강원대학교
AI 소프트웨어학과

머신러닝2
- 다중 회귀분석(Simple) -

KNU 강원대학교

01 다중회귀 회귀(Multiple Linear Regression)

- 종속변수를 설명하기 위해서 두 개 이상의 독립변수가 사용되는 선형회귀모형을 다중선형 회귀모형이라 함

[예제] 예금 유치액을 설명하는데 도움을 주는 요인으로 홍보비용 뿐만 아니라 직원수, 지점의 크기 등을 고려한다면 더 정확한 정보를 얻을 수 있다.

01 다중회귀 회귀(Multiple Linear Regression)

- 종속변수를 설명하기 위해서 두 개 이상의 독립변수가 사용되는 선형회귀모형을 다중선형 회귀모형이라 함

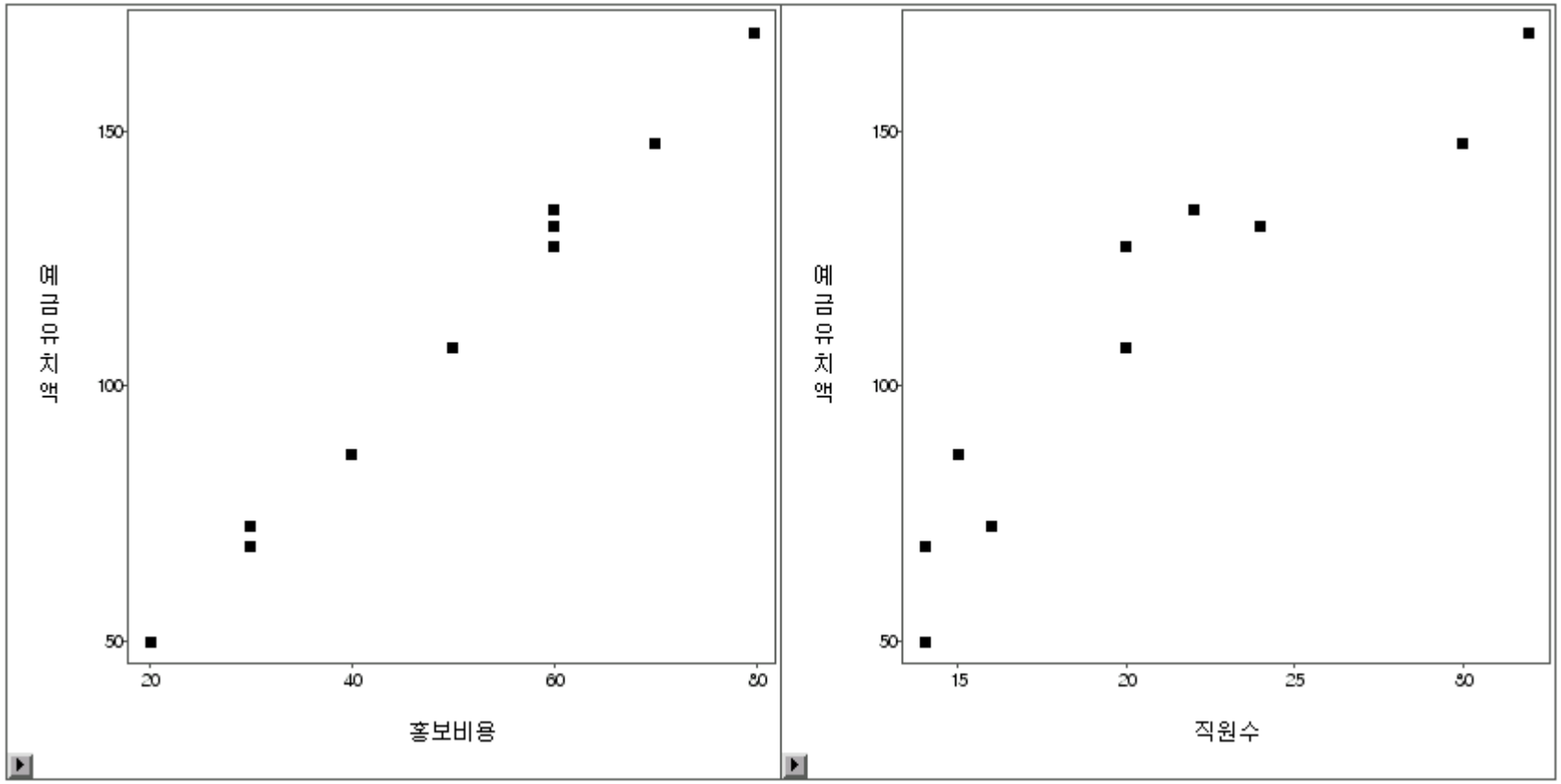
표본지점의 예금유치액

지점번호	홍보비용	직원수	예금유지액
1	40	15	87
2	50	20	108
3	30	14	69
4	60	22	135
5	70	30	148
6	60	24	132
7	30	16	73
8	60	20	128
9	20	14	50
10	80	32	170

Q) 홍보비용과 직원수가 예금유치액에 어떤 영향을 미치는가?

