빅데이터 분석

[R 데이터분석 - 회귀분석]

BigData Analysis

Open API를 활용한 공공데이터 가져오기

```
## XML 패키지설치
install.packages("XML")
library(XML)
## 서비스 URL
serviceURL <- "http://apis.data.go.kr/1160100/service/GetFnCoBasilnfoService/"
## 오퍼레이션명
operation <- "getFnCoOutl"
## 인증키
ServiceKey <- “인증키“ # 인증키는 두개(encoding, decoding) 모두 사용 가능
```

Open API를 활용한 공공데이터 가져오기

```
## 페이지 위치 지정
pageNo <- 1
## 오픈 API 호출 시 얻게 되는 데이터 개수 지정
rows <-10
## 추출 데이터 포맷 지정
type_data_format <- "xml"
## 오픈 API 호출을 위한 최종 URL 생성
url <- paste0(serviceURL, operation, paste0("?serviceKey=", ServiceKey),
paste0("&pageNO=", pageNo), paste0("&numOfRows=", rows), paste0("&resultType=",
type_data_format))
```


Open API를 활용한 공공데이터 가져오기

```
## 오픈 API 호출
xmlDocument <- xmlTreeParse(url, useInternalNodes = TRUE, encoding = "UTF-
8")
## xml root node 획득
rootNode <- xmlRoot(xmlDocument)
## 오픈 API 호출 결과 데이터 개수 확인
numOfRows <- as.numeric(xpathSApply(rootNode, "//numOfRows", xmlValue))
## 전체 데이터 개수 확인
totalCount <- as.number(xpathSApply(rootNode, "//totalCount", xmlValue))
```

Open API를 활용한 공공데이터 가져오기

```
## 총 오픈 API 호출 횟수 계산
loopCount <- round(totalCount/numOfRows, 0)
## 전체 데이터를 저장할 변수 선언
finalTotalData <- data.frame()
## 오픈 API 호출 횟수 계산
if(loopCount*numOfRows < totalCount) {
loopCount <-loopCount + 1
}
```


Open API를 활용한 공공데이터 가져오기

```
# 데이터 확인하기
View(finalTotalData)
# CSV 파일로 저장하기
write.csv(finalTotalData, "OpenAPIData.csv", row.names= FALSE)
```

회귀분석 이해

회귀분석

내 키가 184cm인데...
미래에 우리 아들의 키는 몇 센치가 될까?

회귀분석 시초

F. Galton(1822~1911)의 “아버지와 아들의 키 연관성 연구”

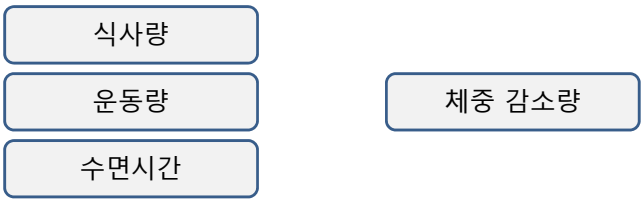
- 아버지의 키가 클 수록, 아들의 키도 크다
- 아버지의 키가 작을 수록, 아들의 키도 작다
- 아버지의 키가 매우 커도, 아들의 키가 매우 작지는 않다
- 아버지의 키가 매우 작아도, 아들의 키가 매우 작지는 않다

그렇다면 대체 아버지와 아들 간의 키에는 어느정도 상관관계가 있는가?
-> 아들의 키가 아버지의 키 수준으로 얼마나 “회귀”하는지 찾기 위한 연구

회귀분석

회귀분석 관련 용어 정리

- 독립변수 : 어떠한 현상을 설명할 때, 현상의 발생에 영향을 미치는 요인
- 종속변수 : 독립변수의 영향에 따라 결정되는 요인
- 예측모형 : 독립변수와 종속변수에 해당하는 자료를 모아 관계를 분석하고, 이를 예측할 수 있는 통계적 방법으로 정리한 것



