floor_no + built_age + urban2
    data = apt3)

Residuals:
    Min      1Q  Median      3Q     Max
-592.05  -95.65   -3.12   85.25  736.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 803.1352     5.4373  147.71   <2e-16 ***
area_m2       0.7448     0.0543   13.72   <2e-16 ***
floor_no      7.0423     0.2013   34.98   <2e-16 ***
built_age   -21.8320     0.1434 -152.25   <2e-16 ***
urban2도시  161.5030     2.6773   60.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.4 on 13304 degrees of freedom
Multiple R-squared:  0.7425,    Adjusted R-squared:  0.7424
F-statistic:  9590 on 4 and 13304 DF,  p-value: < 2.2e-16
```

```
> beta <- lm.beta(lm_2) # 표준화 회귀계수 할당
> beta # 반환

Call:
lm(formula = price_pyung ~ area_m2 + floor_no + built_age + ur
    data = apt3)

Standardized Coefficients::
(Intercept)     area_m2    floor_no   built_age  urban2도시
  0.0000000   0.0636947   0.1685427  -0.7906887   0.2795772

> coef(lm_2) # 비표준화 회귀계수와 비교
(Intercept)     area_m2    floor_no   built_age  urban2도시
803.1352143   0.7447782   7.0423374 -21.8320260 161.5029842
```

연습문제 06

- 최근 거래되어진 아파트의 건축연령(built_age)과 거래지역(urban2)가 평당 거래가격(price_pyung)에 미치는 영향에 대한 회귀분석을 실시하여 그 결과를 summary()함수로 확인하였다.
  - 이 모형의 수정(조정) 결정계수는 얼마인가요?
  - 건축연령이 1년 더 오래되어지게 되면 아파트 평당 거래가격은 얼마나 떨어지게 되나요?
  - 농촌지역에 비하여 도시지역에서 거래되어지는 아파트는 평균적으로 얼마나 더 비싸게 거래되어지나요?

연습문제 07

- 아파트 거래가격(price_pyung)에 영향을 미치는 독립변수로 floor_no, area_m2, built_age, urban, season을 투입하여 선형회귀모형을 구축하여 회귀분석을 실시하였다.

  - 아파트 거래가격은 어느 계절에 거래되었을 때 가장 높은 가격대를 형성하는 지, 계절명을 적으시오.

  - 표준화 회귀계수를 활용하였을 때 어떠한 독립변수가 아파트 가격에 상대적으로 가장 큰 영향력(중요도)을 가지고 있는 지 그 변수명과 표준화 회귀계수의 값을 적으시오.

상호작용과 주효과

- 상호작용(interaction) 효과
  - 어떤 한 독립변수의 종속변수에 대한 영향력이 어떠한 다른 독립변수에 의하여 더 강해지거나 약해지는 효과
    - 보다 더 강해지는 것을 강화효과(amplification effect),
    - 보다 더 약해지는 것을 조절효과(moderation effect)라고도 함
- 주효과(main effect)
  - 독립변수들의 종속변수에 대한 독립적인 효과

(a) $\hat{Y} = \hat{a} + b_1 X_1 + b_2 X_2$

X2의 주효과

(b) $\hat{Y} = \hat{a} + \hat{b_1} X_1 + \hat{b_2} X_2 + \hat{b_3} X_1 X_2$

X1의 주효과

X1과 X2의 상호작용 효과

상호작용과 주효과

- 상호작용(interaction)과 다중공선성
  - 서로 독립성이 있는, 즉 상관성이 낮은 두 독립변수들의 곱(예: X1, X2)을 회귀모형에 추가하는 경우 새로 만들어진 변수 (예: X1*X2) 는 정의상 기존의 두 변수들과 공선성이 높을 수 밖에 없음
  - 즉, 상호작용항은 이의 구성이 되는 기존 변수들과 높은 다중공선성이 발생함
    - 등간 또는 비율척도의 수준으로 측정된 연속 독립변수들의 경우 발생하는 경향이 높음
- 다중공선성 발생 시 대처방안
  - 편차변환(centering)
    - 모형에서 기존 변수들의 각각의 평균을 차감한 편차값을 적용

$$(= X_1 - \overline{X_1}, X_2 - \overline{X_2}\ )$$

상호작용과 주효과

- 상호작용(interaction)과 다중공선성 진단
  - 실습1: 아파트 평당 실거래 가격에 대한 층수(floor_no)와 건축연령(built_age)의 상호작용 효과는? 이 때의 다중공선성은 발생할까?

```
> #* 실습1: 아파트 평당 실거래 가격에 대한 층수(floor_no)와 건축연령(b
> lm_i2 <- lm(price_pyung ~
+               area_m2 + (floor_no + built_age)^2 # 층수(floor
+             + urban2)
> summary(lm_i2) # 회귀분석 결과 반환

Call:
lm(formula = price_pyung ~ area_m2 + (floor_no + built_age)^2 +
    urban2)

Residuals:
    Min      1Q  Median      3Q     Max
-587.81  -96.19   -3.85   85.82  696.13

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       762.96020    5.72304  133.31   <2e-16 ***
area_m2             0.80762    0.05360   15.07   <2e-16 ***
floor_no           10.66541    0.26882   39.67   <2e-16 ***
built_age         -18.97544    0.20106  -94.38   <2e-16 ***
urban2도시        166.39067    2.64948   62.80   <2e-16 ***
floor_no:built_age  -0.38790    0.01942  -19.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136.4 on 13303 degrees of freedom
Multiple R-squared:   0.75,     Adjusted R-squared:  0.7499
F-statistic:  7982 on 5 and 13303 DF,  p-value: < 2.2e-16
```

```
> round(vif(lm_i2), 3) # VIF 값 반환: 소숫점 셋째자리에서 반올림
     area_m2     floor_no    built_age       urban2 floor_no:built_age
       1.118        2.202        2.822        1.119            2.530
> lm <- lm(price_pyung ~ # 상호작용항이 없는 회귀모형
+            area_m2 + floor_no + built_age  + urban2)  # 4개의 독립변수
> round(vif(lm), 3) # 다중공선성 진단하는 vif 반환
  area_m2  floor_no built_age    urban2
    1.114     1.199     1.393     1.110
```

상호작용항이 없는 회귀분석 결과와 비교

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 803.1352     5.4373  147.71   <2e-16 ***
area_m2       0.7448     0.0543   13.72   <2e-16 ***
floor_no      7.0423     0.2013   34.98   <2e-16 ***
built_age   -21.8320     0.1434 -152.25   <2e-16 ***
urban2도시  161.5030     2.6773   60.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 138.4 on 13304 degrees of freedom
Multiple R-squared:  0.7425,    Adjusted R-squared:  0.7424
F-statistic:  9590 on 4 and 13304 DF,  p-value: < 2.2e-16
```

상호작용과 주효과

- 상호작용(interaction)과 다중공선성 진단
  - 실습2: 아파트 평당 실거래 가격에 대한 전용면적(area_m2)과 층수(floor_no)와의 상호작용 효과는? 이 때의 다중공선성은 발생할까?

```
> #* 실습2: 아파트 평당 실거래 가격에 대한 전용면적(area_m2)과 층수(floor_no
> lm_i3 <- lm(price_pyung ~
+               (area_m2 + floor_no)^2  # 전용면적과 층수의 상호작용 효과
+             + built_age + urban2)
> summary(lm_i3) # 회귀분석 결과 반환

Call:
lm(formula = price_pyung ~ (area_m2 + floor_no)^2 + built_age +
    urban2)

Residuals:
    Min      1Q  Median      3Q     Max
-684.52  -95.69   -2.76   84.87  731.81

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)     847.601501   7.595547  111.592  < 2e-16 ***
area_m2           0.170345   0.087480    1.947 0.051527 .
floor_no          2.165892   0.616794    3.512 0.000447 ***
built_age       -21.989095   0.144253 -152.434  < 2e-16 ***
urban2도시      162.405344   2.672579   60.767  < 2e-16 ***
area_m2:floor_no  0.063041   0.007539    8.362  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138 on 13303 degrees of freedom
Multiple R-squared:  0.7438,    Adjusted R-squared:  0.7437
F-statistic:  7726 on 5 and 13303 DF,  p-value: < 2.2e-16
```

```
> round(vif(lm_i3), 3) # 다중공선성 진단하는 vif 반환
       area_m2       floor_no      built_age        urban2 area_m2:floor_no
         2.907         11.317          1.417          1.112         14.026
```

상호작용과 주효과

- 상호작용(interaction)과 다중공선성 진단
  - 실습2: 아파트 평당 실거래 가격에 대한 전용면적(area_m2)과 층수(floor_no)와의 상호작용 효과는? 이때의 다중공선성은 발생할까?
  - 편차 변환(centering)

```
> detach()
> #* 편차변환(centering)
> apt3 <- apt3 %>%
+          mutate(area_m22 = area_m2 - mean(area_m2), # 전용면적 편차변환: area_m22
+                 floor_no2 = floor_no - mean(floor_no)) # 층수 편차변환: floor_no2
> attach(apt3)
```

```
> lm_i4 <- lm(price_pyung ~
+               (area_m22 + floor_no2)^2  # 전용면적과 층수의 상호작용
+             + built_age + urban2)
> summary(lm_i4) # 회귀분석 결과 반환

Call:
lm(formula = price_pyung ~ (area_m22 + floor_no2)^2 + built_age +
    urban2)

Residuals:
    Min      1Q  Median      3Q     Max
-684.52  -95.69   -2.76   84.87  731.81

Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)        916.863726   2.741616  334.425   <2e-16 ***
area_m22             0.727300   0.054200   13.419   <2e-16 ***
floor_no2            6.510828   0.210612   30.914   <2e-16 ***
built_age          -21.989095   0.144253 -152.434   <2e-16 ***
urban2도시         162.405344   2.672579   60.767   <2e-16 ***
area_m22:floor_no2   0.063041   0.007539    8.362   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138 on 13303 degrees of freedom
Multiple R-squared:  0.7438,    Adjusted R-squared:  0.7437
F-statistic:  7726 on 5 and 13303 DF,  p-value: < 2.2e-16
```

```
> round(vif(lm_i4), 3) # 다중공선성 진단하는 vif 반환
     area_m22     floor_no2     built_age       urban2 area_m22:floor_no2
        1.116         1.319         1.417         1.112             1.104
```

편차변환하지 않은 회귀분석 VIF 결과와 비교

```
> round(vif(lm_i3), 3) # 다중공선성 진단하는 vif 반환
     area_m2     floor_no     built_age       urban2 area_m2:floor_no
       2.907       11.317        1.417        1.112           14.026
```