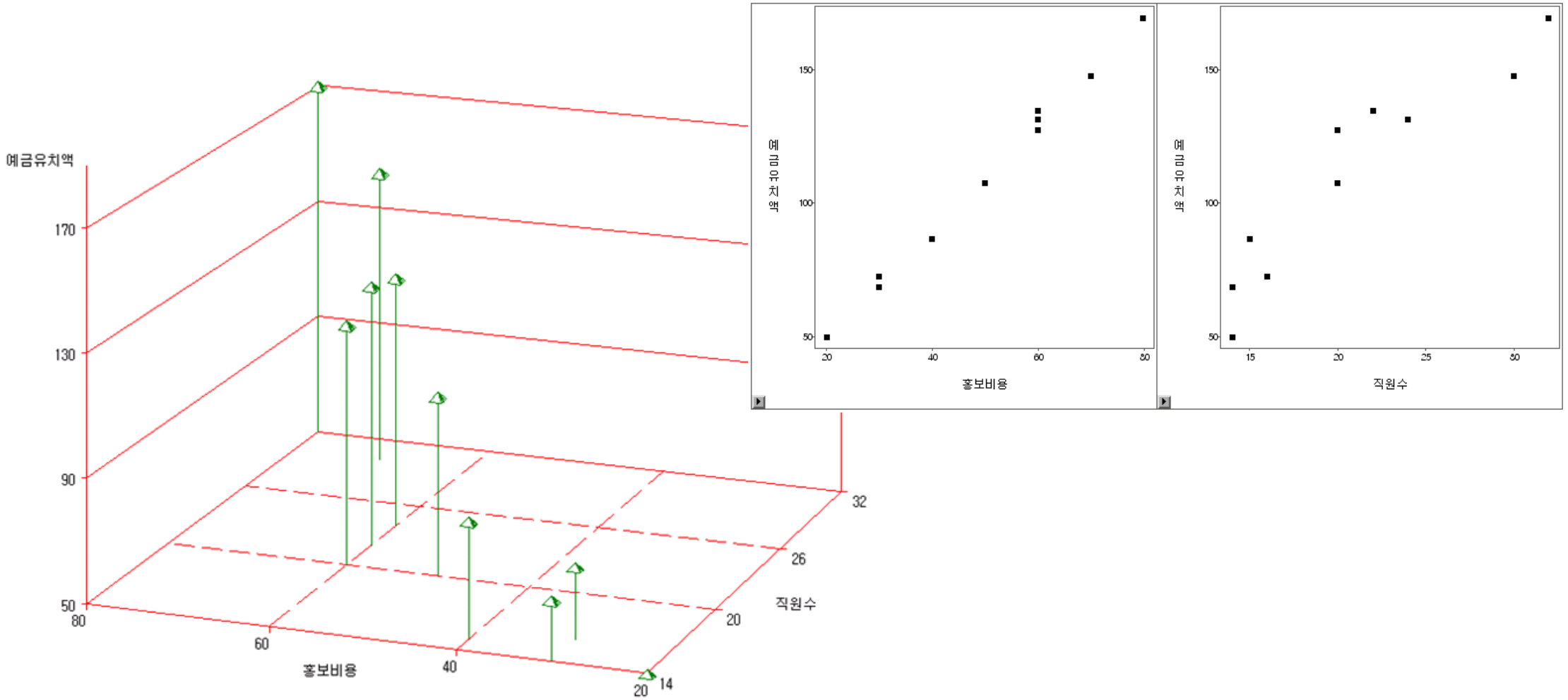
01 다중회귀 회귀(Multiple Linear Regression)

- 종속변수를 설명하기 위해서 두 개 이상의 독립변수가 사용되는 선형회귀모형을 다중선형 회귀모형이라 함



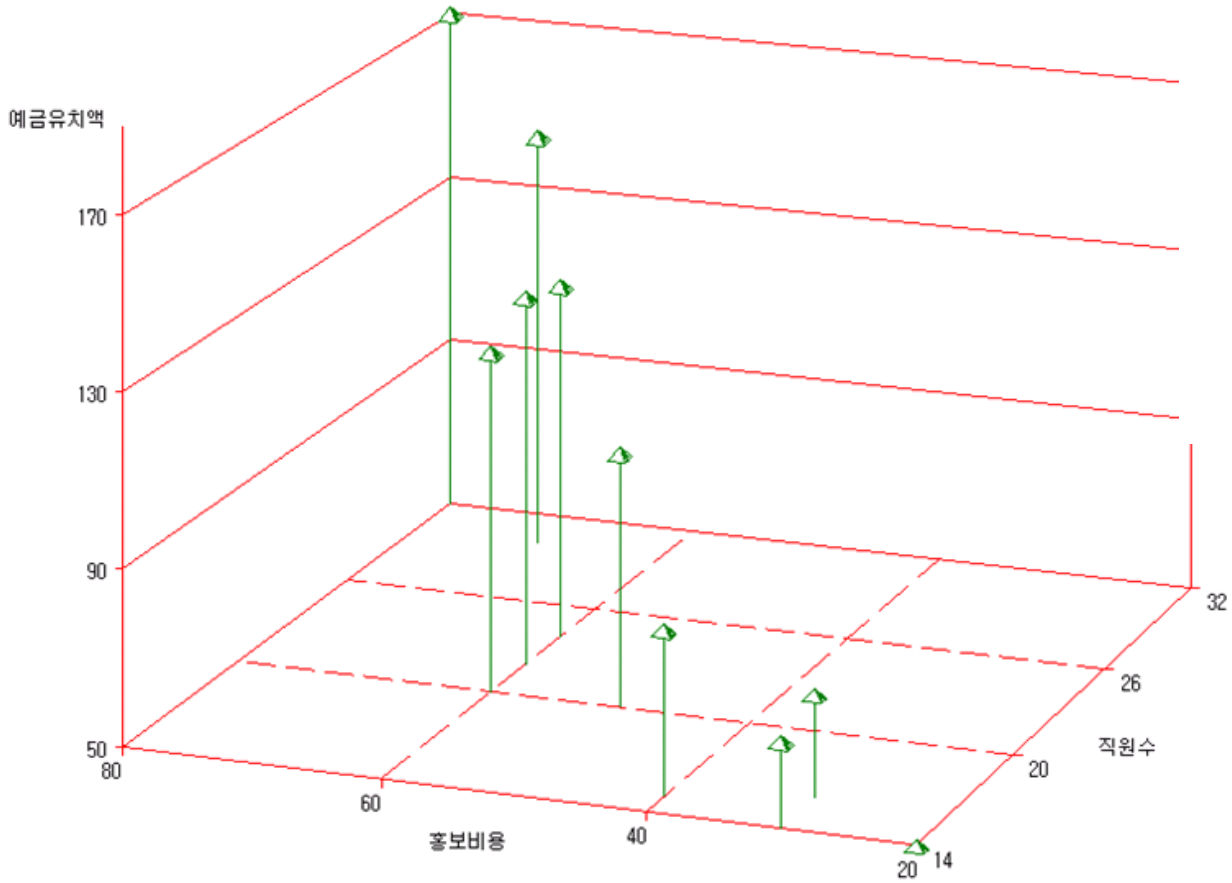
01 다중회귀 회귀(Multiple Linear Regression)