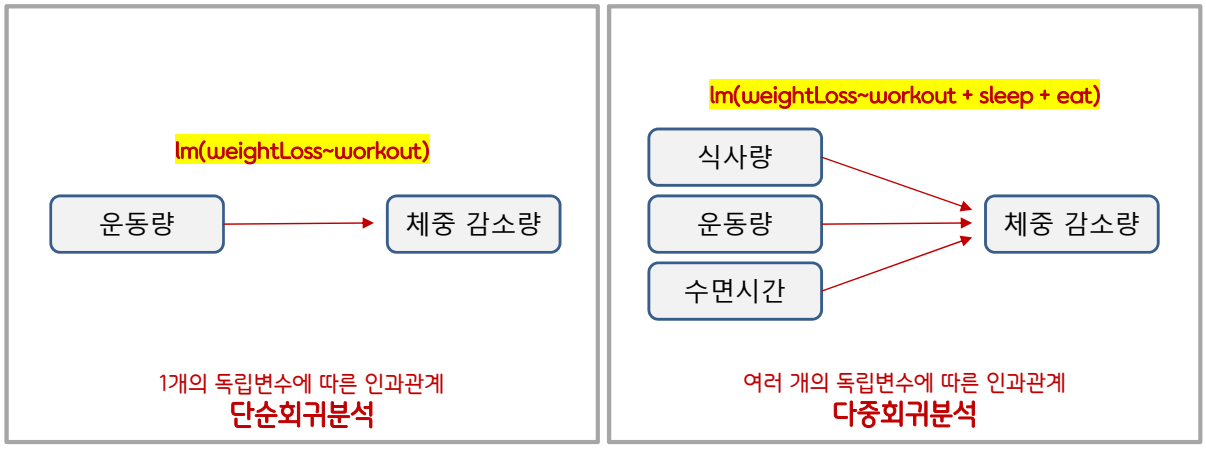
회귀분석

회귀분석 관련 용어 정리

- 회귀분석 : 회귀 이론을 기초로 독립변수가 종속변수에 미치는 영향을 파악하여 예측모형을 도출하는 통계적 방법
  - 회귀식 : 독립변수와 종속변수 사이의 관계를 수학식으로 표현
  - 단순회귀 : 독립변수의 1개인 경우
  - 다중회귀 : 독립변수가 2개 이상인 경우
  - 로지스틱 회귀 : 종속변수의 값의 형태가 연속형 숫자가 아닌 범주형 값인 경우, 이를 분석하기 위해 사용하는 통계적 방법
  - 분류 : 데이터로부터 어떠한 범주를 예측하는 작업

회귀분석

단순회귀분석 vs 다중회귀분석



회귀분석

회귀분석을 위한 lm() 함수

lm() 함수 개요

- 회귀분석에 필수적인 함수로서, 회귀분석 전반에 필요한 도구들을 제공하는 함
- 선형회귀 모형(Linear regression Modelling)라는 의미에서 lm함수로 명칭

lm(formula, data)

- formula : 회귀분석에 필요한 독립변수 & 종속변수를 입력
  - 회귀모형에서 무엇이 독립변수이고 무엇이 종속변수인지 지정하는 것으로, ~ 앞에 있는 게 종속변수이고 ~ 뒤에 있는 게 독립변수이다. 독립변수가 여러 개이면 +으로 연결한다.
  - 단순회귀분석 -> 종속변수 ~ 독립변수
  - 다중회귀분석 -> 종속변수 ~ 독립변수1 + 독립변수2 + … + 독립변수n
- data : 회귀분석 대상(데이터프레임) 입력

단순회귀분석

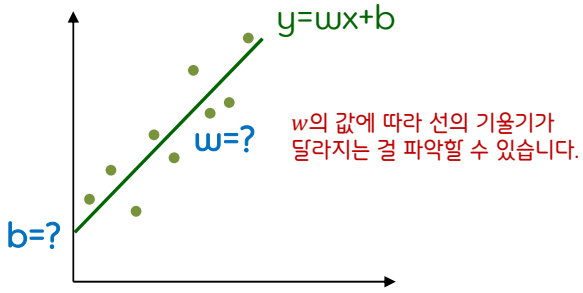
단순회귀분석

단순회귀분석 개념

- 독립변수(x)와 종속변수(y) 사이의 선형관계를 파악하여 예측에 활용하는 통계적 방법
- 독립변수와 종속변수에 대해 수집한 데이터를 활용하여, 인과관계를 가장 잘 설명하는 w와 b를 찾는 게 단순회귀분석의 목표
- 단순회귀식

  y=wx + b (w, b는 상수)

  - x -> 독립변수(영향을 주는 값)
  - y -> 종속변수(영향을 받는 값)
  - w -> 단순회귀선의 기울기
  - b -> 단순회귀선의 절편(y축과 단순회귀선이 닿는 지점)



단순회귀분석

단순회귀분석 개념

- **(예시) 신입생의 수능 성적으로 대학 논술 성적을 예측하기**
  - 독립변수(x) : 수능 성적
  - 종속변수(y) : 대학 논술 성적
  - 목표

    수능 성적과 대학 논술 성적간의 인과관계를 설명해줄 단순회귀식($y = wx + b$)에 w와 b를 찾아 채워넣어, 단순회귀식을 완성하는 것

단순회귀분석

단순회귀분석 실습

만일 차량 중량을 x를 가진 차량이 있을 때, 연비 y가 어느정도 일까?