상호작용과 주효과

- 상호작용(interaction)과 다중공선성 진단
  - 실습2: 도시와 농촌지역에서 각각 거래된 아파트의 건축연령과 상호작용은 평당 실거래 가격에 어떠한 영향을 미칠까?

```
> #* 실습3: 도시와 농촌지역에서 각각 거래된 아파트의 건축연령과 상호작용
> lm_i5 <- lm(price_pyung ~
+               area_m2 + floor_no
+             + (built_age + urban2)^2 )  # 건축연령과 거래지역의
> summary(lm_i5) # 회귀분석 결과 반환

Call:
lm(formula = price_pyung ~ area_m2 + floor_no + (built_age +
    urban2)^2)

Residuals:
    Min      1Q  Median      3Q     Max
-595.18  -95.71   -2.86   84.35  740.55

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         796.58618    5.96001  133.66  < 2e-16 ***
area_m2               0.75254    0.05436   13.84  < 2e-16 ***
floor_no              7.06472    0.20144   35.07  < 2e-16 ***
built_age           -21.36225    0.22644  -94.34  < 2e-16 ***
urban2도시          171.77725    4.67554   36.74  < 2e-16 ***
built_age:urban2도시  -0.71157    0.26550   -2.68  0.00737 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.4 on 13303 degrees of freedom
Multiple R-squared:  0.7426,    Adjusted R-squared:  0.7425
F-statistic:  7677 on 5 and 13303 DF,  p-value: < 2.2e-16
```

```
> round(vif(lm_i5), 3) # 편차변환 모형 vif 반환
     area_m2     floor_no     built_age        urban2 built_age:urban2
       1.117        1.201        3.477        3.386            6.523
```

연습문제 08

- 아파트 거래가격(price_pyung)에 영향을 미치는 독립변수로 built_age와 season, 그리고 이 두 변수의 상호작용항을 투입하여 모형을 구축하고 회귀분석을 실시하였다.

  - 건축연령이 오래된 아파트일수록 어느 계절에 거래되어지면 아파트 가격이 상대적으로 보다 낮게 거래될까요? 그 계절명은?

  - 그 상호작용 회귀계수값은?

회귀모형 비교 및 최적 모형 선택

- 다중회귀모형의 구축과 모형들의 비교
  - 다중회귀모형은 수 많은 독립변수들로 구성되어 어떤 종속변수의 결과에 영향을 미치는 결정요인들로 구성되어진다.
    - 이때 어떤 독립변수들로 회귀모형을 구축하여야 할까?
  - 회귀모형의 구성 원칙
    - 가장 알고 싶어하는 변수들은 반드시 모형에 포함되어야 함
    - 종속변수에 영향을 미치는 중요한 다른 독립변수들도 포함되어야
      - 이를 통제변수라고 함
    - 종속변수에 영향을 미칠 것이라고 예상하였지만 실제 분석결과 통계적으로 유의하지 않은 변수들은 모형에 포함되어도 되고, 그러하지 않아도 됨
      - 만약 그러한 경우에는 다른 중요한 설명변수들과의 다중공선성이 크지 않아야 됨
- 회귀모형의 구축에서 변수의 추가될 때의 모형들의 비교와 선택 기준
  - 추가된 회귀계수의 통계적 유의성으로 판단
    - 유의하지 않을 경우 이를 포함하지 않은 모형이 더 적합
  - 모형의 비교
    - 수정(조정)결정계수(Adjusted determination coefficient)
      - 값이 클수록 좋음
    - AIC (Akaike Information Criterion)
      - 값이 적을수록 좋음
    - BIC (Bayesian Information Criterion)
      - 값이 적을수록 좋음
    - anova() 활용

회귀모형 비교 및 최적 모형 선택

- 모형의 비교
  - 수정(조정)결정계수(Adjusted determination coefficient, adj. R^2)
    - 독립 변수가 추가되면 결정 계수(R-squared)의 값은 항상 증가
      - 즉, 유의하지 않거나 우연의 일치로 유의한 독립변수의 추가 또는 독립변수들간 공선성으로 인하여 결정계수(R^2)는 커지는 경향
    - 결정계수(R^2)의 조정 또는 수정이 필요
      - 독립변수들의 개수 K에 따라 결정계수의 값을 조정

$$R^2_{adj} = 1 - \frac{n-1}{n-K}(1-R^2) = \frac{(n-1)R^2 + 1 - K}{n-K}$$

  - 정보량 규준(information criterion)
    - 최대 우도에 독립 변수의 갯수에 대한 손실(penalty)분을 반영하는 방법
    - 손실 가중치의 계산 법에 따라 AIC (Akaike Information Criterion)와 BIC (Bayesian Information Criterion) 두 가지를 사용
    - AIC (Akaike Information Criterion)
      - 값이 적을수록 좋음
    - BIC (Bayesian Information Criterion)
      - 값이 적을수록 좋음

$$\mathrm{AIC} = -2\log L + 2K$$

$$\mathrm{BIC} = -2\log L + K\log n$$

  - anova() 함수 활용

회귀모형 비교 및 최적 모형 선택

- 모형의 비교
  - 실습1: 아파트 평당 실거래가격(price_pyung)에 영향을 미치는 독립변수들을 1개, 2개, 3개로 추가하였을 대, 어떤 회귀모형이 아파트 가격의 결정모형으로 보다 더 적합한 지를 진단하시오.
  - 조정결정계수와 AIC, BIC 비교

lm_1 결과

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 398.50986    6.95674   57.28   <2e-16 ***
area_m2       3.88316    0.09561   40.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 257.2 on 13307 degrees of freedom
Multiple R-squared:  0.1103,    Adjusted R-squared:  0.1102
F-statistic:  1650 on 1 and 13307 DF,  p-value: < 2.2e-16
```

lm_2 결과

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 304.29382    6.42337   47.37   <2e-16 ***
area_m2       2.93879    0.08695   33.80   <2e-16 ***
floor_no     18.03145    0.31071   58.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 229.8 on 13306 degrees of freedom
Multiple R-squared:   0.29,     Adjusted R-squared:  0.2899
F-statistic:  2717 on 2 and 13306 DF,  p-value: < 2.2e-16
```

```
> #### 모형 비교
> #* 실습1: 아파트 평당 실거래가격(price_pyung)에 영향을 미치는 독립변수들을 0개, 1
시오.
> lm_1 <- lm(price_pyung ~ area_m2) # 독립변수가 1개인 모형
> lm_2 <- lm(price_pyung ~ area_m2
+            + floor_no) # 독립변수가 1개 더 추가된 모형
> lm_3 <- lm(price_pyung ~ area_m2
+            + floor_no + built_age) # 독립변수가 2개 더 추가된 모형
> summary(lm_1) ; summary(lm_2); summary(lm_3) # 3개의 회귀모형 분석결과 반환
```

lm_3 결과

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 835.7210     6.1054  136.88   <2e-16 ***
area_m2       1.1080     0.0609   18.20   <2e-16 ***
floor_no      7.7273     0.2268   34.07   <2e-16 ***
built_age   -19.1667     0.1539 -124.51   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 156.2 on 13305 degrees of freedom
Multiple R-squared:  0.6721,    Adjusted R-squared:  0.672
F-statistic:  9089 on 3 and 13305 DF,  p-value: < 2.2e-16
```

```
> AIC(lm_1, lm_2,lm_3)
      df      AIC
lm_1   3 185502.9
lm_2   4 182502.1
lm_3   5 172223.2
```

```
> BIC(lm_1, lm_2,lm_3)
      df      BIC
lm_1   3 185525.4
lm_2   4 182532.1
lm_3   5 172260.7
```

회귀모형 비교 및 최적 모형 선택

- 모형의 비교
  - 실습1: 아파트 평당 실거래가격(price_pyung)에 영향을 미치는 독립변수들을 1개, 2개, 3개로 추가하였을 대, 어떤 회귀모형이 아파트 가격의 결정모형으로 보다 더 적합한 지를 진단하시오.
  - **ANOVA로 모델간의 비교 및 평가**
    - F통계량은 두 모델의 SSR과 자유도를 통해 구할 수 있으며, 두 모델의 차이가 있는지를 검정하기 위한 통계량

```
> anova(lm_1,lm_2) # ANOVA로 모델간의 비교 및 평가
Analysis of Variance Table

Model 1: price_pyung ~ area_m2
Model 2: price_pyung ~ area_m2 + floor_no
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1  13307 880520527
2  13306 702670990  1 177849537 3367.8 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(lm_2,lm_3) # ANOVA로 모델간의 비교 및 평가
Analysis of Variance Table

Model 1: price_pyung ~ area_m2 + floor_no
Model 2: price_pyung ~ area_m2 + floor_no + built_age
  Res.Df       RSS Df Sum of Sq     F    Pr(>F)
1  13306 702670990
2  13305 324542494  1 378128497 15502 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

회귀모형 비교 및 최적 모형 선택

- 모형의 비교와 최적 모형 선택
  - Adj. R^2, AIC, BIC 비교로 최적 모형 선택
- 단계별 회귀(Stepwise Regression) 선택법
  - 전진선택법(Forward selection)
    - starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
  - 후진선택법(Backward selection or backward elimination)
    - starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
  - 단계별 선택법: 위의 2가지 방식 조합
    - start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection)
- 다수의 모형 비교와 선택 함수 제공 패키지: "MASS", "leaps"

회귀모형 비교 및 최적 모형 선택

- 모형의 비교와 최적 모형 선택
  - Adj. R^2, AIC, BIC 비교로 최적 모형 선택
- 단계별 회귀(Stepwise Regression) 선택법
  - 전진선택법(Forward selection)
    - starts with no predictors in the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.
  - 후진선택법(Backward selection or backward elimination)
    - starts with all predictors in the model (full model), iteratively removes the least contributive predictors, and stops when you have a model where all predictors are statistically significant.
  - 단계별 선택법: 위의 2가지 방식 조합
    - start with no predictors, then sequentially add the most contributive predictors (like forward selection). After adding each new variable, remove any variables that no longer provide an improvement in the model fit (like backward selection)
- 다수의 모형 비교와 선택 함수 제공 패키지: "MASS", "leaps"

회귀와 예측

회귀모형 비교 및 최적 모형 선택

- 패키지: "MASS"

```
> library("MASS")
> library(leaps)
> full.model <- lm(price_pyung ~ # 종속변수
+                  area_m2 + floor_no +  built_age + urban + season, # 설명변수
+                  data=apt3) # 데이터 객체
> summary(full.model) # 모형 결과 반환

Call:
lm(formula = price_pyung ~ area_m2 + floor_no + built_age + urban +
    season, data = apt3)

Residuals:
    Min      1Q  Median      3Q     Max
-562.06  -94.53   -2.70   82.99  754.30

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 1001.87043    6.28576  159.387  < 2e-16 ***
area_m2        0.67770    0.05378   12.602  < 2e-16 ***
floor_no       6.65249    0.19998   33.265  < 2e-16 ***
built_age    -22.11746    0.14271 -154.986  < 2e-16 ***
urban면      -206.40671    4.15943  -49.624  < 2e-16 ***
urban읍      -148.64461    2.96750  -50.091  < 2e-16 ***
season겨울    -20.36472    3.42494   -5.946 2.82e-09 ***
season봄      -30.96942    3.29747   -9.392  < 2e-16 ***
season여름    -40.40854    3.35506  -12.044  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136.7 on 13300 degrees of freedom
Multiple R-squared:  0.749,     Adjusted R-squared:  0.7488
F-statistic:  4960 on 8 and 13300 DF,  p-value: < 2.2e-16
```