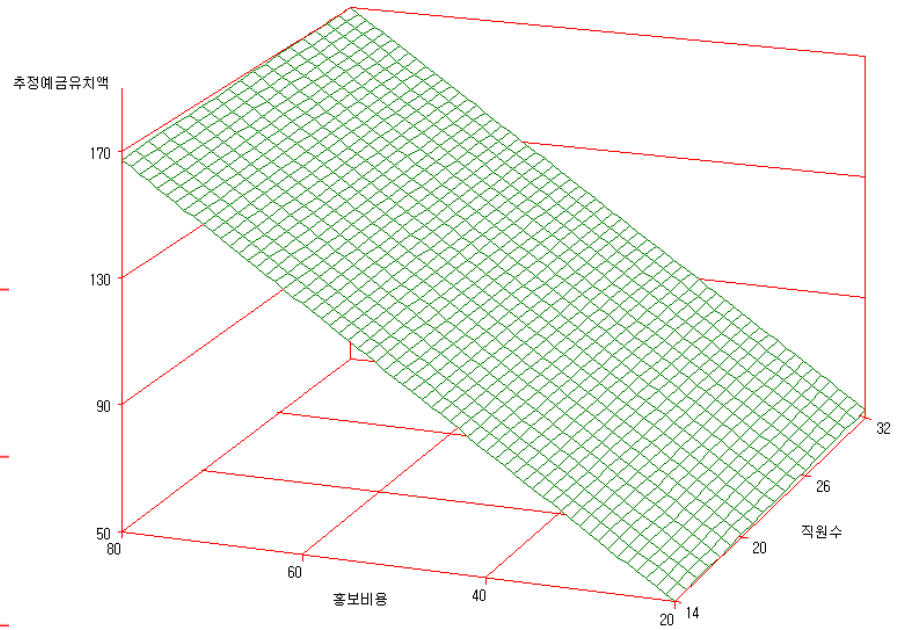
- 종속변수를 설명하기 위해서 두 개 이상의 독립변수가 사용되는 선형회귀모형을 다중선형 회귀모형이라 함



01 다중회귀 회귀(Multiple Linear Regression)