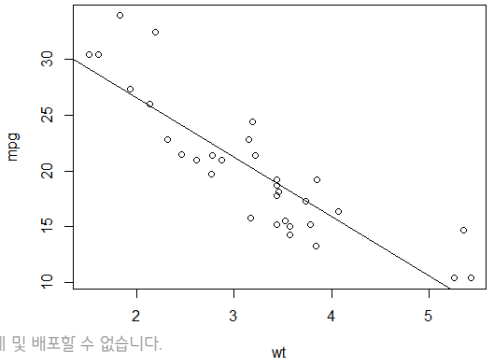
1. 기존의 mtcars 데이터로 단순회귀분석 모형 만들기
2. 해당 모형으로 새로운 x값(차량중량)을 대입해보고, y(연비) 예측해보기
3. 해당 모형에 기존 mtcars 데이터를 대입해보고, 회귀모형 예측 값과 실제 값 간의 차이(오차)를 구하여 회귀모형의 신뢰도 알아보기

단순회귀분석

단순회귀분석 예측모형 만들기

(기존 mtcars 데이터를 통해, 차량 중량을 바탕으로 연비를 예측하는 모형 만들기)

```
data(mtcars)
plot(mpg~wt, data = mtcars)  #차량 중량(x)과 연비(y) 간의 산점도를 통해 선형관계 확인

model <- lm(mpg~wt, mtcars) #회귀모형 생성하기
abline(model)  #회귀선을 산점도 위에 표시

coef(model)[1]  # 회귀결과 추출 : b값을 출력(37.28)
coef(model)[2]  # 회귀결과 추출 : w값을 출력(-5.34)
```



차량중량 - 연비 단순회귀분석 회귀식

$$mpg = -5.34 * wt + 37.28$$

단순회귀분석

단순회귀분석 예측모형 만들기

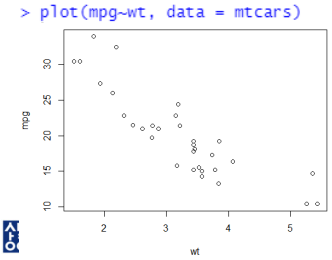
1. mtcars 데이터 확인

```
> data(mtcars)
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

head (데이터프레임)
- head는 데이터의 **처음 6행의** 값을 보여주는 함수

2. 차량 중량(wt)와 연비(mpg) 사이의 관계를 나타내는 산점도 그리기

```
> plot(mpg~wt, data = mtcars)
```



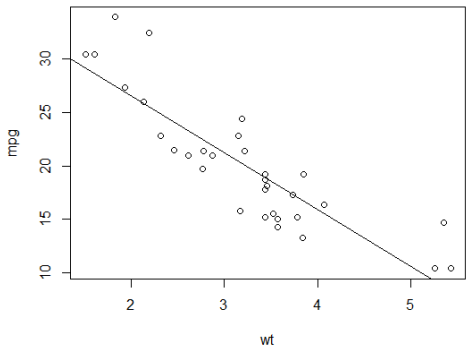
plot (y축변수~x축변수, data=데이터프레임)
- 변수 분포를 보여주는 함수

단순회귀분석

단순회귀분석 실습

3. 예측모형 생성 & 회귀선 그리기

```
> model <- lm(mpg~wt, mtcars)
> abline(model)
```



lm (y축변수~x축변수, 데이터프레임)
- 두변수의 선형관계를 나타내는 선의 식 (회귀식)을 찾는 함수

abline(회귀식)
- 산점도 위에 회귀선을 그리는데 사용

단순회귀분석

단순회귀분석 실습

4. 회귀식($y = wx + b$)에 필요한 회귀계수(w) 및 회귀상수(b) 구하기

```
> coef(model)[1]
(Intercept)
   37.28513
> coef(model)[2]
       wt
-5.344472
```

회귀상수(b)

회귀계수(w)

coef(회귀식) / coef(lm(y축변수~x축변수))
- 회귀 계수를 추출하는 함수 (b값 w값) # y=wx+b에서 w와 b

coef(회귀식) [1] / coef(lm(y축변수~x축변수)) [1]
- 회귀계수로 사용되는 b값 추출

coef(회귀식) [2] / coef(lm(y축변수~x축변수)) [2]
- 회귀계수로 사용되는 w값 추출

단순회귀분석

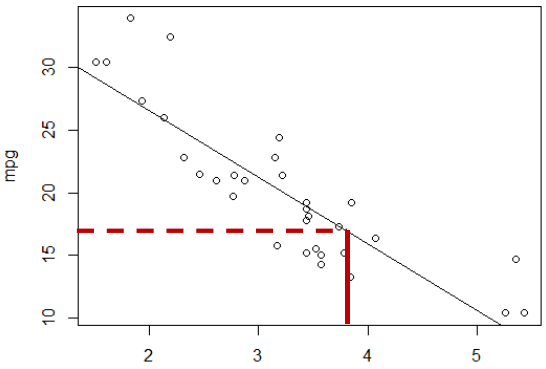
단순회귀분석 회귀모형으로 예측값 구하기

(새로운 차량 중량 값을 대입하여, 연비 값을 예측해보기)

