



9주차: 모델링과 예측 : 선형회귀

ChulSoo Park

School of Computer Engineering & Information Technology
Korea National University of Transportation

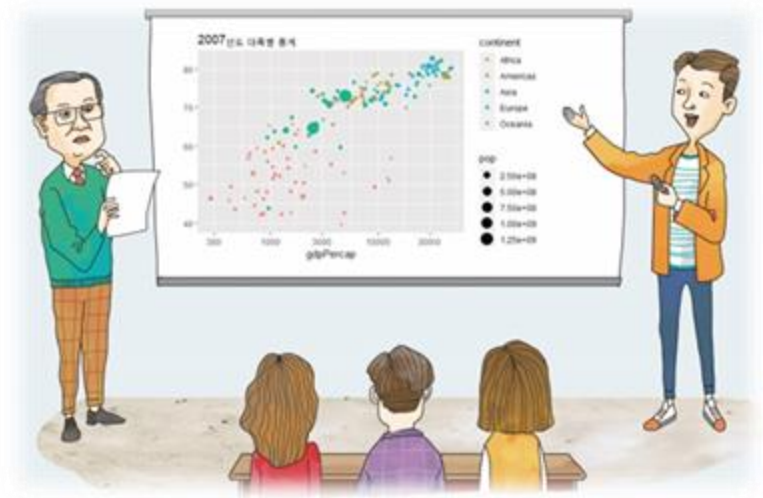
학습목표 (9주차)

- ❖ 모델링과 예측 이해
- ❖ 회귀 분석 개념 이해
- ❖ 단순회귀분석, 분산 분석(ANOVA) 실행
- ❖ 모델의 통계량 이해
- ❖ t-검정과 분산분석

07

CHAPTER

모델링과 예측



CONTENTS

7.1 모델링과 예측이란?