결과 동일
∴ 모든 변수 투입 모형이 최적

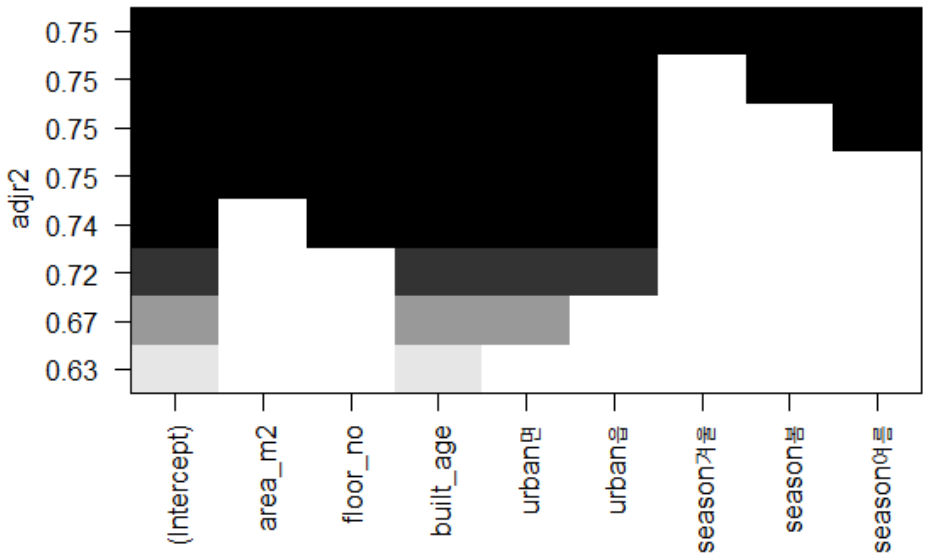
```
> ?stepAIC # Choose a model by AIC in a Stepwise Algorithm
> step.model <- stepAIC(full.model, # Stepwise regression model
+                       direction = "backward", # "both", "backward", "forward" 중 선택
+                       trace = FALSE) # 단계별 선택별 과정 추적 내용 제시(여기서는 F이므로 미제시)
> summary(step.model) # stepAIC 결과 반환

Call:
lm(formula = price_pyung ~ area_m2 + floor_no + built_age + urban +
    season, data = apt3)

Residuals:
    Min      1Q  Median      3Q     Max
-562.06  -94.53   -2.70   82.99  754.30

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 1001.87043    6.28576  159.387  < 2e-16 ***
area_m2        0.67770    0.05378   12.602  < 2e-16 ***
floor_no       6.65249    0.19998   33.265  < 2e-16 ***
built_age    -22.11746    0.14271 -154.986  < 2e-16 ***
urban면      -206.40671    4.15943  -49.624  < 2e-16 ***
urban읍      -148.64461    2.96750  -50.091  < 2e-16 ***
season겨울    -20.36472    3.42494   -5.946 2.82e-09 ***
season봄      -30.96942    3.29747   -9.392  < 2e-16 ***
season여름    -40.40854    3.35506  -12.044  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136.7 on 13300 degrees of freedom
Multiple R-squared:  0.749,     Adjusted R-squared:  0.7488
F-statistic:  4960 on 8 and 13300 DF,  p-value: < 2.2e-16
```

회귀모형 비교 및 최적 모형 선택

- 패키지: “leaps”

```
> ?regsubsets() # "leaps" 패키지: functions for model selection
> lm_fw <-regsubsets(price_pyung ~ # 종속변수
+                      area_m2 + floor_no +  built_age + urban +  season , # 설명변
+                    method = "forward", # forward selection, backward selection
+                    data=apt3)
> summary(lm_fw) # 결과 반환
Subset selection object
Call: regsubsets.formula(price_pyung ~ area_m2 + floor_no + built_age +
    urban + season, method = "forward", data = apt3)
8 variables  (and intercept)
           Forced in Forced out
area_m2        FALSE      FALSE
floor_no       FALSE      FALSE
built_age      FALSE      FALSE
urban면        FALSE      FALSE
urban읍        FALSE      FALSE
season겨울     FALSE      FALSE
season봄       FALSE      FALSE
season여름     FALSE      FALSE
1 subsets of each size up to 8
Selection Algorithm: forward
         area_m2 floor_no built_age urban면 urban읍 season겨울 season봄 season여름
1  ( 1 ) " "     " "      "*"       " "     " "     " "       " "      " "
2  ( 1 ) " "     " "      "*"       "*"     " "     " "       " "      " "
3  ( 1 ) " "     " "      "*"       "*"     "*"     " "       " "      " "
4  ( 1 ) " "     "*"      "*"       "*"     "*"     " "       " "      " "
5  ( 1 ) "*"     "*"      "*"       "*"     "*"     " "       " "      " "
6  ( 1 ) "*"     "*"      "*"       "*"     "*"     " "       " "      "*"
7  ( 1 ) "*"     "*"      "*"       "*"     "*"     " "       "*"      "*"
8  ( 1 ) "*"     "*"      "*"       "*"     "*"     "*"       "*"      "*"
```

```
> plot(lm_fw, scale="adjr2") # 투입변수별 조정계수 그래프 생성
```



연습문제 09

- 아래와 같이 아파트 평당 거래가격에 영향을 미치는 설명변수들을 세개의 모형에 각각 달리 투입하여 회귀분석을 실행하고자 한다.
  - lm_e09_1 <- lm(price_pyung ~ floor_no) # 1번
  - lm_e09_2 <- lm(price_pyung ~ floor_no + built_age) # 2번
  - lm_e09_3 <- lm(price_pyung ~ floor_no + season) # 3번
- 이들 중 최적의 회귀모형을 분산분석(anova())으로 채택하고자 할 때 어떤 모형이 최적인 지 그 결과를 실행하고, 최적의 모형은 어떤 것인지 적으시오.
- 그리고 각각의 조정결정계수 값을 적으시오.

연습문제 10

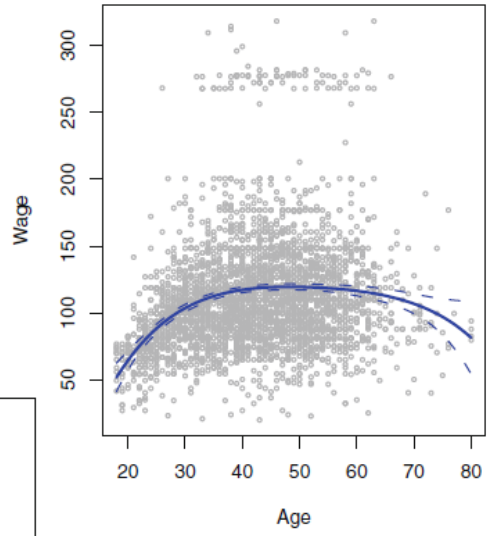
- 아래와 같이 아파트 평당 거래가격에 영향을 미치는 설명변수들을 세개의 모형에 각각 달리 투입하여 회귀분석을 실행하고자 한다.
  - lm_e09_1 <- lm(price_pyung ~ floor_no) # 1번
  - lm_e09_2 <- lm(price_pyung ~ floor_no + built_age) # 2번
  - lm_e09_3 <- lm(price_pyung ~ floor_no + season) # 3번
- 이들 중 최적의 회귀모형을 AIC 통계량으로 선택하고자 한다. 이 때 각각의 AIC 값을 적으시오.

비선형 회귀모형의 개요

- 선형 회귀모형의 장점과 단점
  - 단순하여 해석과 추론, 예측이 쉬움
  - 현실 문제에서 무리한 선형성 가정에 따른 정보의 손실
- 비선형 회귀모형의 정의
  - 설명변수와 종속변수의 비선형 관계를 추정하는 모형
  - 선형회귀모형의 해석력은 가능한 잃지 않으면서 선형의 가정을 완화시켜 추정하는 모형
- 선형회귀와의 유사성과 차이점
  - 유사성
    - 하나의 종속변수와 하나 이상의 설명변수 사이의 관계를 수학적으로 설명
    - 곡선형태의 관계를 모형화
      - 선형에서는 종속변수에 log, exp 등의 함수를 적용하여 선형으로 모형화 가능
    - 잔차(오차)의 제곱합(SSE)을 최소화하는 추정선을 도출
  - 차이점
    - 비선형 회귀모형에서는 모수가 필요하지 않음
    - 선형회귀모형은 하나의 방정식이 기본 형태이지만 비선형 회귀모형은 여러가지 방정식이 사용될 수 있음
    - 비선형 회귀분석에서는 선형회귀분석과 다른 절차를 사용하여 잔차의 제곱합을 최소화 함
      - 연속적 근사값 추정으로 모형 적합도 도달

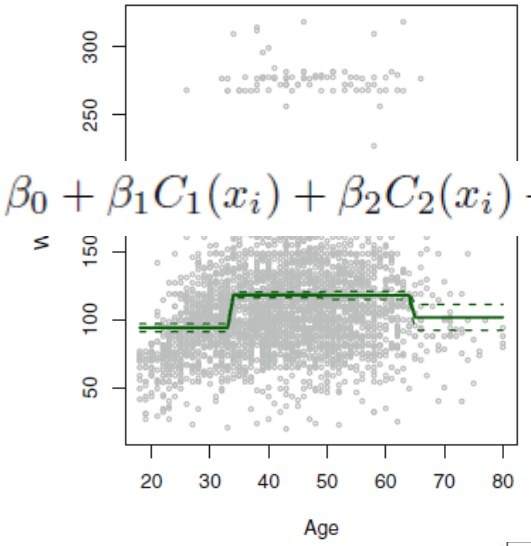
비선형 회귀모형의 개요

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + .. \beta_d x_i^d + \epsilon_i$$