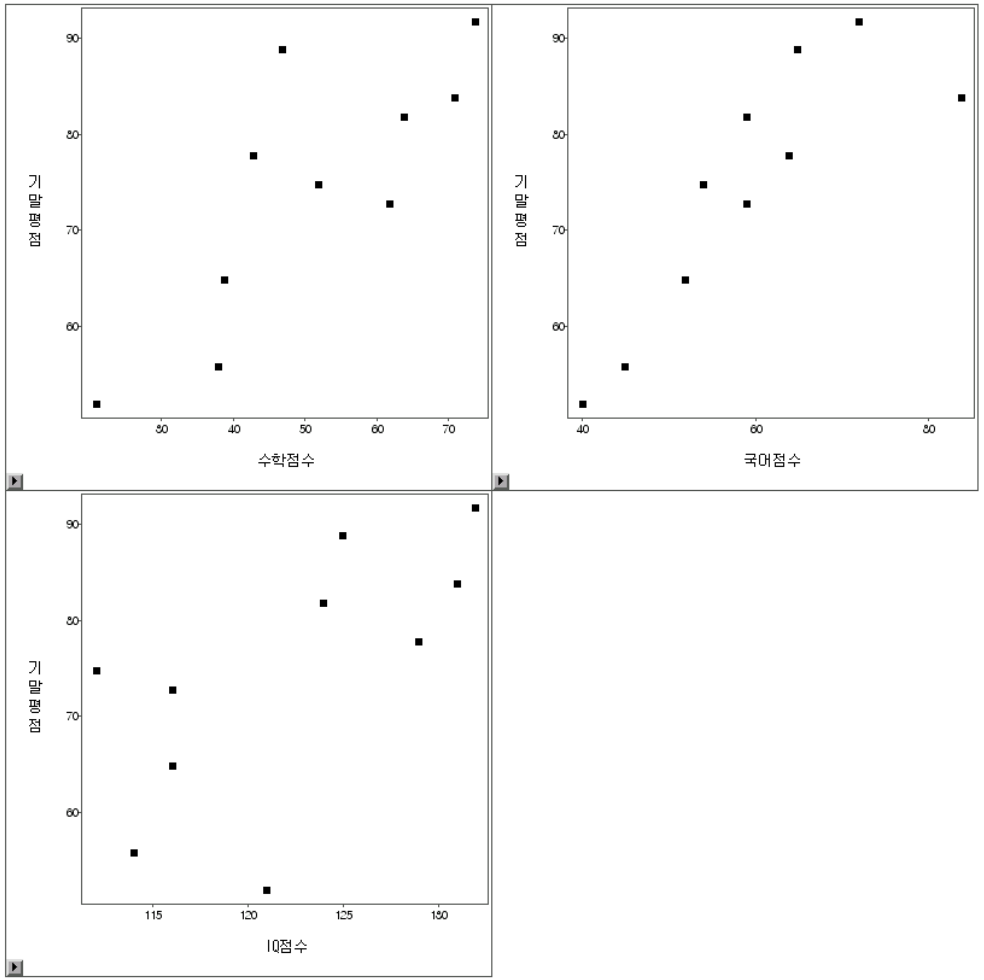
- **초평면에 대해 설명해야 함**



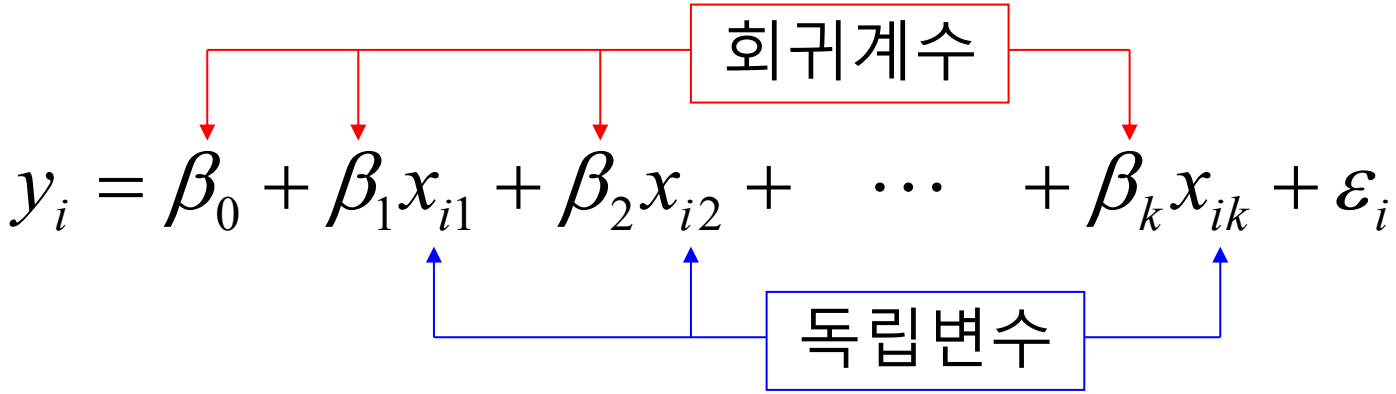


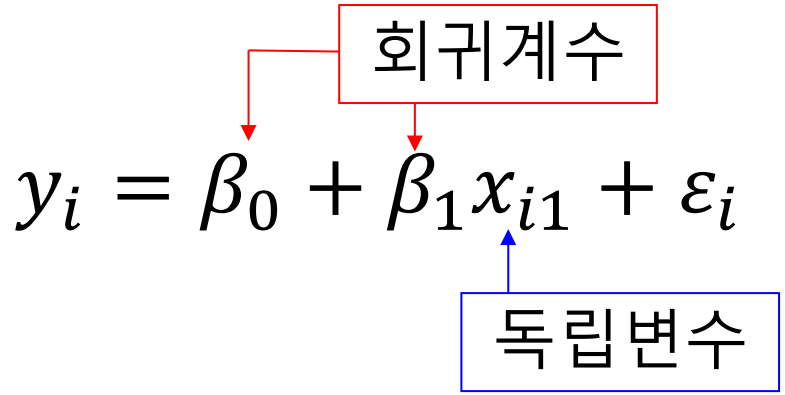
01 다중회귀 회귀(Multiple Linear Regression)

- 수학, 국어, 그리고 IQ가 기말성적에 어떤 영향을 미치는가?



다중회귀선형모형


회귀계수
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$
독립변수


회귀계수
$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$
독립변수

01 다중회귀 회귀(Multiple Linear Regression)

<단순선형회귀모형>

모형 : $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$

가정 : $\varepsilon_i \sim iidN(0, \quad \sigma^2)$

y = 예금유치액

x_{i1}=홍보비용

회귀계수에 대한 정보 ?

<다중 선형회귀모형>

모형: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

가정: $\varepsilon_i \sim iidN(0, \quad \sigma^2)$

y = 예금유치액

x_{i1}=홍보비용, x_{i2}=직원수

01 다중회귀 회귀(Multiple Linear Regression)