7.2 현실 세계의 모델링

7.3 단순 선형 회귀

7.4 단순 선형 회귀 적용:cars 데이터

7.5 모델의 통계량 해석

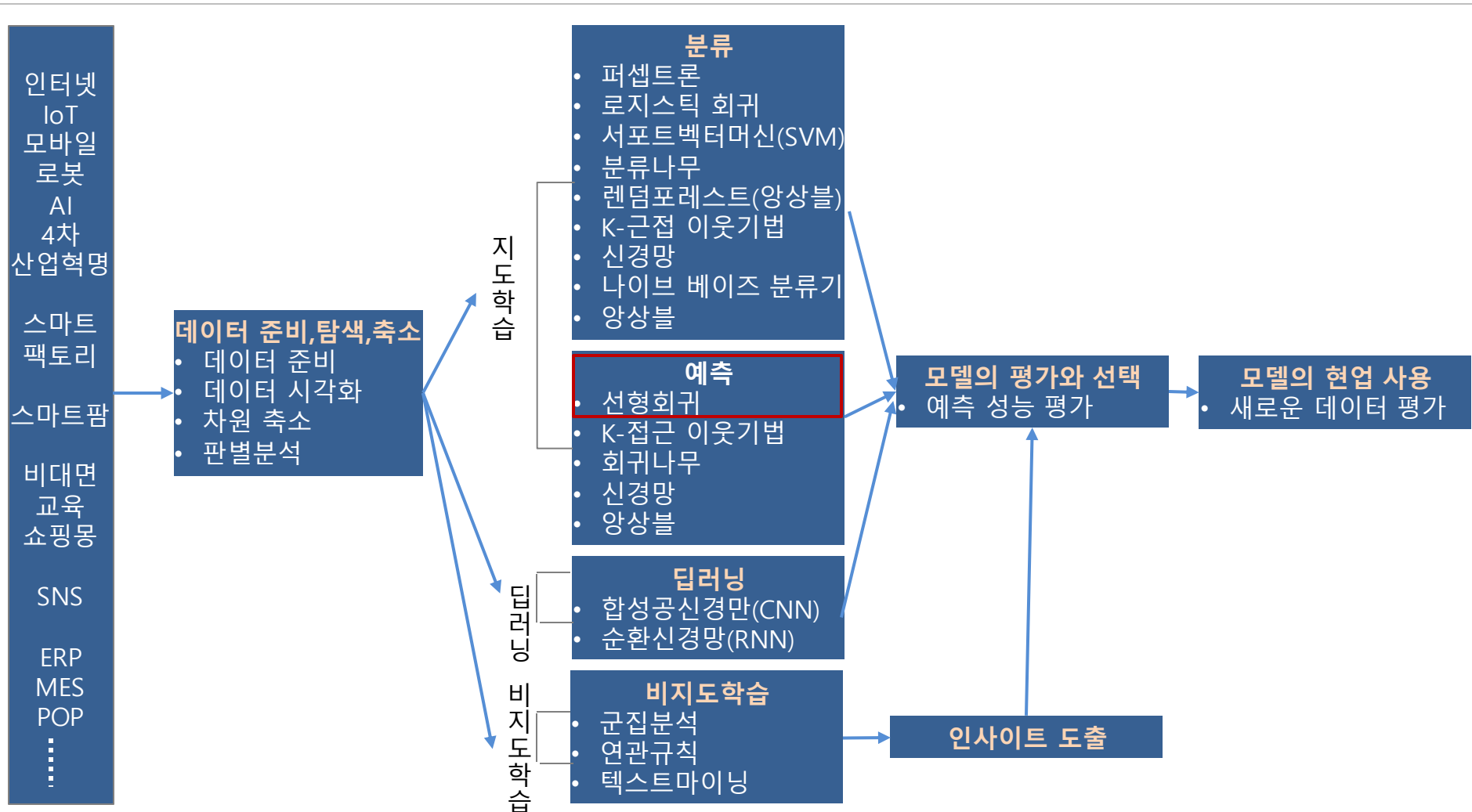
7.6 다중 선형 회귀

7.7 다중 선형 회귀의 적용:trees 데이터

t-검정과 분산 분석

요약

■ 데이터 분석 Process에서 이번주 교육 위치

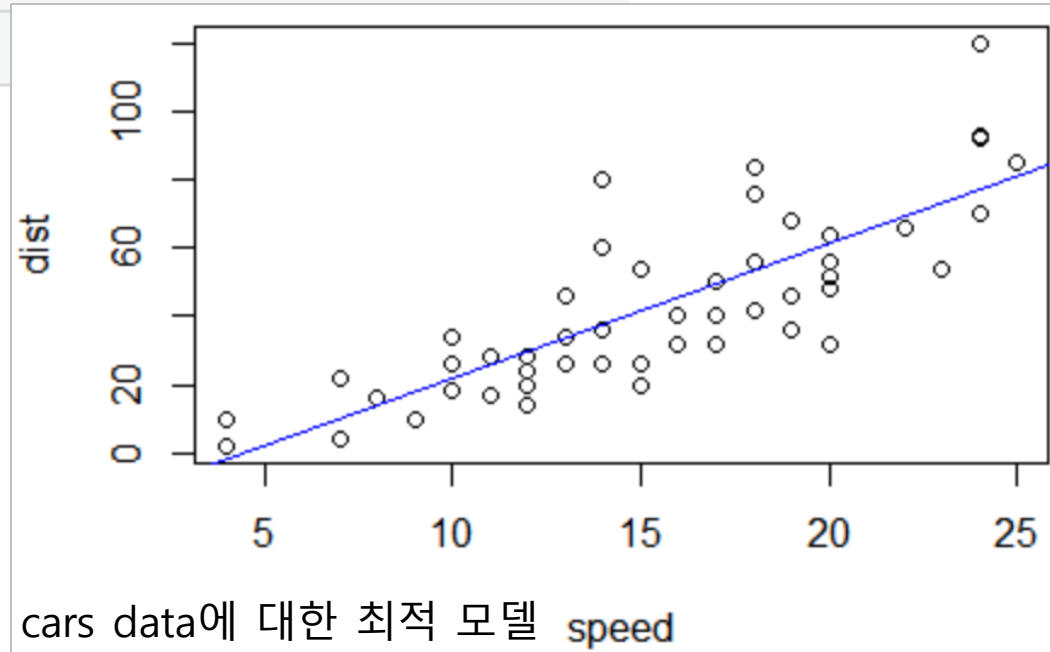


7.4 단순 선형 회귀의 적용 : cars data

■ 모델 적합

Console C:/RSources/ ➡

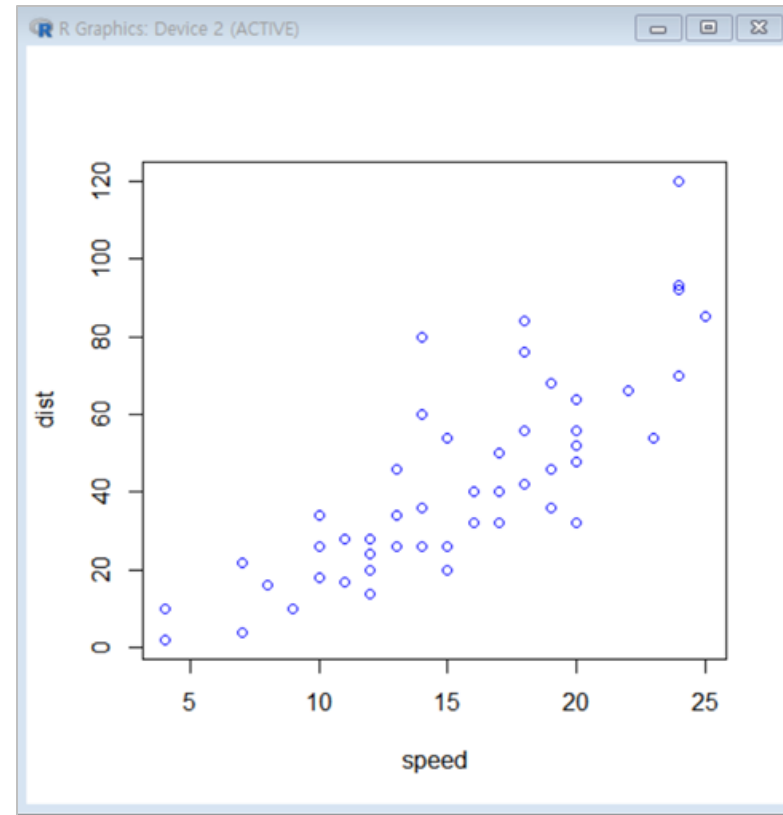
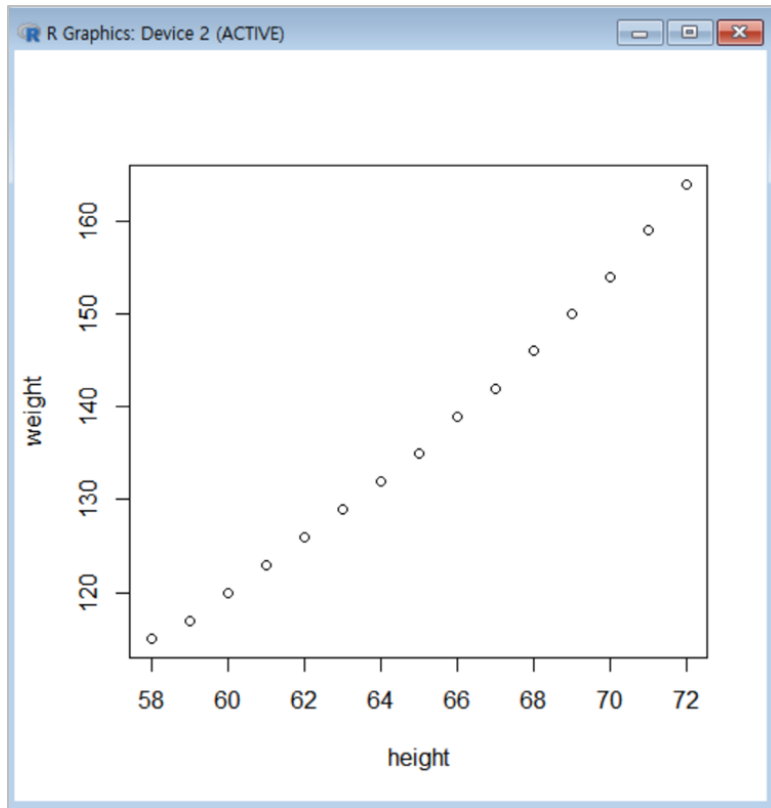
```
> car_model = lm(dist~speed,data=cars)
> coef(car_model)
(Intercept)      speed
-17.579095      3.932409
> abline(car_model,col='blue')
```



최적 모델 : $\text{dist} = -17.579095 + 3.932409 \times \text{speed}$

7.5 모델의 통계량 해석

women, cars 데이터의 시각화 정보



7.5 모델의 통계량 해석

- women 데이터로 모델의 통계량 해석 해보기
 - 먼저 women 데이터의 확인

Console C:/RSources/ ↗

```
> str(women)
'data.frame':  15 obs. of  2 variables:
 $ height: num  58 59 60 61 62 63 64 65 66 67 ...
 $ weight: num  115 117 120 123 126 129 132 135 139 142 ...

> women
  height weight
1     58    115
2     59    117
3     60    120
4     61    123
5     62    126
6     63    129
7     64    132
8     65    135
9     66    139
10    67    142
11    68    146
12    69    150
13    70    154
14    71    159
15    72    164
```