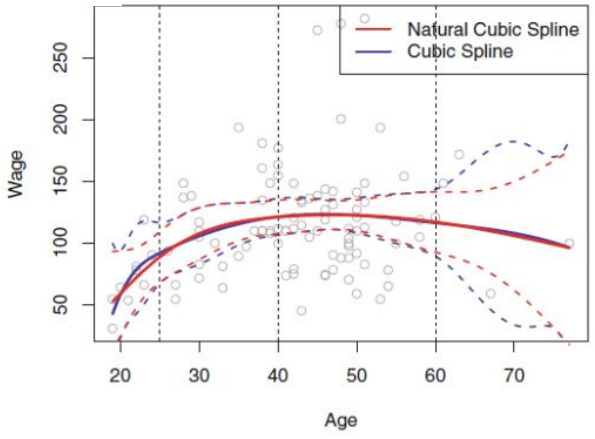


- 비선형 회귀모형의 종류
  - **다항회귀모형(Polynomial regression model)**
    - 기존의 변수의 다차항(x^2, x^3등)을 추가하여 non-linear data에 적합을 할 수 있도록 선형모델을 확장
  - 단계함수 모형(Step functional model, piecewise constant regression)
    - 변수를 K개의 부분으로 나누어, 질적변수(즉, constant)로 하여 추정
  - **스플라인 모형(spline model)**
    - 높은 수준의 고차 항을 추가하는 것은 회귀모형에 바람직하지 않음
    - 일련의 조각별 다항함수식이라 할 수 있음
    - 위의 두 방식의 확장으로, 전체 X를 똑같이 K개의 범주로 나누되 각 범주내에서 다항적합을 추정
      - K개의 knots가 있을때, 이에 대하여 각각 매끄럽게 선형적합(natural cubic spline)
    - 다항적합은 양 옆의 범주의 다항함수와 매끄럽게(smoothly) 연결되도록 한다는 제약을 조건으로 함
    - 통계적으로 단순하고 표준적인 모수 추론이 가능
      - 이 방법은 부드러운 함수의 LM 표현 방법

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i$$



$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

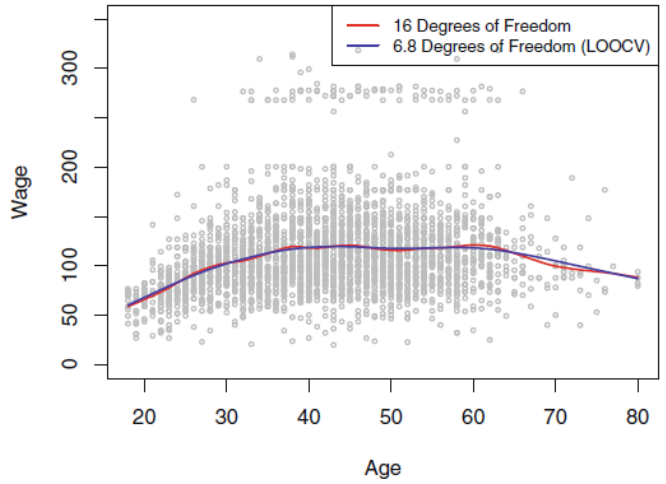


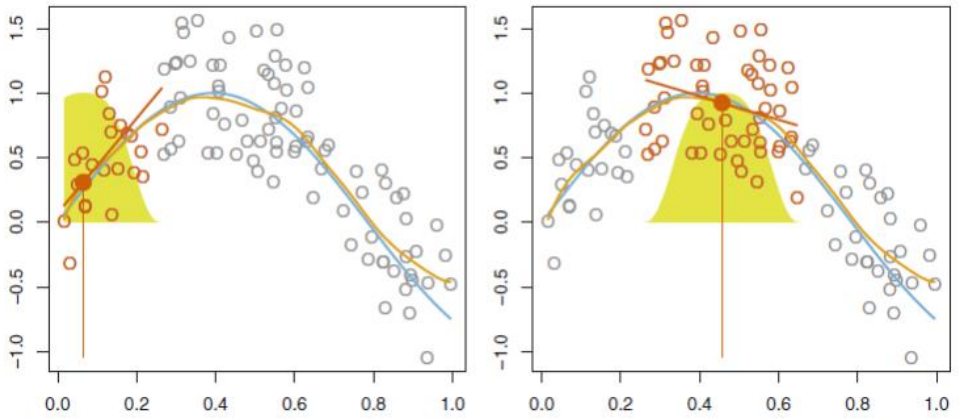
출처: 비선형모델(Local regression, Smoothing splines, GAM) 이해하기
https://godongyoung.github.io/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D/2018/02/14/ISL-Moving-Beyond-Linearity_ch7.html

비선형 회귀모형의 개요

- 비선형 회귀모형의 종류
  - 스무딩 스플라인 모형(smoothing spline model)
    - regression splines과 비슷하지만, smooth penalty를 포함한 SSE식을 최소화하는 방식으로 적합
    - 모든 관측치 xi에 knot를 두기에 knot의 갯수나 위치를 지정해야 하는 문제 해결
  - 국지적 회귀모형(local regression model)
    - spline방식과 유사하지만 각 범주가 겹칠수 있는 방식으로 더욱 유연한 적합을 가능하게 함
    - 각 특정 target point x0에서 그 근처의 관측자료들만을 토대로 적합을 시켜 flexible한 적합을 하고자 하는 접근방식
    - 가까운 k개의 자료(nearest neighbor)들만이 가중치를 갖고, 나머지는 가중치가 0이 되도록 가중치 Ki0=K(xi,x0)을 설정한다.

$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{n}(y_i - \hat{g}_{\lambda}^{(-i)}(x_i))^2 = \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_{\lambda}(x_i)}{1 - \{\mathrm{S}_{\lambda}\}_{ii}}\right]^2$$

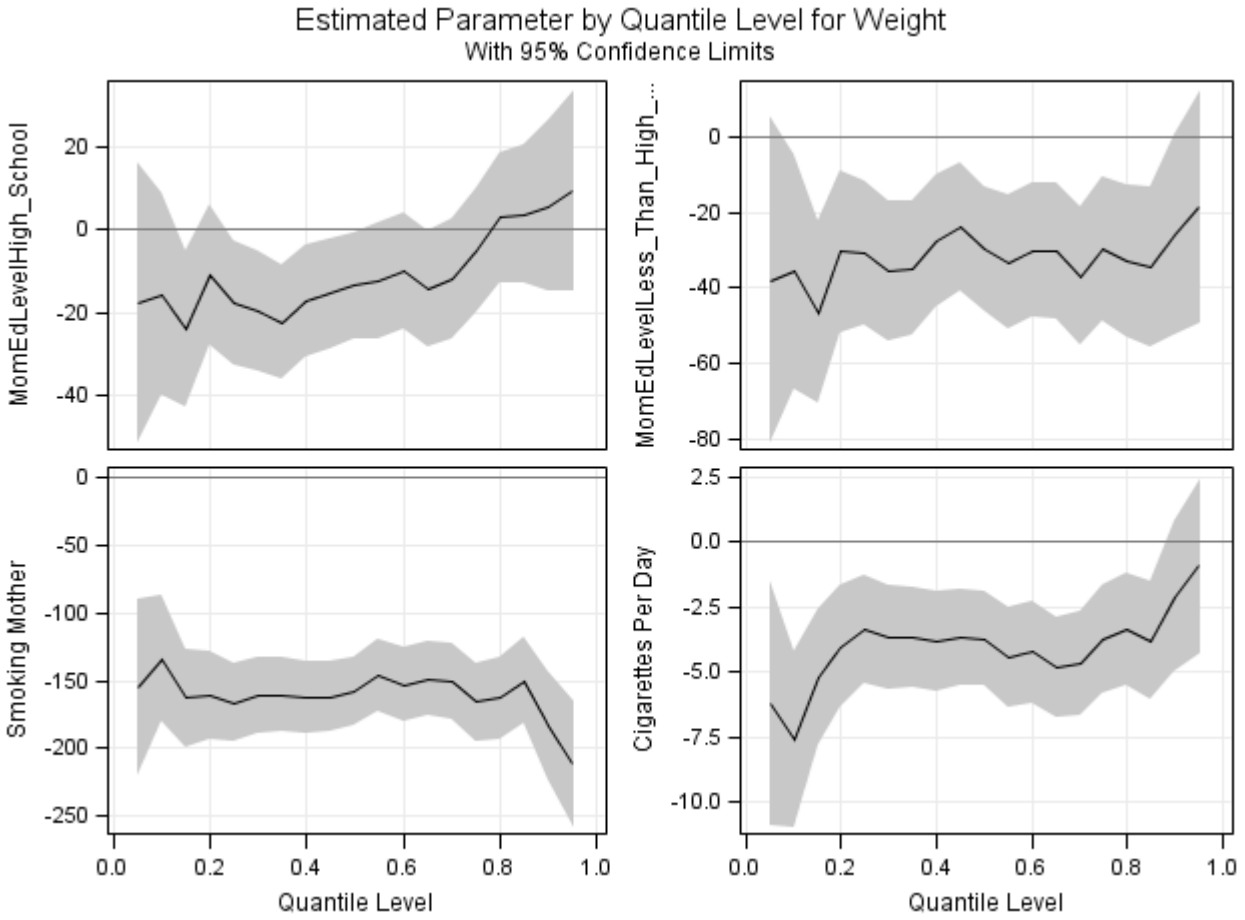




출처: 비선형모델(Local regression, Smoothing splines, GAM) 이해하기
https://godongyoung.github.io/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D/2018/02/14/ISL-Moving-Beyond-Linearity_ch7.html

비선형 회귀모형의 개요

- 분위회귀모형(quantile regression model)
  - 국지적 회귀모형과 유사
  - 이분산일 경우, 구간(분위)별 분석 수행
    - 소득수준별 소비성향을 분석하거나 혈당치(sugar level)별 치료효과 등을 분석할 때
    - 의료비 지출수준별로 만성질환, 보험가입, 건강식음료 지출 등의 차이가 있는 지를 분석할 때 등등
  - 종속변수의 값을 순위로 대별하여 특정 구간별로 효과의 크기가 다를 수 있음을 가정
    - Y의 분위수 수준의 결과를 알고자 할 때 적합



출처: https://towardsdatascience.com/an-introduction-to-quantile-regression-eca5e3e2036a

비선형 회귀모형의 개요

a Gaussian linear model:

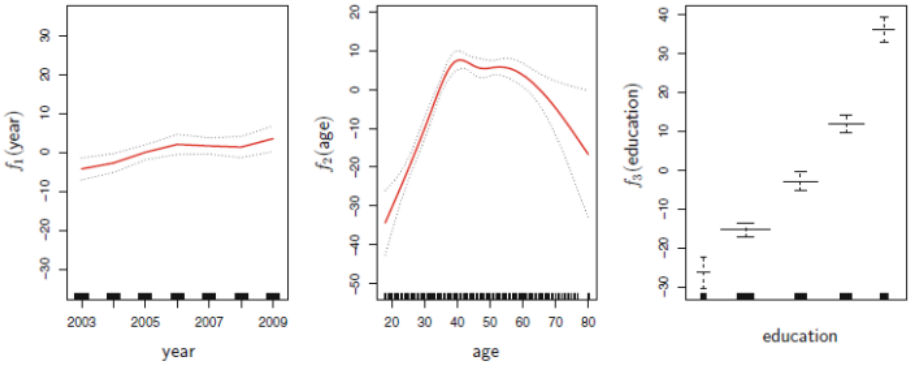
$$y = \beta_0 + x_1\beta_1 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

a GAM is the presence of a smoothing term:

$$y = \beta_0 + f(x_1) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

- 일반화 가법 모형(Generalized additive model, GAM)
  - 위의 비선형 회귀식들을 여러 개의 예측(독립, 설명)변수들에 적용할 수 있게 하는 방식.
  - 기존의 선형모델에서 가법성은 유지하면서도 각 변수에 non-linear한 적합을 가능하게 하는 방법
  - 질적변수와 양적변수에 모두 가능

$$\texttt{wage} = \beta_0 + f_1(\texttt{year}) + f_2(\texttt{age}) + f_3(\texttt{education}) + \epsilon$$



- 일반화 가법모형(GAM)의 장단점
  - 기존의 선형적합에서는 할 수 없던 비선형적합을 자동적으로 수행
    - 각 변수를 연구자가 변형할 필요가 없음
  - 비선형적합이므로 예측력이 높음
  - 가법적인 모델이므로 다른 변수들이 고정되어 있을때의 한 변수의 영향을 알 수있음
    - 그러므로 추론의 측면에서도 강점이 있음
  - Xj에 대한 함수, 즉 fj의 smooth정도가 어느 정도인지를 degrees of freedom으로 간단하게 표현할 수 있음
  - GAMs는 반응변수Y가 질적변수일때도 사용이 가능
  - 그러나 가법적인 모델이라는 점에서 중요한 상호작용을 잡아낼 수 없다는 단점

출처: 비선형모델(Local regression, Smoothing splines, GAM) 이해하기
https://godongyoung.github.io/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D/2018/02/14/ISL-Moving-Beyond-Linearity_ch7.html

다항회귀모형

- 실습1: 아파트 건축연령이 평당 거래가격에 미치는 영향에 대한 비선형 관계를 다항함수식으로 표현하여 그 결과를 확인하여 보자.

```
> lm_p1 <- lm(price_pyung ~
+               built_age) # 건축연령 일차함수형태
> summary(lm_p1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1026.1592     2.7768   369.5   <2e-16 ***
built_age    -21.9605     0.1451  -151.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 165.3 on 13307 degrees of freedom
Multiple R-squared:  0.6326,    Adjusted R-squared:  0.6325
F-statistic: 2.291e+04 on 1 and 13307 DF,  p-value: < 2.2e-16
```

```
> lm_p2 <- lm(price_pyung ~
+               poly(built_age, 2)) # 건축연령 이차함수형태
> summary(lm_p2)

(Intercept)           666.146      1.424  467.67   <2e-16 ***
poly(built_age, 2)1 -25020.599    164.326 -152.26   <2e-16 ***
poly(built_age, 2)2   2082.479    164.326   12.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 164.3 on 13306 degrees of freedom
Multiple R-squared:  0.6369,    Adjusted R-squared:  0.6369
F-statistic: 1.167e+04 on 2 and 13306 DF,  p-value: < 2.2e-16
```

```
> lm_p3 <- lm(price_pyung ~
+               poly(built_age, 3)) # 건축연령 3차함수형태
> summary(lm_p3)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            666.146      1.397  476.97   <2e-16 ***
poly(built_age, 3)1 -25020.599    161.121 -155.29   <2e-16 ***
poly(built_age, 3)2   2082.479    161.121   12.93   <2e-16 ***
poly(built_age, 3)3   3728.910    161.121   23.14   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 161.1 on 13305 degrees of freedom
Multiple R-squared:  0.651,     Adjusted R-squared:  0.6509
F-statistic:  8273 on 3 and 13305 DF,  p-value: < 2.2e-16
```

다항회귀모형

- 실습2: 아파트 건축연령이 평당 거래가격에 미치는 영향에 대한 비선형 관계를 2차항 함수식으로 표현하여 그 결과를 확인하여 보자. 이 때 통제변수로 area_m2 + floor_no + built_age + urban2를 추가하자.

```
> lm_p4 <- lm(price_pyung ~
+               area_m2 + floor_no + built_age + urban2) # 건축연령 1차함수형태
> summary(lm_p4)

Call:
lm(formula = price_pyung ~ area_m2 + floor_no + built_age + urban2)

Residuals:
    Min      1Q  Median      3Q     Max
-592.05  -95.65   -3.12   85.25  736.05

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 803.1352     5.4373  147.71   <2e-16 ***
area_m2       0.7448     0.0543   13.72   <2e-16 ***
floor_no      7.0423     0.2013   34.98   <2e-16 ***
built_age   -21.8320     0.1434 -152.25   <2e-16 ***
urban2도시  161.5030     2.6773   60.32   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.4 on 13304 degrees of freedom
Multiple R-squared:  0.7425,    Adjusted R-squared:  0.7424
F-statistic:  9590 on 4 and 13304 DF,  p-value: < 2.2e-16
```

```
> lm_p5 <- lm(price_pyung ~
+               area_m2 + floor_no + poly(built_age, 2) + urban2) # 건축연령 2차함수형태
> summary(lm_p5)

Call:
lm(formula = price_pyung ~ area_m2 + floor_no + poly(built_age,
    2) + urban2)

Residuals:
    Min      1Q  Median      3Q     Max
-590.45  -93.82   -5.66   84.87  653.35

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.361e+02  4.311e+00  101.16   <2e-16 ***
area_m2              8.607e-01  5.377e-02   16.00   <2e-16 ***
floor_no             7.000e+00  1.983e-01   35.31   <2e-16 ***
poly(built_age, 2)1 -2.483e+04  1.609e+02 -154.33   <2e-16 ***
poly(built_age, 2)2  2.798e+03  1.373e+02   20.38   <2e-16 ***
urban2도시           1.638e+02  2.639e+00   62.07   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 136.3 on 13303 degrees of freedom
Multiple R-squared:  0.7503,    Adjusted R-squared:  0.7502
F-statistic:  7994 on 5 and 13303 DF,  p-value: < 2.2e-16
```

분위회귀모형

- 실습. 건축연령과 아파트 평당 가격에 대한 선형회귀모형과 분위회귀모형을 각각 적용하여 그래프를 그려보자.

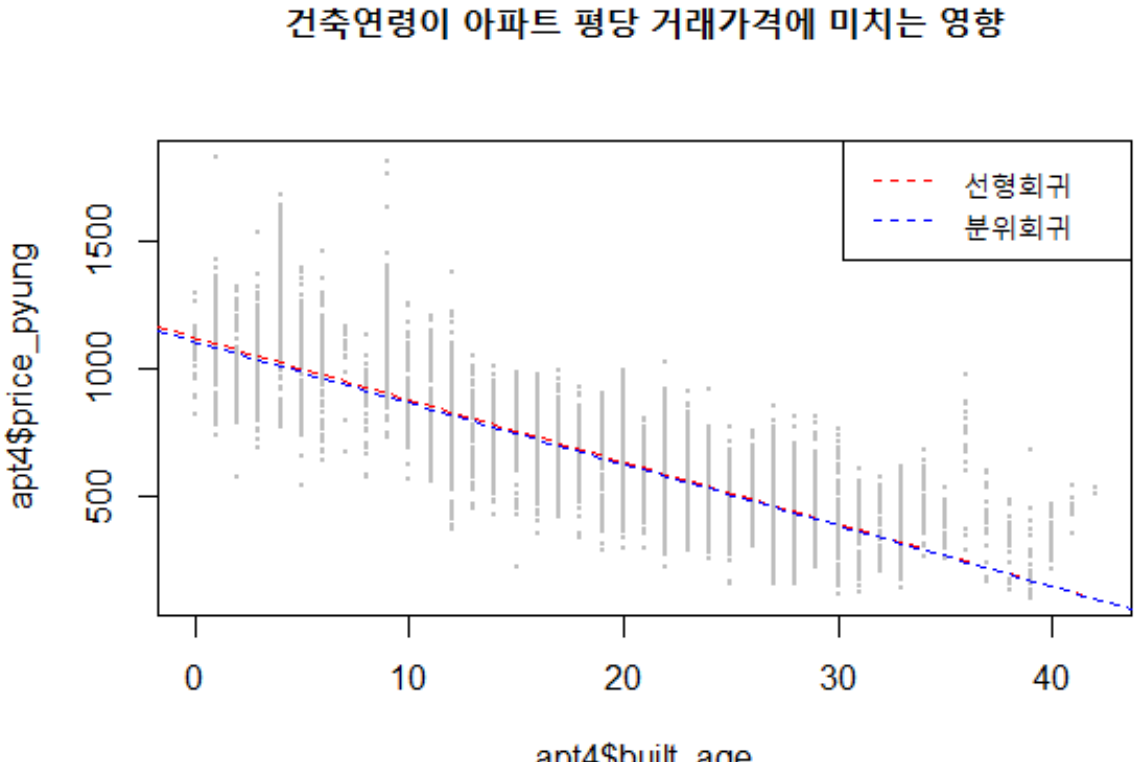
```
> library(quantreg)
> ?rq # Quantile Regression
> lm_0 <- lm(price_pyung ~ # 선형회귀모형
+              built_age, data=apt4) # 건축연령 독립변수
> summary(lm_0) # 선형회귀 결과

Call:
lm(formula = price_pyung ~ built_age, data = apt4)

Residuals:
    Min      1Q  Median      3Q     Max
-536.26 -111.83  -11.99   96.31  905.39

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1123.2654     3.7113   302.7   <2e-16 ***
built_age    -24.3669     0.1798  -135.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 157.3 on 8846 degrees of freedom
Multiple R-squared:  0.6748,    Adjusted R-squared:  0.6748
F-statistic: 1.836e+04 on 1 and 8846 DF,  p-value: < 2.2e-16
```

```
> lm_q <- rq(price_pyung ~ # 분위회귀모형
+              built_age, data=apt4)  # 건축연령 독립변수
> summary(lm_q) # 분위회귀 결과

Call: rq(formula = price_pyung ~ built_age, data = apt4)

tau: [1] 0.5

Coefficients:
            Value      Std. Error t value    Pr(>|t|)
(Intercept) 1107.34081    4.98838  221.98424    0.00000
built_age    -24.10302    0.22350 -107.84281    0.00000
```