<단순선형회귀모형>

모형 : $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$

가정 : $\varepsilon_i \sim iidN(0, \quad \sigma^2)$

<다중 선형회귀모형>

모형: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

가정: $\varepsilon_i \sim iidN(0, \quad \sigma^2)$

최소제곱법(OLS)은 회귀모형의

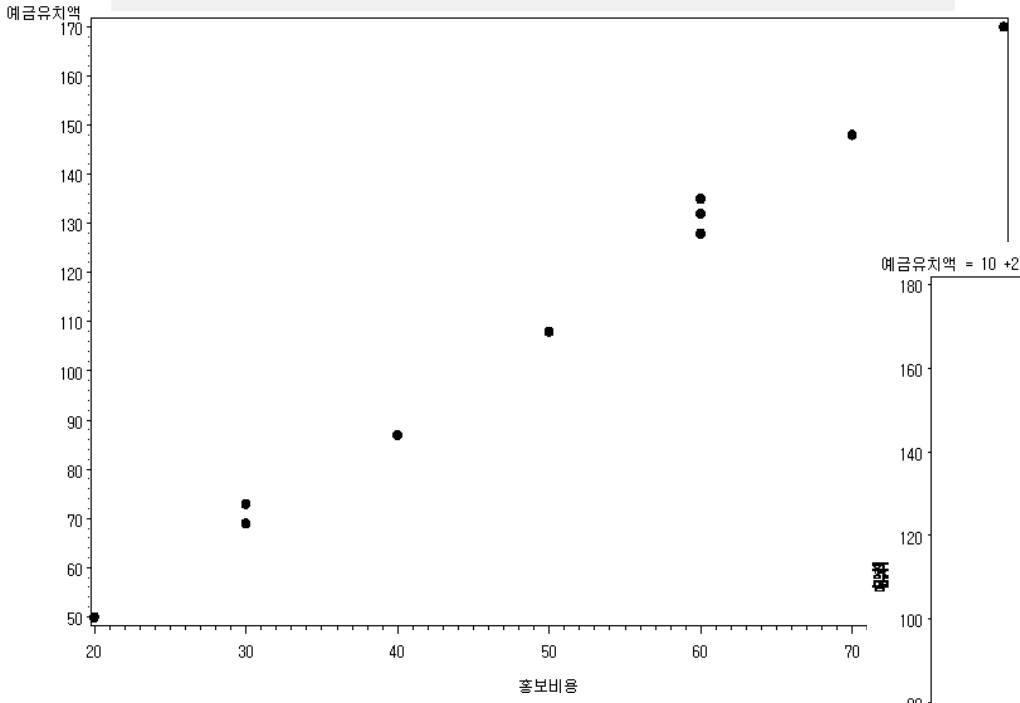
오차 제곱의 합 $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ 을 최소로

하는 회귀계수를 이들의 추정치로 하는 것

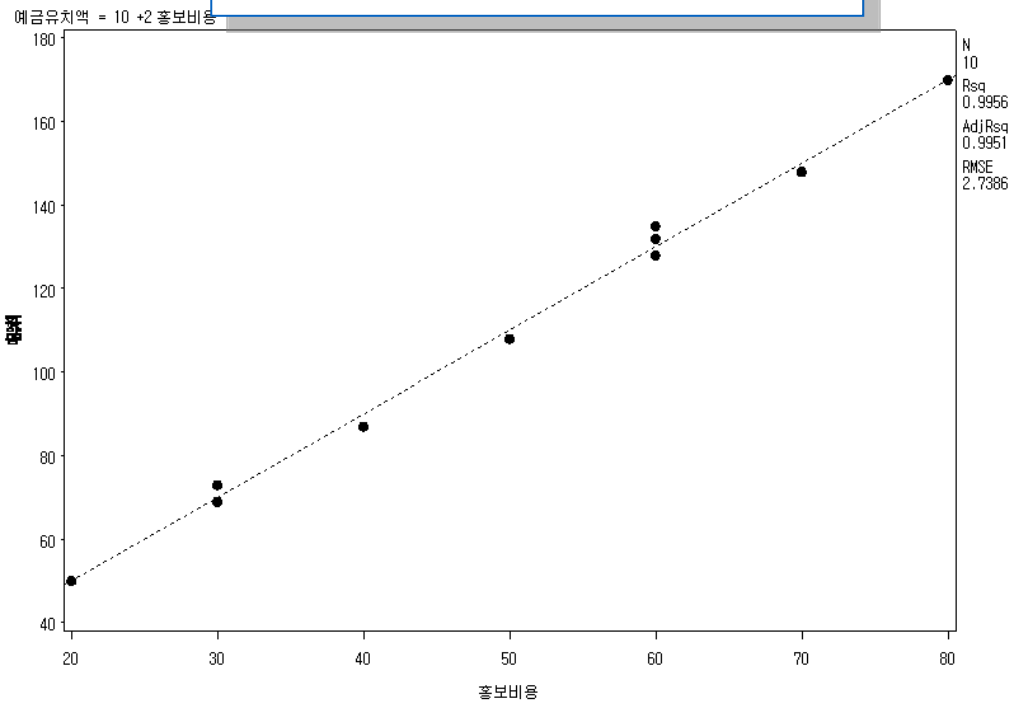
01 다중회귀 회귀(Multiple Linear Regression)

홍보비용과 예금유치액에 대한 산점도

$$y = \beta_0 + \beta_1 x_1 + \varepsilon, \quad y = 예금유치액, x_1 = 홍보비용$$



$$\hat{y} = 10 + 2 \times 홍보비용$$



01 다중회귀 회귀(Multiple Linear Regression)