```
b <- coef(model)[1]  #b값 대입
w <- coef(model)[2]  #w값 대입
wtSample <- 3.8  #예측하고자 하는 독립변수 대입

equation <- w * wtSample + b  #회귀식 만들기
print(equation)  #회귀식에 독립변수 대입한 결과 출력
```

만일 차량 무게가 3.8 파운드일 때,
연비가 갤런 당 16.7 마일이라 예측할 수 있음

$mpg = -5.34 * 3.8 + 37.28 = 16.97$



단순회귀분석

단순회귀분석 회귀모형으로 예측값 구하기

1. 회귀계수 & 회귀상수를 각각 변수에 대입하기

```
> b <- coef(model)[1]
> w <- coef(model)[2]
```

2. 예측하고자 하는 값(차량 중량)을 입력하고, 회귀식을 만들어 대입

```
> wtSample <- 3.8
> equation <- w * wtSample + b
```

3. 예측 값(연비) 결과 확인

```
> print(equation)
       wt
16.97613
```

단순회귀분석

회귀모형 오차 구하기

(기존 실제 데이터와 회귀모형 예측 데이터 간의 차이 구하기)

```
wtData <- mtcars[,"wt"]  #전체 차량 중량 데이터 선택
mpgPred <- w * wtData + b  #wtData를 회귀선에 대입하여 전체 차량 연비 예측값 도출
mpgData <- mtcars[,  " mpg " ]  #전체 차량 연비 데이터 선택

compare <- data.frame(mpgPred, mpgData, mpgPred - mpgData)
#차량 연비 예측값, 차량 연비 실제값, 예측값과 실제값 간의 차이 계산값을 담은 데이터프레임 생성
colnames(compare) <- c( " 예상 " ,  " 실제 " ,  " 오차 " )  #데이터프레임 열 이름 재정의
head(compare)
```


단순회귀분석

회귀모형 오차 구하기

1. 독립변수 데이터(차량 중량)들을 모두 선택하여, 예측 값들을 도출

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4

$$mpgPred = w * wtDATA + b$$

	mpgPred
1	23.282611
2	21.919770
3	24.885952
4	20.102650
5	18.900144
6	18.793255
7	18.205363
8	20.236262
9	20.450041
10	18.900144
11	18.900144
12	15.533127
13	17.350247
14	17.083024
15	9.226650
16	8.296712
17	8.718926
18	25.527289
19	28.653805
20	27.478021

차량 중량 데이터 — 예측모형 대입 — 예측 연비 값 도출

단순회귀분석

회귀모형 오차 구하기

2. 독립변수 데이터(차량 중량)들을 모두 선택하여, 예측 값들을 도출

	mpgPred
1	23.282611
2	21.919770
3	24.885952
4	20.102650
5	18.900144
6	18.793255
7	18.205363
8	20.236262
9	20.450041
10	18.900144
11	18.900144
12	15.533127
13	17.350247
14	17.083024
15	9.226650
16	8.296712
17	8.718926
18	25.527289
19	28.653805
20	27.478021

예측 연비 값

−

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4

실제 연비 데이터 (mpgData)

$$mpgPred - mpgData$$

=

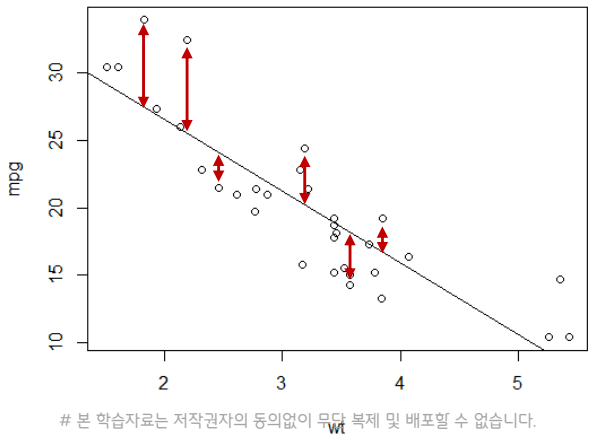
	mpgPred...mpgData
1	2.2826106
2	0.9197704
3	2.0855521
4	-1.2973499
5	0.2001440
6	0.6932545
7	3.9053627
8	-4.1637381
9	-2.3499593
10	-0.2998560
11	1.1001440
12	-0.8668731
13	0.0502472
14	1.8830236
15	-1.1733496
16	-2.1032876
17	-5.9810744
18	-6.8727113
19	-1.7461954
20	-6.4219792

예측 오차

단순회귀분석

회귀모형 오차 구하기

3. 회귀모형의 예측값과 실제값의 차이



단순회귀분석

회귀모형 오차 구하기

1. 독립변수 데이터(차량 중량)들을 모두 선택하여, 예측 연비 값들을 도출