7.5 모델의 통계량 해석

■ women 데이터로 모델의 통계량 해석 해보기

- 설명 변수 : height(키), 반응 변수 : weight(몸무게)
- 모델링한 다음 모델을 가시화 하면

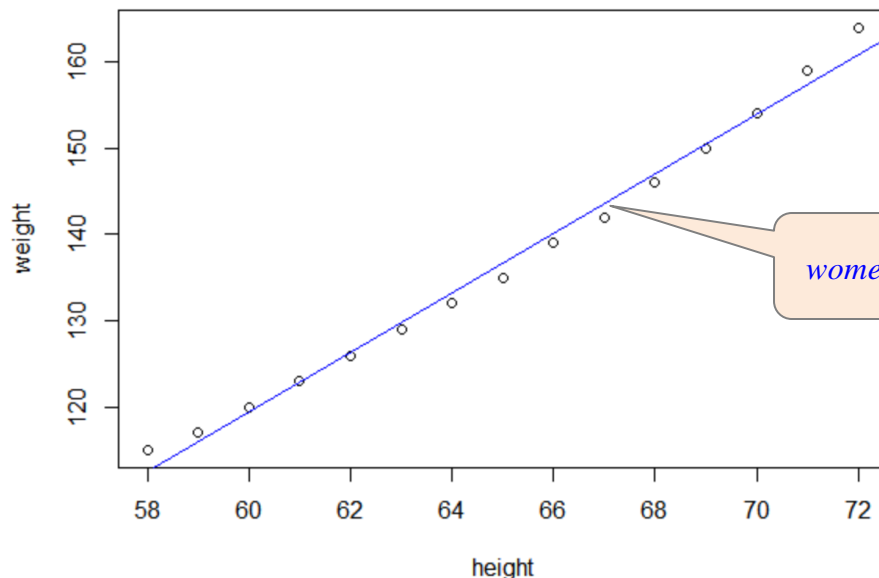
```

Console C:/RSources/
> women_model = lm(weight ~ height, data=women)
> coef(women_model)
(Intercept)      height
   -87.51667     3.45000
> plot(women)
> abline(women_model, col='blue')
  
```

반응 변수

설명 변수

최적 모델 : $\text{weight} = -87.51667 + 3.45000 \times \text{height}$



women data의 선형 회귀 모델 그래프

7.5 모델의 통계량 해석

■ women 데이터로 모델의 통계량 해석 해보기

- summary 함수로 모델의 상세 내용을 살펴보면,

```

Console C:/Rsources/
> summary(women_model)

Call:
lm(formula = weight ~ height, data = women)

Residuals:
    Min       1Q   Median       3Q      Max
-1.7333 -1.1333 -0.3833  0.7417  3.1167

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height       3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.525 on 13 degrees of freedom
Multiple R-squared:  0.991,    Adjusted R-squared:  0.9903
F-statistic: 1433 on 1 and 13 DF, p-value: 1.091e-14