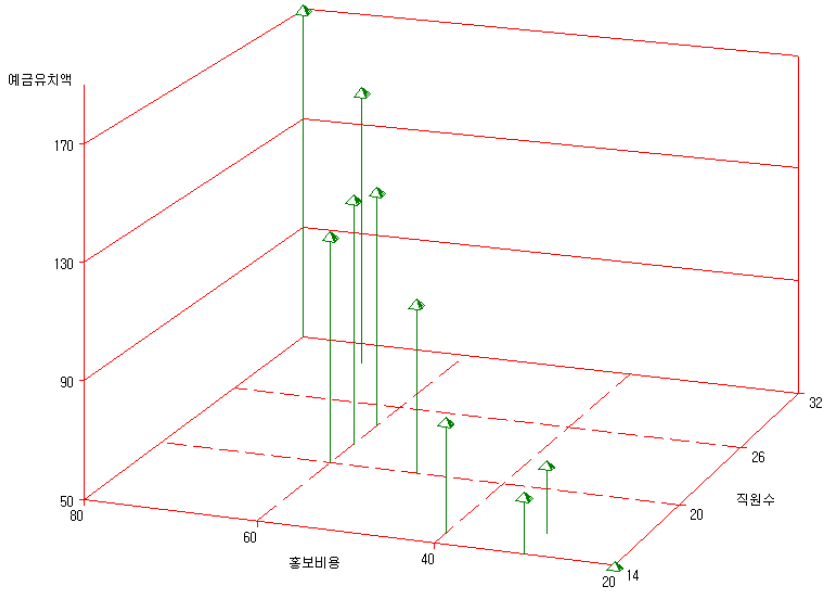
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

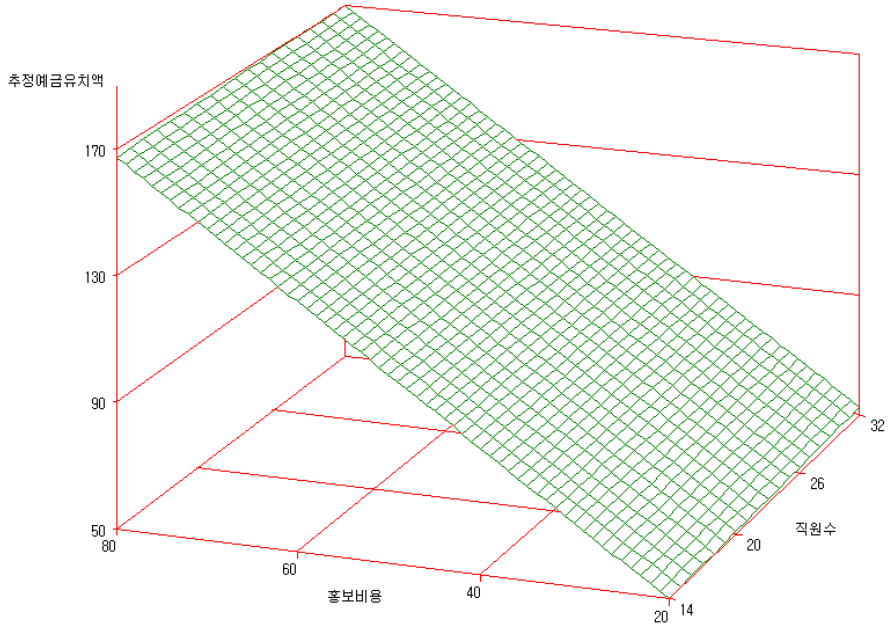
y = 예금유치액, x_1=홍보비용, x_2=직원수

홍보비용, 직원수 그리고 예금유치액에 대한 산점도

추정된 다중회귀모형

$$\hat{y} = 9.23 + 1.95 \times \text{홍보비용} + 0.15 \times \text{직원수}$$





01 다중회귀 회귀(Multiple Linear Regression)

다중회귀 코드

```
install.packages("rstatix") #통계 테스트
install.packages("skimr") #데이터의 흐름 요약

library(rstatix)
library(skimr)

#데이터 불러오기
df <- read.csv("diabetes.csv", header = TRUE, na = ".", stringsAsFactors = TRUE)
skim(df)

#불필요 변수 제거
df <- select(df, -c(var1)) #하나면 제거(순수한 열 이름만 적기)
df <- select(df, -c(var1, var2, var3)) #여러 개 제거
skim(df)
```

01 다중회귀 회귀(Multiple Linear Regression)

다중회귀 코드

```
#다중회귀분석
result <- lm(df$Diabetes ~ ., data = df)
result1 <- lm(df$Diabetes ~ 1, data = df) # 다른 변수 고려 X(단순 평균으로 예측)

#ANOVA분석을 활용해 두 집단의 의미있는 차이가 있는지 확인(잔차제곱합 차이의 검증)
anova_result <- anova(result, result1)
print(anova_result)

#회귀모델
summary(result)

#모델 저장
saveRDS(model, "regression_model.rds")

#모델 불러오기
loaded_model <- readRDS("regression_model.rds")
```

01 다중회귀 회귀(Multiple Linear Regression)

다중회귀 코드

```
#데이터를 csv에서 불러오기
new_data <- read.csv("diabetes_test.csv")

#모델에 새로운 데이터 추가
predicted<- predict(loaded_model, newdata = new_data)
head(predicted)

#예측 결과 출력
results <- data.frame(new_data, Predicted = predicted)

#예측 결과를 새로운 csv 파일로 저장 (옵션)
write.csv(results, "predicted_values.csv", row.names = FALSE)
```

01 다중회귀 회귀(Multiple Linear Regression)

다중회귀 코드

```
#RMSE 계산

rmse <- sqrt(mean((results$Diabetes - results$Predicted)^2))

print(paste("RMSE:", round(rmse, 3)))
```