```
> wtData <- mtcars[,"wt"]
> mpgPred <- w * wtData + b
```

2. 종속변수 데이터(실제 연비 값)들을 모두 선택하여, 예측 연비 값 - 실제 연비 값 계산을 실행해 오차 도출

```
> mpgData <- mtcars[, "mpg"]
```

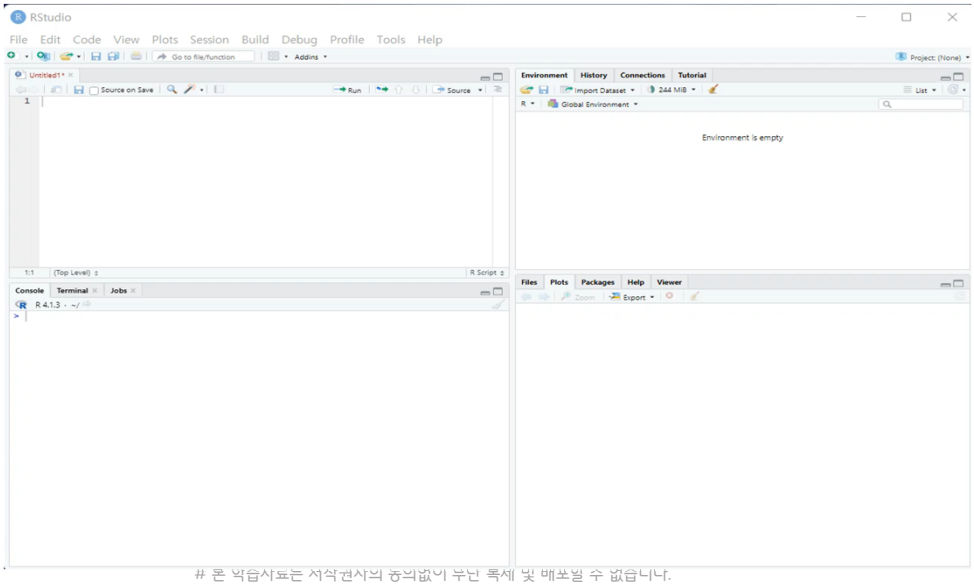

단순회귀분석

회귀모형 오차 구하기

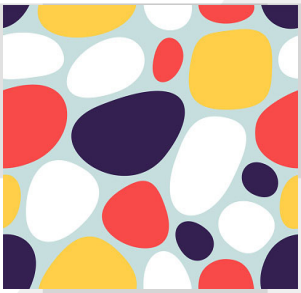
3. 예측 연비 값, 실제 연비 값, 오차 값을 모두 데이터프레임으로 담고, 열 명칭 수정