분위회귀모형

- 실습. 건축연령과 아파트 평당 가격에 대한 선형회귀모형과 분위회귀모형을 각각 적용하여 그래프를 그려보자.

```
> plot(apt4$built_age, apt4$price_pyung, # 추정 회귀식 작성
+      pch = 16,
+      cex=0.3, col="gray", # 점의 크기 0.3 색상 = gray
+      main = "건축연령이 아파트 평당 거래가격에 미치는 영향")
> abline(lm(price_pyung ~ # 선형회귀 추세선
+             built_age, data = apt4), col = "red", lty = 2)
> abline(rq(price_pyung ~ # 분위회귀 추세선(평균값)
+             built_age, data = apt4), col = "blue", lty = 2)
> legend("topright", legend = c("선형회귀", "분위회귀"), col = c("red", "blue"), lty = 2)
```

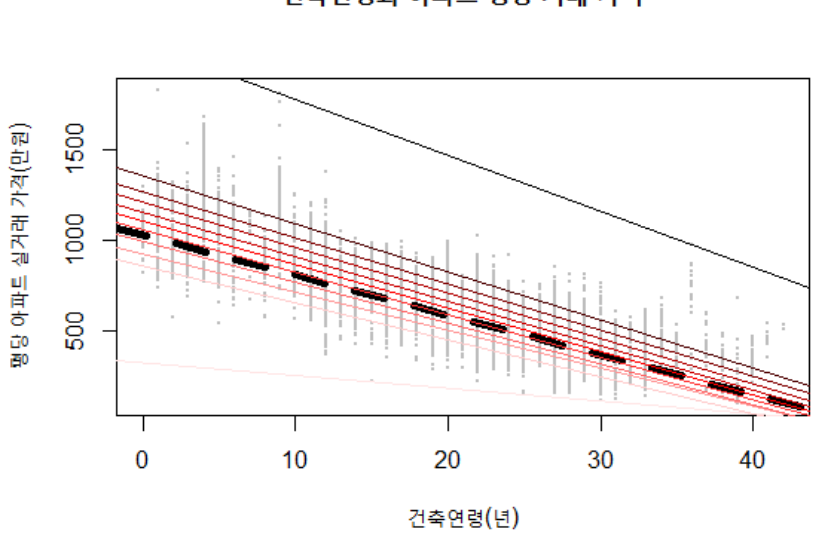


분위회귀모형

- 건축연령과 아파트 평당 가격에 대한 선형회귀모형과 분위회귀모형을 각각 적용하여 그래프를 그려보자.

```
> lm_q2 <- rq(price_pyung ~ # 분위회귀
+              built_age, # 건축연령 독립변수
+            tau = seq(0, 1, by = 0.1), # 분위의 구분을 10%분위로 11개 구간 작성
+            data=apt4)
> lm_q2$coefficients # 분위회귀계수 확인
               tau= 0.0  tau= 0.1  tau= 0.2  tau= 0.3  tau= 0.4   tau= 0.5   tau= 0.6   tau= 0.7   tau= 0.8   tau= 0.9  tau= 1.0
(Intercept) 326.960558 861.51302 919.92874 996.82499 1057.5319 1107.34081 1155.19005 1206.94645 1268.16893 1357.50382 2087.0930
built_age    -7.030893 -20.38594 -20.93423 -22.54644  -23.4882  -24.10302  -24.52349  -24.97986  -25.46765  -26.56878  -30.8595
> coef(lm_q2) # 또 다른 회귀계수 확인 방법
               tau= 0.0  tau= 0.1  tau= 0.2  tau= 0.3  tau= 0.4   tau= 0.5   tau= 0.6   tau= 0.7   tau= 0.8   tau= 0.9  tau= 1.0
(Intercept) 326.960558 861.51302 919.92874 996.82499 1057.5319 1107.34081 1155.19005 1206.94645 1268.16893 1357.50382 2087.0930
built_age    -7.030893 -20.38594 -20.93423 -22.54644  -23.4882  -24.10302  -24.52349  -24.97986  -25.46765  -26.56878  -30.8595
```

```
> # plotting different quantiles
> colors <- c("#ffe6e6", "#ffcccc", "#ff9999", "#ff6666", "#ff3333",
+             "#ff0000", "#cc0000", "#b30000", "#800000", "#4d0000", "#000000")
> plot(price_pyung ~
+        built_age, data = apt4,
+      pch = 16,
+      col = "gray",
+      cex = 0.2,
+      main = "건축연령과 아파트 평당 거래 가격",
+      xlab = "건축연령(년)",
+      ylab = "평당 아파트 실거래 가격(만원)")
> abline(lm(price_pyung ~ built_age ),
+        lwd = 5,
+        col = "black",
+        lty = 2)
> for (j in 1:ncol(lm_q2$coefficients)) {
+   abline(coef(lm_q2)[, j], col = colors[j])
+ }
```



스플라인 회귀모형

- 실습1: 건축연령(built_age)과 아파트 평당 거래가격(price_pyung)의 관계를 비선형으로 가정하고, 이를 스플라인 회귀분석을 시행하고, 이를 선형회귀와 그 결과를 비교하여 보시오.

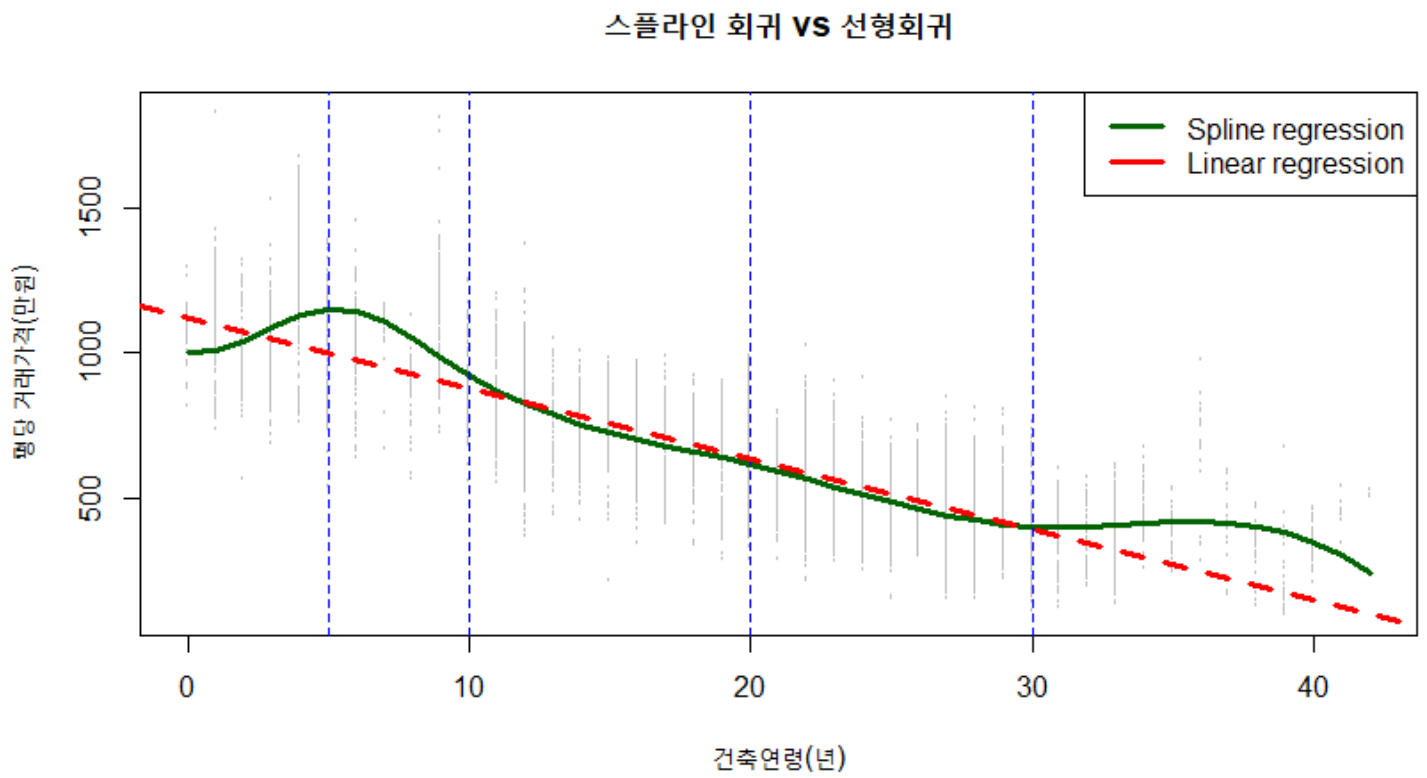
```
> lm <- lm(price_pyung ~ built_age,
+          data = apt4)
> summary(lm)

Call:
lm(formula = price_pyung ~ built_age, data = apt4)

Residuals:
    Min      1Q  Median      3Q     Max
-536.26 -111.83  -11.99   96.31  905.39

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1123.2654     3.7113   302.7   <2e-16 ***
built_age    -24.3669     0.1798  -135.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 157.3 on 8846 degrees of freedom
Multiple R-squared:  0.6748,    Adjusted R-squared:  0.6748
F-statistic: 1.836e+04 on 1 and 8846 DF,  p-value: < 2.2e-16
```

```
> # Generating Test Data
> age.grid<-seq(from=agelims[1], to = agelims[2])
> lm_s <- lm(price_pyung ~ bs(built_age, knots=c(5, 10, 20, 30))) # 4개의 조각으로 
> summary(lm_s)

Call:
lm(formula = price_pyung ~ bs(built_age, knots = c(5, 10, 20,
    30)))

Residuals:
    Min      1Q  Median      3Q     Max
-608.53 -102.93   -8.88   95.96  822.44

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                                1001.39      13.23  75.677   <2e-16 ***
bs(built_age, knots = c(5, 10, 20, 30))1    -23.66      24.18  -0.978    0.328
bs(built_age, knots = c(5, 10, 20, 30))2    300.43      15.22  19.744   <2e-16 ***
bs(built_age, knots = c(5, 10, 20, 30))3   -257.26      17.61 -14.611   <2e-16 ***
bs(built_age, knots = c(5, 10, 20, 30))4   -341.75      14.92 -22.911   <2e-16 ***
bs(built_age, knots = c(5, 10, 20, 30))5   -726.04      18.66 -38.910   <2e-16 ***
bs(built_age, knots = c(5, 10, 20, 30))6   -485.51      26.38 -18.402   <2e-16 ***
bs(built_age, knots = c(5, 10, 20, 30))7   -755.41      35.72 -21.145   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 145.9 on 8840 degrees of freedom
Multiple R-squared:  0.7203,    Adjusted R-squared:  0.7201
F-statistic:  3252 on 7 and 8840 DF,  p-value: < 2.2e-16
```