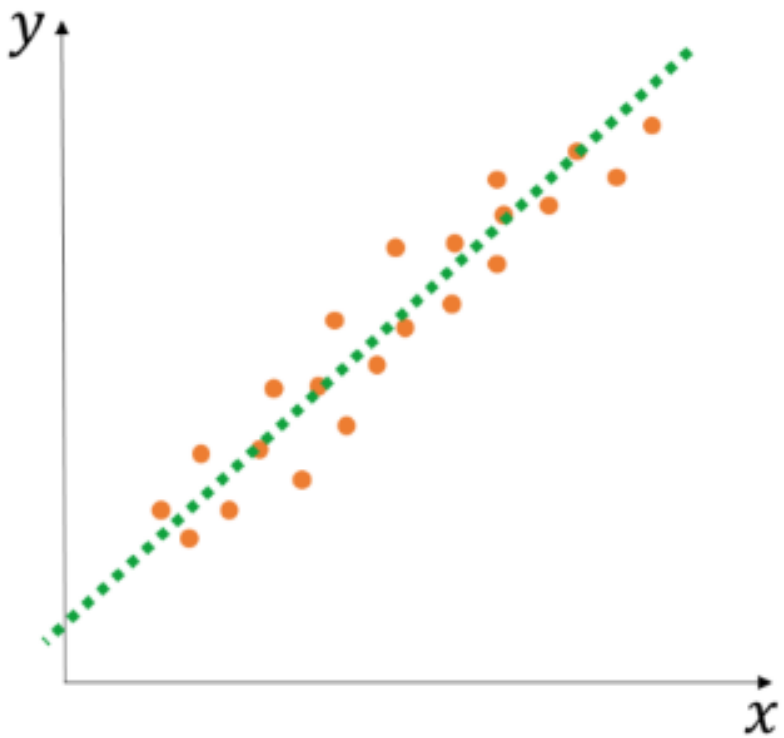
01 다중회귀 회귀(Multiple Linear Regression)

모형의 타당성&신뢰성 검정

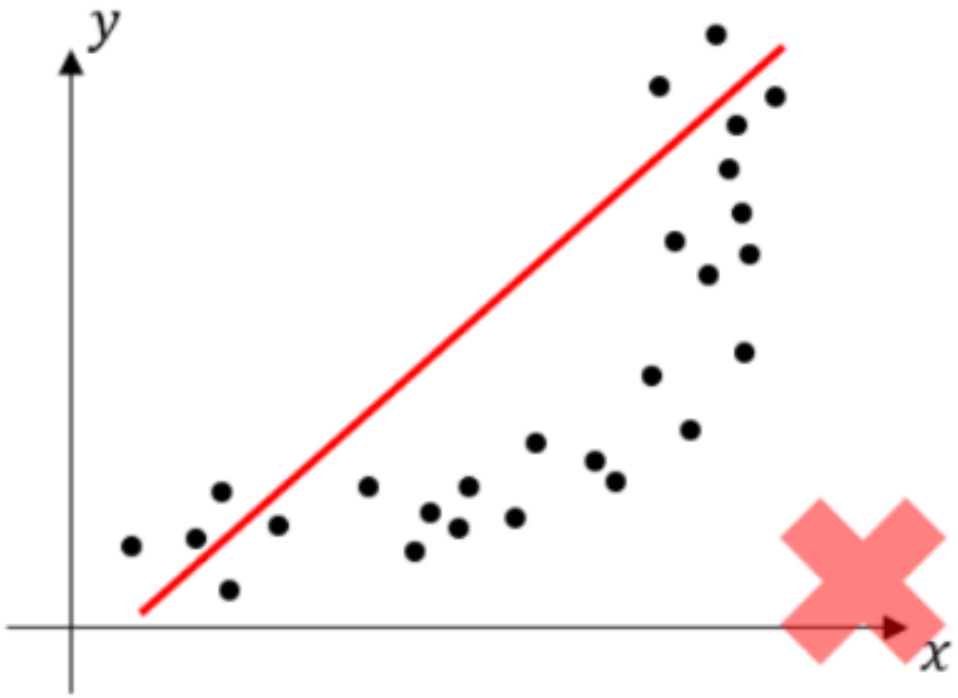
- 1) 선형성 → 예측하고자 하는 종속변수 y와 독립변수 x간의 선형성을 이루어야 함
- 2) 등분산성 → 오차의 분산이 같다는 것을 의미하고 특정한 패턴 없이 고르게 분포하는 것을 의미함
- 3) 독립성 → 오차 사이에는 서로 영향을 주지 않으며 오차간의 상관 없이 독립적이어야 함
- 4) 정규성 → 잔차가 정규성을 만족하는지 여부 (오차가 평균이 0인 정규분포)
- 5) 다중공선성 → 회귀 모델에서 두 개 이상의 독립변수가 서로 높은 상관관계가 있는 상황

01 다중회귀 회귀(Multiple Linear Regression)

- 1) 선형성 → 예측하고자 하는 종속변수 y와 독립변수 x간의 선형성을 이루어야 함



선형성 만족 O



선형석 만족 X

01 다중회귀 회귀(Multiple Linear Regression)