```
> compare <- data.frame(mpgPred, mpgData, mpgPred - mpgData)
> colnames(compare) <- c("예상", "실제", "오차")
> head(compare)
       예상 실제       오차
1 23.28261 21.0  2.2826106
2 21.91977 21.0  0.9197704
3 24.88595 22.8  2.0859521
4 20.10265 21.4 -1.2973499
5 18.90014 18.7  0.2001440
6 18.79325 18.1  0.6932545
```

단순회귀분석



다중 회귀분석



다중회귀분석

다중회귀분석 개념

- 여러 개의 독립변수(x)와 종속변수(y) 사이의 선형관계를 파악하여, 예측에 활용하는 통계적 방법
- 여러 개의 독립변수와 종속변수에 대해 수집한 데이터를 활용하여, 인과관계를 가장 잘 설명하는 w와 b를 찾는 게 다중회귀분석의 목표
- 다중회귀식

  $y=w_1x_1 + w_2x_2 + w_3x_3 + \dots w_nx_n + b$ (w, b는 상수)

  - x_n -> 독립변수(영향을 주는 값)
  - y -> 종속변수(영향을 받는 값)
  - w_n -> 회귀계수(회귀선의 기울기)
  - b -> 회귀상수(y축과 회귀선이 닿는 지점)

다중회귀분석

다중회귀분석 개념

- **(예시) 신입생의 수능 성적, 출석일 수, 상장 개수로 대학 논술 성적을 예측하기**
  - 독립변수(x) : 수능 성적, 출석일 수, 상장 개수
  - 종속변수(y) : 대학 논술 성적
  - 목표

    수능 성적, 출석일 수, 상장 개수와 대학 논술 성적 간의 인과관계를 설명해줄 다중회귀식($y = w_1x_1 + w_2x_2 + w_3x_3 + b$)에 w와 b를 찾아 채워 넣어, 다중회귀식을 완성하는 것

다중회귀분석

다중회귀분석 주의사항

- **독립변수와 종속변수 간의 높은 상관관계**
  예를 들어 학생의 평균 칫솔질 횟수(독립변수)으로 대학 논술 성적(종속변수)을 예측하려는 건 비논리적
  설령 유의도가 높게 측정되더라도, 이를 가지고 인과관계가 있다고 할 수 없음
  참고로 독립변수와 종속변수 간의 관계에 대한 논문 등 자료로 논리성 근거 확보가 필요
- **선택한 독립변수 간에는 서로 낮은 상관관계를 보여야 함**
  만일 수능 성적과 내신 성적을 둘 다 독립변수로 설정할 경우, 두 변수 간에 상관관계가 높기 때문에(즉 내신 성적이 높은 경우에 수능 성적이 높은 양의 상관 관계) 다중공선성 문제가 발생할 수 있음.
- **독립변수 개수는 적을 수록 유리함**
  회귀분석 모형 복잡도를 낮출 수록 복잡도가 하락하여 예측 성능 보장

다중회귀분석

다중회귀분석 주의사항 - 다중공선성(Multicollinearity)

- 독립변수들이 서로 간에 강한 상관관계가 있어 상호 영향을 주기에, 종속변수 예측 값에 부정적인 영향을 주는 현상
- 일례로 수능 점수(독립변수 #1)와 내신 점수(독립변수 #2)를 사용하여 대학논술 성적(종속변수)를 예측하려 할 때
  - 수능 점수와 내신 점수 간에는 강력한 양의 상관관계가 있기 때문에, 서로 영향을 주고 있음
    (수능 점수가 높으면 내신 점수 높을 가능성 올라가며, 반대 사례도 마찬가지임)
  - 따라서 두 독립변수를 같이 사용할 경우, 다중공선성으로 종속변수 추정에 오류를 발생시킬 수 있음

```
sqrt(vif(fit))>2
```

다중회귀분석

다중공선성 확인

- 1. vif() 함수 이용 # vif : variance inflation factor

install.packages("car")

library(car)

vif(변수명)

sqrt(vif(변수명))>2 # 일반적으로 vif 제곱급>2 다중공선성 문제 있는 것으로 봄

다중회귀분석

다중공선성 확인

- 2. corrplot() 함수 이용 변수들간의 상관관계 확인 # 상관관계 행렬 correlation matrix plot

install.packages("corrplot")

library(corrplot)

corrplot(cor(변수명), method = "shade", addCoef.col = "black")

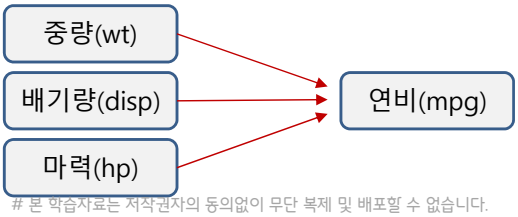
다중회귀분석

다중회귀분석 실습

만일 차량 중량 x1, 배기량 x2, 마력 x3를 가진 차량이 있을 때,

연비 y가 어느정도 일까?

1. 기존의 mtcars 데이터로 다중회귀분석 예측모형 만들어서 시각화하기
2. 결과값을 확인하고, 변수선택 진행하여 최종 회귀모형 도출



다중회귀분석

colnames(데이터프레임) <- c(“”, “”, “”,...)
- 데이터프레임 열 명칭 재설정

다중회귀분석 실습

```
data(mtcars)
df <- data.frame(mtcars$wt, mtcars$disp, mtcars$hp)
#독립변수 데이터들을 바탕으로 DF 생성
colnames(df) <- c( " 중량 " ,  " 배기량 " ,  " 마력 " )  #DF 열 명칭 재설정