스플라인 회귀모형

- 실습1: 건축연령(built_age)과 아파트 평당 거래가격(price_pyung)의 관계를 비선형으로 가정하고, 이를 스플라인 회귀분석을 시행하고, 이를 선형회귀와 그 결과를 비교하여 보시오.

```
> #Plotting the Regression Line to the scatterplot
> par(mfrow = c(1,1))
> plot(built_age, price_pyung,
+      col="gray",
+      cex=0.3,
+      xlab="건축연령(년)",
+      ylab="평당 거래가격(만원)",
+      main= "스플라인 회귀 vs 선형회귀")
> points(age.grid,
+        predict(lm_s, newdata = list(built_age=age.grid)),
+        col="darkgreen",
+        lwd=3,type="l")
> #adding cutpoints
> abline(v=c(5, 10,20,30),
+        lty=2,
+        col="blue")
> #adding linear regression modeling line
> abline(lm,
+        lty=2,
+        lwd=3,
+        col="red")
> legend("topright",
+        c("Spline regression","Linear regression"),
+        col=c("darkgreen", "red"),
+        lwd=3)
```

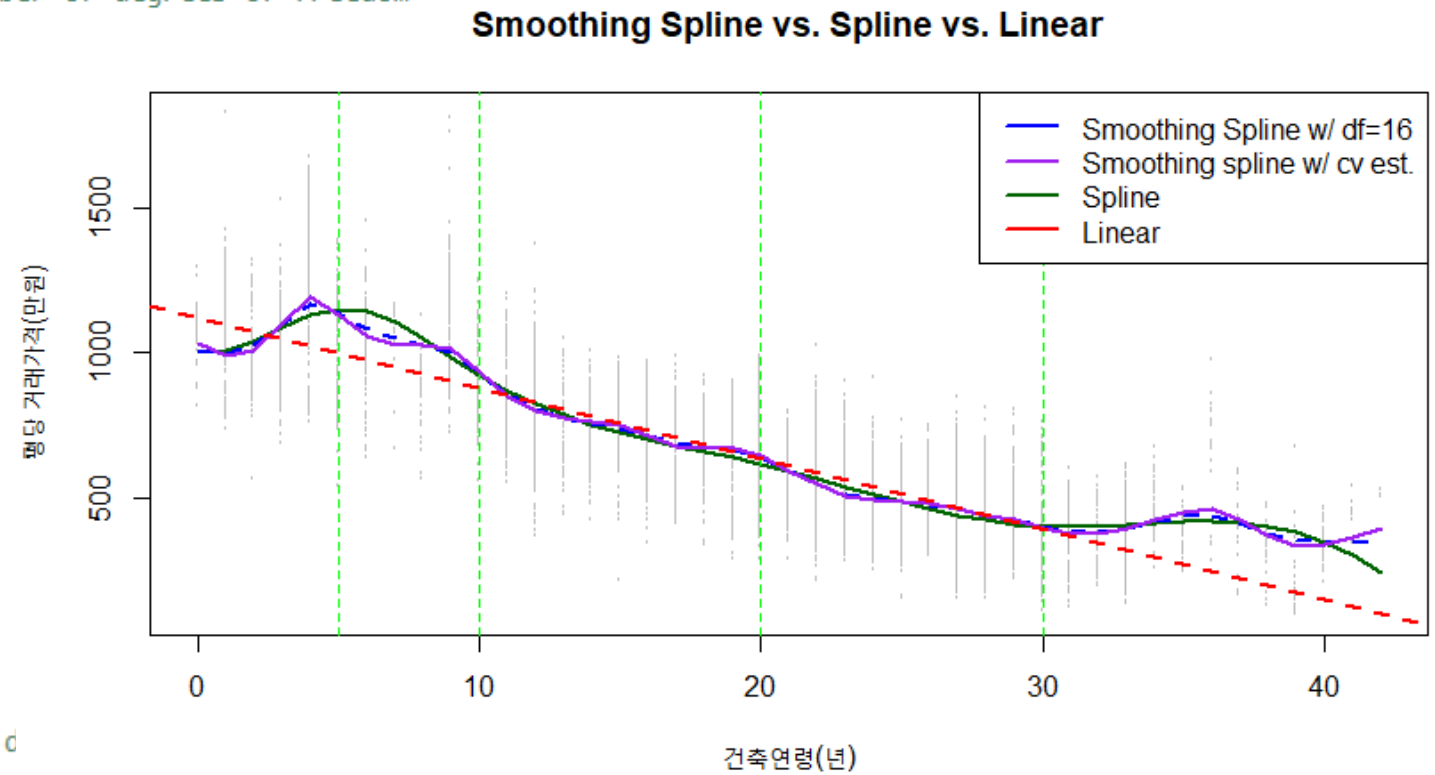


스플라인 회귀모형

- 실습2: 건축연령(built_age)과 아파트 평당 거래가격(price_pyung)의 관계를 비선형으로 가정하고, 이를 스무딩 스플라인 회귀분석을 시행하고, 이전이 결과들과 그래프로 비교하여 보시오.

```
#fitting smoothing splines using smooth.spline(X,Y,df=...)
?smooth.spline() # Fit a Smoothing Spline
lm_ss <-smooth.spline(built_age, price_pyung,
                      df=16) # df = the desired equivalent number of degrees of freedom
```

16차 k(다항함수형태)



Knots(df)의 개수 추정하여 적용

```
lm_ss2 <- smooth.spline(built_age, price_pyung,
                        cv = TRUE)
# ordinary leave-one-out (TRUE) or 'generalized' cross-valid
```

일반화가법모형: GAM

- 실습1. 층수(floor_no)와 아파트 평당 거래가격(price_pyung)의 관계를 비선형으로 가정하고, GAM으로 추정하고, 선형회귀 결과와 비교하여 보자. (1)

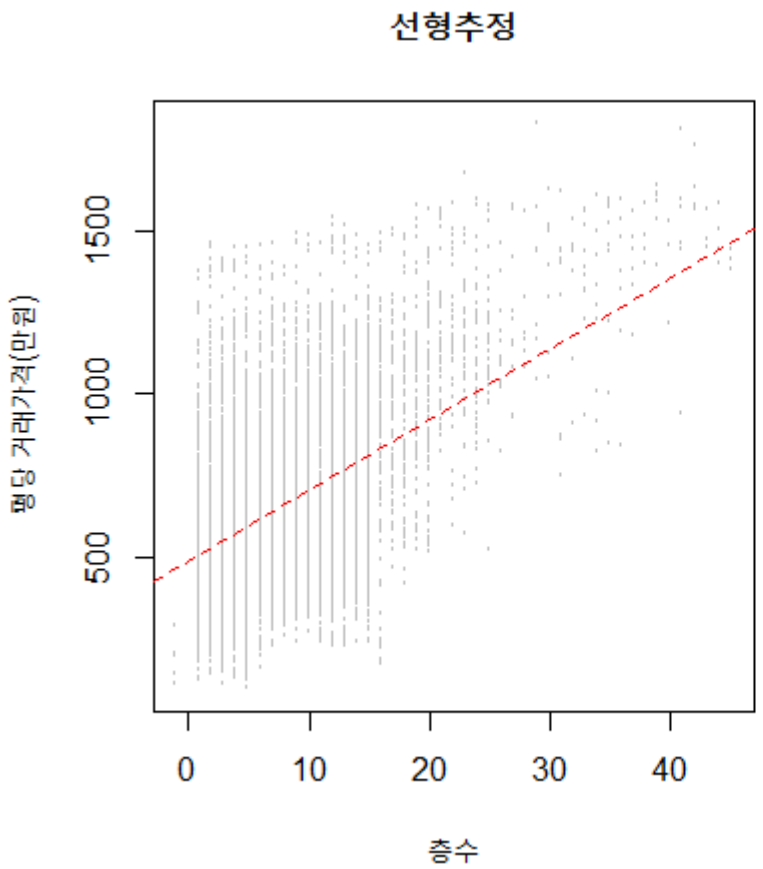
```
> lm <- lm(price_pyung ~
+              floor_no)
> summary(lm)

Call:
lm(formula = price_pyung ~ floor_no)

Residuals:
    Min      1Q  Median      3Q     Max
-658.60 -178.06  -37.93  157.76  932.06

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 489.0121     4.4557  109.75   <2e-16 ***
floor_no     21.6320     0.4242   50.99   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 242.5 on 8846 degrees of freedom
Multiple R-squared:  0.2272,    Adjusted R-squared:  0.2271
F-statistic:  2600 on 1 and 8846 DF,  p-value: < 2.2e-16
```

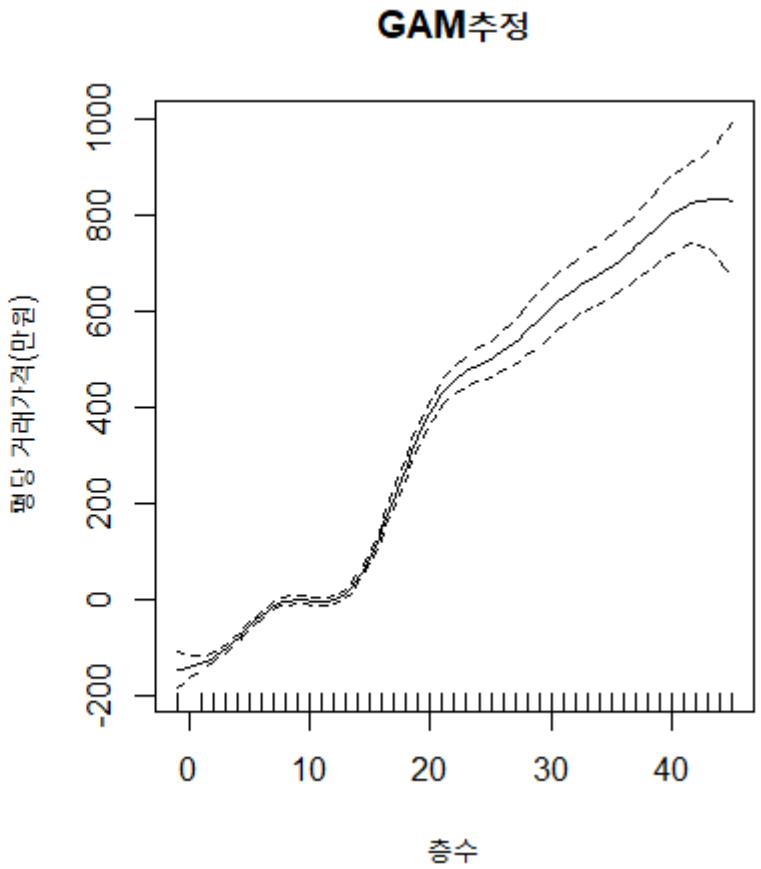
```
> lm_gam <- gam(price_pyung ~ s(floor_no), # 층수 =
+               data=apt4)
> summary(lm_gam) # gam 결과 반환

Family: gaussian
Link function: identity

Formula:
price_pyung ~ s(floor_no)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  674.331      2.517   267.9   <2e-16 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.

Approximate significance of smooth terms:
               edf Ref.df     F p-value
s(floor_no) 7.783  8.466 372.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.

R-sq.(adj) =  0.263   Deviance explained = 26.4%
GCV =  56128  Scale est. = 56072     n = 8848
```