- **1) 선형성 → 예측하고자 하는 종속변수 y와 독립변수 x간의 선형성을 이루어야 함**

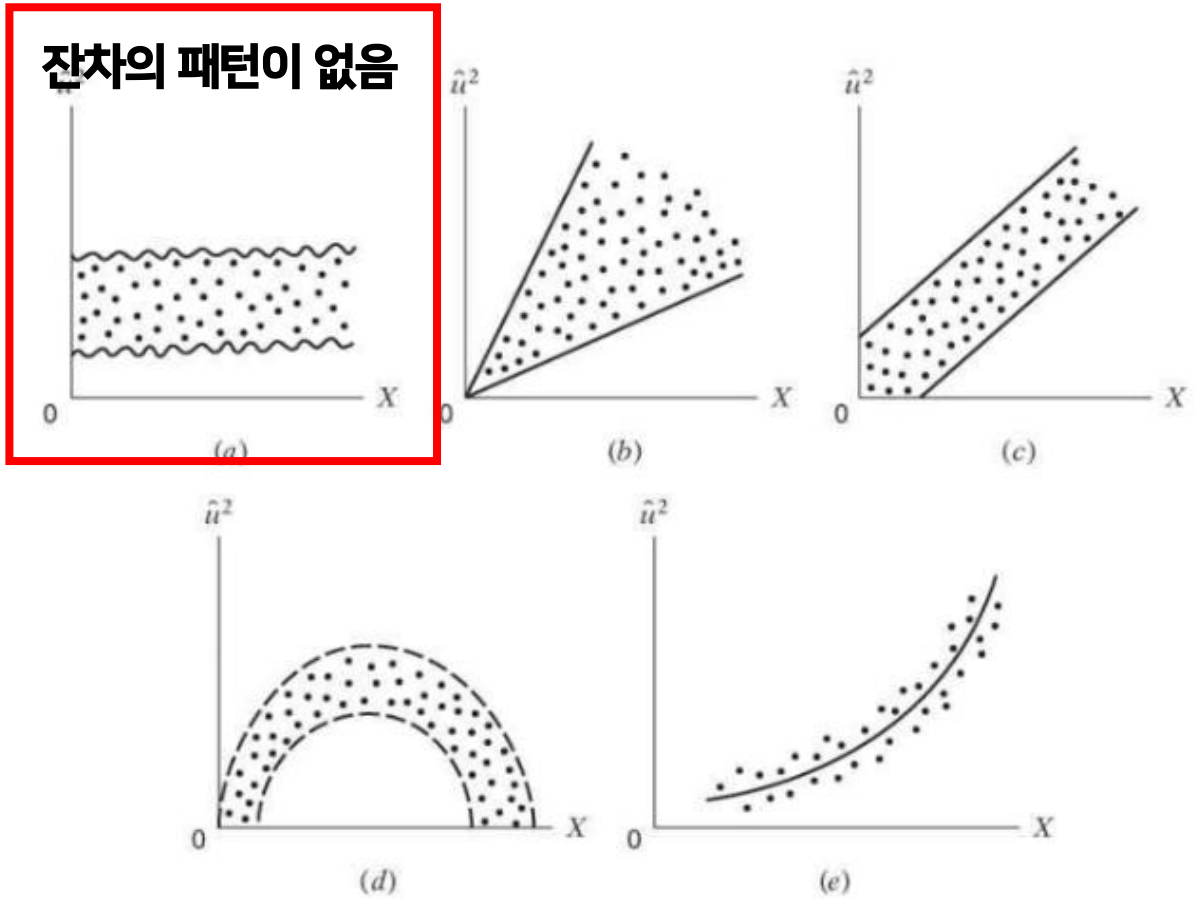
```
df$Fitted    <- fitted(result)       # 적합값
df$Residuals <- resid(result)        # 잔차

library(ggplot2)

ggplot(df, aes(x = Fitted, y = Residuals)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residuals vs Fitted (선형성 확인)",
       x = "적합값 (Fitted Values)",
       y = "잔차 (Residuals)") +
  theme_minimal()
```

01 다중회귀 회귀(Multiple Linear Regression)

- 2) 등분산성 → 오차의 분산이 같다는 것을 의미하고 특정한 패턴 없이 고르게 분포하는 것을 의미함



- 다트 게임에서 항상 같은 거리에서 던지면, 명중점들이 과녁 주위에 비슷한 폭으로 흩어짐 = 등분산
- 수산물 시장에서 항목별로 무게가 달라진다. (무거울수록 오차가 점점 커지면 이분산)

- **2) 등분산성 → 오차의 분산이 같다는 것을 의미하고 특정한 패턴 없이 고르게 분포하는 것을 의미함**

```
df$StdResid <- sqrt(abs(df$Residuals))  # 표준화된 잔차의 크기 근사치

ggplot(df, aes(x = Fitted, y = StdResid)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "loess", se = FALSE, color = "blue") +
  labs(title = "Scale-Location Plot (등분산성 확인)",
       x = "적합값 (Fitted Values)",
       y = "Standardized Residuals") +
  theme_minimal()

#등분산성 : p>0.05보다 크면, 등분산성 가정을 유지 ⟷ 반대는 이분산성

library(lmtest)
bptest(result)
```

01 다중회귀 회귀(Multiple Linear Regression)

- 3) 독립성 → 오차 사이에는 서로 영향을 주지 않으며 오차간의 상관 없이 독립적이어야 함
  - 버스가 지연되면 다음 정거장 도착도 연쇄적으로 지연됨 → 잔차(+지연)가 줄줄이 이어짐 = 독립성 위반
  - 현실 예(집단): 같은 반 학생들의 시험 채점에서, 채점량이 많아 피곤해 이후의 시험에 점수를 후하게/박하게 = 독립성 위반

- Durbin Watson 검정
  - → 검정통계량 D-W Statistic값이 0~4의 값을 가지며 0으로 가까울 수록(잔차의) 양의 상관관계 4에 가까울수록 음의 상관관계
  - → 1.5~2.5사이 일때 잔차는 독립
  - → 0 근사 : 양의 자기상관
  - → 4 근사 : 음의 자기상관

```
#독립성 : p>0.05보다 크면 잔차의 독립성을 가짐

library(car)
durbinWatsonTest(result)
```