plot(df, pch = 16, col = "blue", main = "산점도 매트릭스")  #3:3 산점도 매트릭스 그리기

model <- lm(mpg ~ wt + disp + hp, data = mtcars)  #다중회귀분석 예측모형 만들기
summary(model)  #예측모형 결과 도출
```

다중회귀분석

다중회귀분석 실습

1. mtcars 데이터 불러오고, 필요한 데이터들을 추출하여 DF 생성 및 열 명칭 재설정

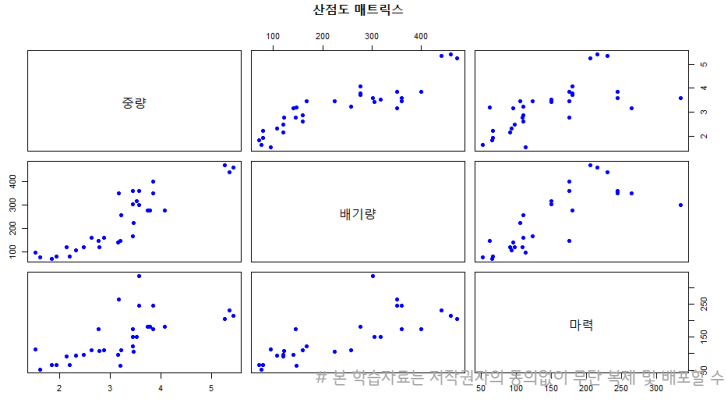
```
> data(mtcars)
> df <- data.frame(mtcars$wt, mtcars$disp, mtcars$hp)
> colnames(df) <- c("중량", "배기량", "마력")
```


다중회귀분석

다중회귀분석 실습

2. DF로 산점도 매트릭스 그리기(pch = 점 크기 설정, col = 색상, main = 제목)

```
> plot(df, pch = 16, col = "blue", main = "산점도 매트릭스")
```



다중회귀분석

다중회귀분석 실습

3. 다중회귀분석 예측모형을 만들고, 이를 summary() 함수를 통해 해석

```
> model <- lm(mpg ~ wt + disp + hp, data = mtcars)
> summary(model)
```

(다음 슬라이드에서 summary 해석)

다중회귀분석

summary() 함수를 통한 다중회귀분석 예측모형 해석

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

Residuals:
   Min     1Q Median     3Q    Max
-3.891 -1.640 -0.172  1.061  5.861

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
wt          -3.800891   1.066191  -3.565  0.00133 **
disp        -0.000937   0.010350  -0.091  0.92851
hp          -0.031157   0.011436  -2.724  0.01097 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

mpg를 설명하는 데 얼마나 중요한 독립변수인지 나타내며, 별(*) 개수가 많을 수록 유리함.
아무런 표시가 없는 경우 유의성이 없는 독립변수임.

R-square는 다중선형회귀모형이 mpg를 얼마나 잘 설명하는지 나타냄. 0.65 이상이면 잘 설명하는 것으로 간주함.

p-value는 예측모형이 얼마나 유의한지 나타냄.
0.05 미만이면 유의하다고 판단함.

다중회귀분석

다중회귀분석 실습

4. stepAIC() 함수를 통해 유의미한 독립변수 만으로 회귀모형을 만드는 변수선택을 진행하여, 새로운 회귀모형을 만들고 최종 회귀식 도출

* AIC Akaike Information Criterion

* install.packages("MASS")

변수선택을 거쳐, 새로운 회귀모형을 만들도록 합니다.

```
> library(MASS)
> newModel <- stepAIC(model)
Start:  AIC=65.83
mpg ~ wt + disp + hp

       Df Sum of Sq    RSS    AIC
- disp  1     0.057 195.05 63.840
<none>              194.99 65.831
- hp    1    51.692 246.68 71.356
- wt    1    88.503 283.49 75.806

Step:  AIC=63.84
mpg ~ wt + hp

       Df Sum of Sq    RSS    AIC
<none>              195.05 63.840
- hp    1    83.274 278.32 73.217
- wt    1   252.627 447.67 88.427
```

변수선택 과정에서 무의미한 disp 변수를 제거하고
다시 회귀모형을 만든 것을 알 수 있습니다.

다중회귀분석

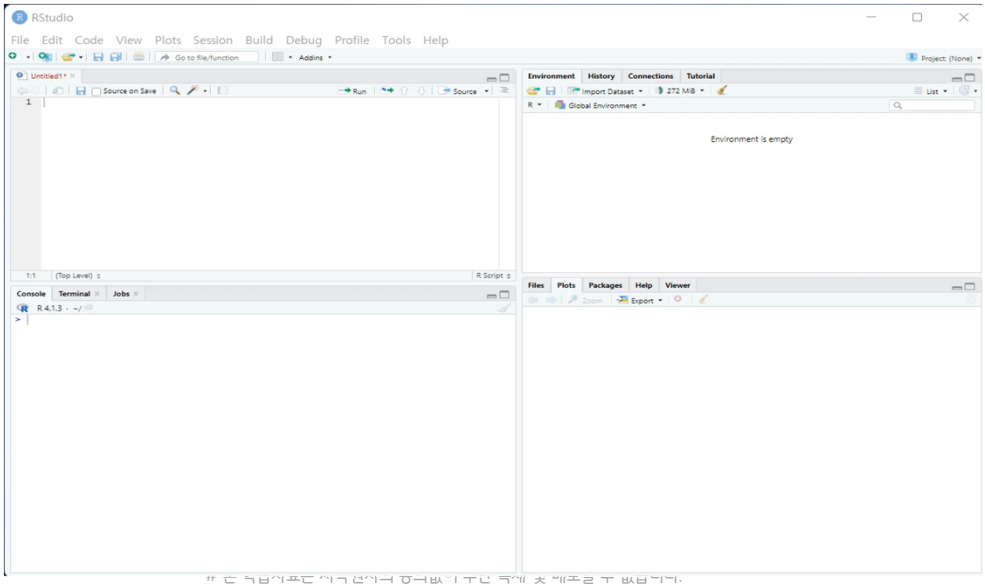
다중회귀분석 실습