일반화가법모형: GAM

- 실습1. 층수(floor_no)와 아파트 평당 거래가격(price_pyung)의 관계를 비선형으로 가정하고, GAM으로 추정하고, 선형회귀 결과와 비교하여 보자. (2)

```
> par(mfrow=c(1,2))
> plot(price_pyung ~
+        floor_no,
+      cex=0.3,
+      col="gray",
+      xlab ="층수",
+      ylab = "평당 거래가격(만원)",
+      main = "선형추정")
> abline(lm,
+        lty =2,
+        col="red")

> plot(lm_gam, se=TRUE, # gam graph
+      xlab ="층수",
+      ylab = "평당 거래가격(만원)",
+      main = "GAM추정")
```





일반화가법모형: GAM

- 실습2. 층수(floor_no)와 아파트 평당 거래가격(price_pyung)의 관계를 비선형으로 가정하고, GAM으로 추정하고, 이후 선형관계로 가정한 추가 독립변수로 built_age와 area_m2를 차례로 투입하여 그 결과를 비교하여 보자. (1)

```
> lm_gam1 <- gam(price_pyung ~ s(floor_no), # 층수 =
+                data=apt4)
> summary(lm_gam1) # gam 결과 반환

Family: gaussian
Link function: identity

Formula:
price_pyung ~ s(floor_no)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  674.331      2.517   267.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Approximate significance of smooth terms:
             edf Ref.df     F p-value
s(floor_no) 7.783  8.466 372.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

R-sq.(adj) =  0.263   Deviance explained = 26.4%
GCV =  56128  Scale est. = 56072     n = 8848
```

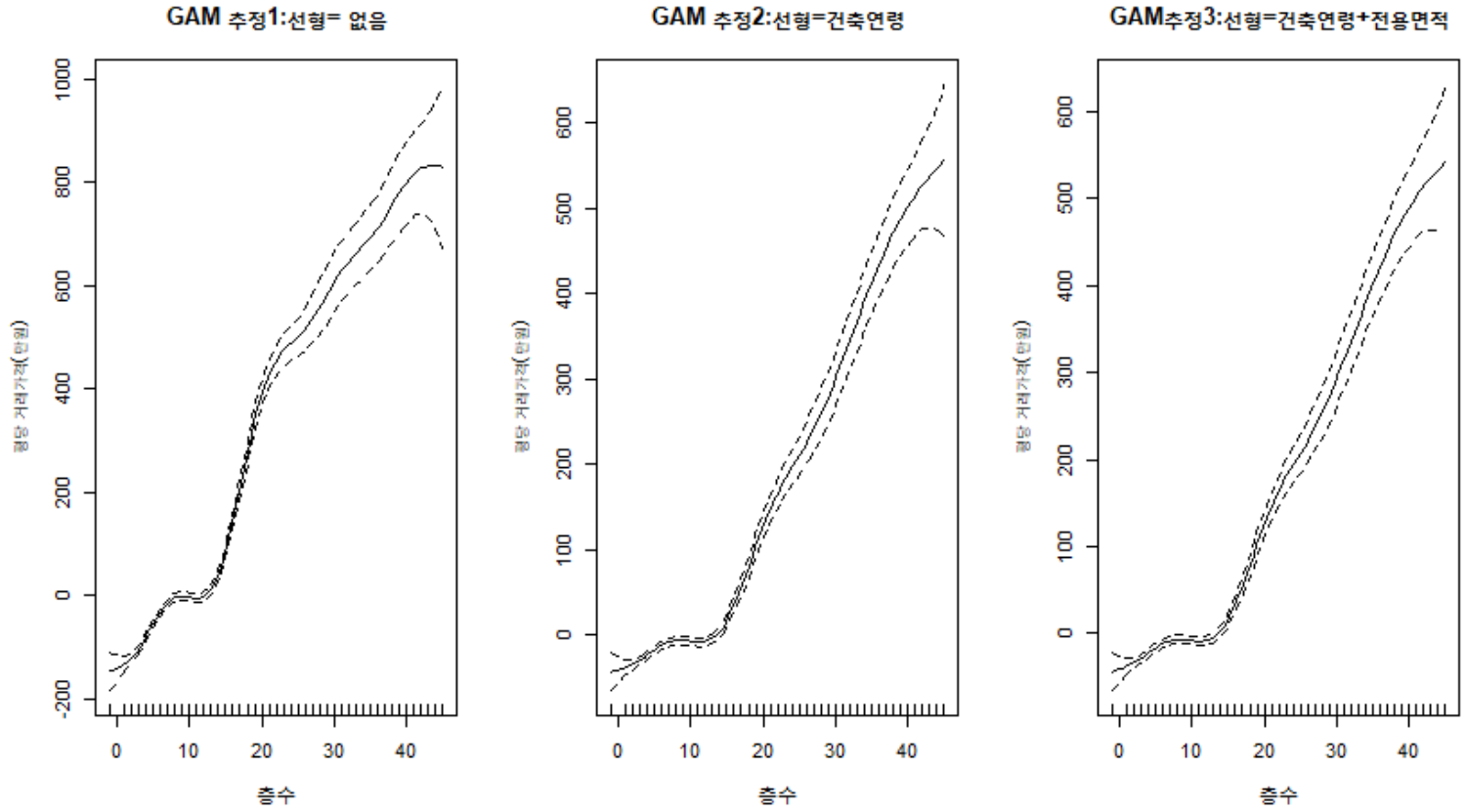
```
> lm_gam2 <- gam(price_pyung ~
+                s(floor_no) + built_age) # 선형 건축연령 독립변
> summary(lm_gam2)

Family: gaussian
Link function: identity

Formula:
price_pyung ~ s(floor_no) + built_age

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1079.8149     3.7260   289.8   <2e-16 ***
built_age    -22.0085     0.1838  -119.8   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Approximate significance of smooth terms:
             edf Ref.df     F p-value
s(floor_no) 7.03  7.874 174.5  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

R-sq.(adj) =  0.718   Deviance explained = 71.9%
GCV =  21439  Scale est. = 21417     n = 8848
```

```
> lm_gam3 <- gam(price_pyung ~
+                s(floor_no) + built_age + area_m2)
> summary(lm_gam3)

Family: gaussian
Link function: identity

Formula:
price_pyung ~ s(floor_no) + built_age + area_m2

Parametric coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) 1033.11017    6.35693  162.517   <2e-16 ***
built_age    -21.67204    0.18668 -116.091   <2e-16 ***
area_m2        0.58471    0.06462    9.049   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Approximate significance of smooth terms:
             edf Ref.df     F p-value
s(floor_no) 6.979  7.832 167.6  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

R-sq.(adj) =  0.721   Deviance explained = 72.1%
GCV =  21247  Scale est. = 21223     n = 8848
```

일반화가법모형: GAM

- 실습2. 층수(floor_no)와 아파트 평당 거래가격(price_pyung)의 관계를 비선형으로 가정하고, GAM으로 추정하고, 이후 선형관계로 가정한 추가 독립변수로 built_age와 area_m2를 차례로 투입하여 그 결과를 비교하여 보자. (2)

```
> par(mfrow=c(1,3))
> plot(lm_gam1, se=TRUE, # gam graph
+      xlab ="층수",
+      ylab = "평당 거래가격(만원)",
+      main = "GAM 추정1:선형= 없음")
> plot(lm_gam2, se=TRUE,  # gam graph
+      xlab ="층수",
+      ylab = "평당 거래가격(만원)",
+      main = "GAM 추정2:선형=건축연령")
> plot(lm_gam3, se=TRUE,
+      xlab ="층수",
+      ylab = "평당 거래가격(만원)",
+      main = "GAM추정3:선형=건축연령+전용면적")
```



연습문제 11

- 아파트 층수(floor_no)가 평당 아파트 거래가격(price_pyung)에 미치는 영향이 아파트 가격의 분위별로 다를 것으로 예상된다. 75분위, 50분위, 25분위인 아파트 평당 가격 구간별로 층수가 아파트 평당 거래가격에 미치는 영향을 분위회귀모형의 회귀계수로 확인하여 보자.

  - 75분위 구간과 25분위 구간의 회귀계수를 적으시오.

  - 이 결과로 볼 때, 아파트 평당 가격은 비싼 아파트 일수록 높은 층수에서 아파트 가격이 보다 비싸게 거래되는 지를 진단하시오.

연습문제 12

- 아파트 전용면적(area_m2)이 클수록 가격은 높다고 알려져 있다. apt4 데이터로 이에 대한 이차다항식으로 단순회귀모형을 적용하여 그 결과를 확인하시오.

연습문제 13

- 층수(floor_no)와 건축연령(built_age)가 아파트 평당 거래가격(price_pyung)의 관계를 비선형으로, 전용면적(area_m2)를 선형으로 가정하여 GAM으로 추정하여 보자. 그리고 이들 세 변수 모두 선형으로 가정한 선형회귀 분석결과와 비교하여 보자.

요약

- 평균 차이 검정
  - 평균차이 검정 방법론 개요
  - 데이터 가공하기
  - 두 집단 평균차이 검정: t-test
    - 독립 두 표본 평균 검정
    - 대응 두 표본 평균 검정
    - 한 표본 평균 일치 검정
  - 둘 이상의 평균차이 검정: ANOVA
- 상관관계분석
  - 상관관계의 개념
  - 상관관계 검정절차
  - 상관관계 검정방법
  - 다중 상관관계와 산점도 매트릭스
- 선형회귀와 예측
  - 회귀와 예측 모형의 개요
  - 단순선형회귀모형
  - 다중선형회귀모형
    - 표준화회귀계수(Beta)
  - 상호작용과 주효과
  - 회귀모형 비교와 최적 모형 선택
    - 모형비교1: adj.R2, AIC, BIC
    - 모형비교2: 분산분석(anova)
    - 모형선택: 단계별 선택법(Stepwise regression): “MASS”, “leaps”
- 비선형 회귀와 예측
  - 비선형 회귀모형의 개요
  - 다항선형회귀모형
  - 분위회귀모형
  - (스무딩)스플라인회귀모형
  - 일반화가법모형(GAM)

끝

- 질의와 토의(Question & Discussion)
  - 이번 강의 내용을 시청하고, 실행하면서 궁금한 점이나 어려운 점에 대하여 토의해봅시다.

- 다음 차 강의주제
  - **분류분석과 머신러닝 기법**