```
> summary(newModel)

Call:
lm(formula = mpg ~ wt + hp, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

Estimate 열의 값들을 통해 최종 회귀식이
$mpg = (-3.87) * wt + (-0.03) * hp + 37.22$ 라는 걸
알 수 있습니다.

또한 변수선택법을 거쳐 만들어진 새로운 회귀모형이
R^2 및 P-value 모두 소폭 상승한 것을 확인할 수 있습니다.

다중회귀분석



SUMMARY

- **회귀분석 이해**
  - 독립변수 / 종속 변수 / 예측모형
  - 단순회귀 / 다중회귀
- **lm() 함수**
  - 회귀분석에 필수적인 함수
  - lm(formula, data) / lm(y축 변수~x축 변수, 데이터프레임)

SUMMARY

- **단순회귀분석**
  - 독립변수(x)와 종속변수(y) 사이의 선형관계를 파악하여 예측에 활용하는 통계적 방법
  - y=wx + b (w, b는 상수)

data(데이터프레임)
plot(y축 변수~x축 변수, data=데이터프레임)
lm(y축 변수~x축 변수, 데이터프레임)
abline(회귀식l)
coef(회귀식) [1] b값 출력 [2] w값 출력

SUMMARY

- **다중회귀분석**
  - 여러 개의 독립변수(x)와 종속변수(y) 사이의 선형관계를 파악하여,예측에 활용하는 통계적 방법
  - y=w1x1+ w2x2+ w3x3+ …wnxn+b (w, b는 상수)

data(데이터프레임)
data.frame(데이터프레임$변수1, 데이터프레임$변수2, 데이터프레임$변수3)
colnames(뉴데이터프레임) <- c(변경할 문구)
lm(y축 변수~x축 변수1+x축 변수2+x축 변수3+…,, 데이터프레임)
library(MASS)
stepAIC(회귀식)
summary(회귀모델)

연습문제

- trees 데이터셋에 대해 다음의 문제를 해결하는 R 코드를 작성하시오.
  - 다중선형 회귀모델을 이용하여 trees 데이터셋의 나무 둘레(Girth)와 나무의 키(Height)로 나무의 볼륨을 예측하시오.


```
mod <- lm(Volume~Girth+Height , data=trees)
summary(mod)
```

회귀모델

Volume = -57.9877 + 4.7082 * Girth + 0.3393 * Height

- lbench 패키지의 BostonHousing 데이터셋은 보스턴 지역의 지역 정보 및 평균 주택가격(medv) 정보가 저장되어 있다. 다른 변수들을 이용하여 medv를 예측하는 모델을 만드시오.(단 chas 변수는 모델을 만들 때 제외한다.)
  - 전체 변수를 이용하여 평균 주택가격(medv)을 예측하는 회귀모델을 만들고 회귀식을 나타내시오.


```
library(mlbench)
data(BostonHousing) # 데이터셋 불러오기

ds <- BostonHousing[,-4] # chas 제거
mod <- lm(medv~., data=ds)
summary(mod)
```

medv = 36.891960
−0.113139 * crim
+0.047052 * zn
+0.040311 * indus
−17.366999 * nox
+3.850492 * rm
+0.002784 * age
−1.485374 * dis
+0.328311 * rad
−0.013756 * tax
−0.990958 * ptratio
+0.009741 * b
−0.534158 * lstat

- mtcars 데이터셋에서 다른 변수들을 이용하여 연비(mpg)를 예측하는 다중 회귀모델을 만드시오.
  - 전체 변수를 이용하여 연비(mpg)를 예측하는 회귀모델을 만들고 회귀식을 나타내시오.

```
mod <- lm(mpg~., data=mtcars)
summary(mod)
```



```
# 회귀식
mpg = 12.30337
-0.11144 * cyl
+0.01334 * disp
-0.02148 * hp
+0.78711 * drat
-3.7153 * wt
+0.82104 * qsec
+0.31776 * vs
+2.52023 * am
+0.65541 * gear
-0.19942 * carb
-0.19942 * carb
```

감사합니다.