```

- height 변수의 계수의 p-값이 0.05보다 작은 1.09e-14이므로 통계적으로 유의미한 모델링이 되었음을 확인

7.5 모델의 통계량 해석

■ cars 데이터의 모델과 비교해 보면

```

Console C:/RSources/ ↗
> summary(car_model)

call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min       1Q   Median       3Q      Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.5791     6.7584  -2.601   0.0123 *
speed           3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

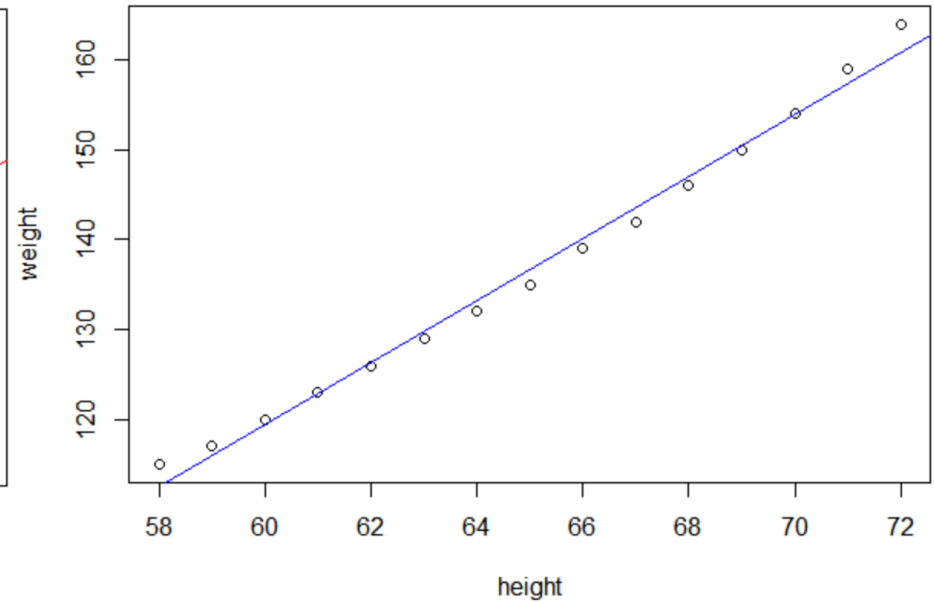
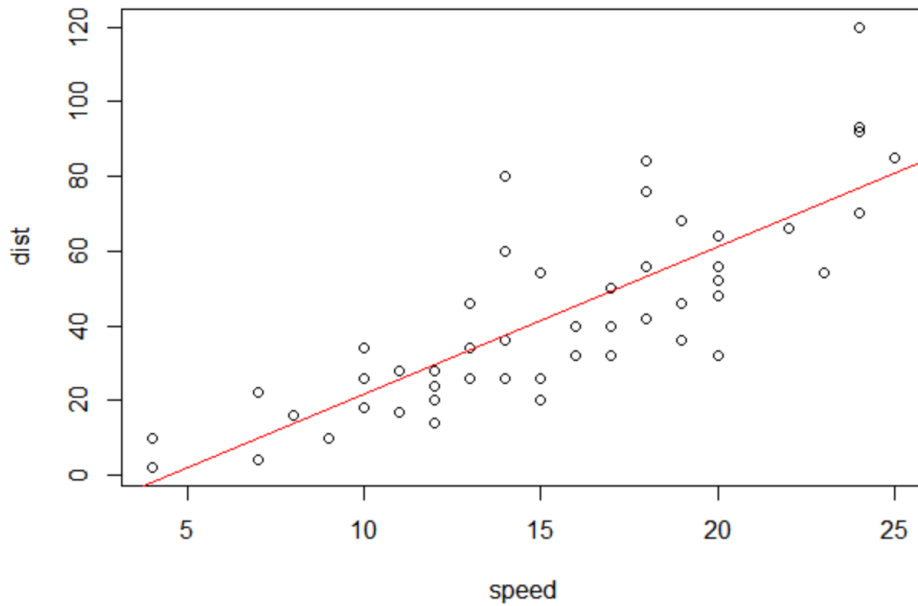
Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

```

- (Intercept), 즉 절편의 p-값은 0.0123으로 높은 편임
- women 데이터는 절편의 p-값이 $1.71e-9$ 으로 매우 작음 ← women 데이터의 모델이 cars 모델보다 예측을 더 잘 할 것으로 기대됨

7.5 모델의 통계량 해석

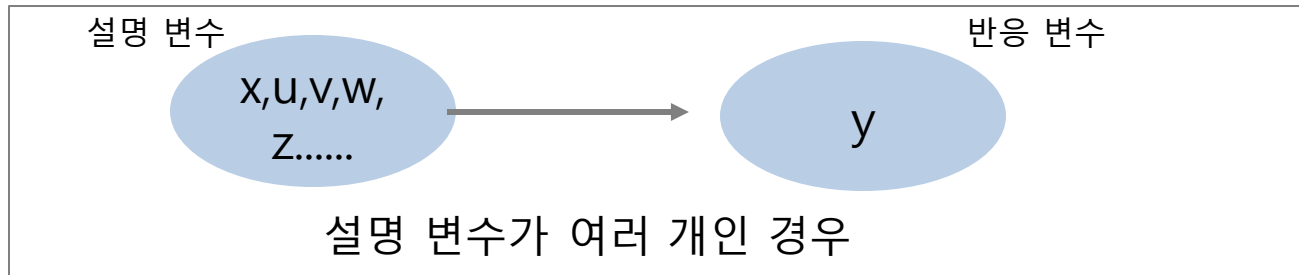
- 모델링 결과 그림에서 보는 바와 같이 점들이 상하로 퍼져있어 절편에 대한 오차가 커서 나타나는 현상



7.6 다중 선형 회귀

■ 현실 세계의 데이터는 설명 변수가 여러 개

- 월급에 영향을 미치는 변수로 판매 대수뿐 아니라 근무 연수, 직급 등
- 제동 거리에 영향을 미치는 변수로 속도뿐 아니라 날씨나 브레이크의 종류 등
- 매출에 영향을 미치는 변수 브랜드, 광고, 날씨, 경쟁사 프로모션, 영업 능력, 경영진 의지 등
- 일반적으로 표시하면,



■ 다중 선형 회귀(multiple linear regression)

- 설명 변수가 2개 이상인 선형 회귀
- 설명 변수가 2개인 경우에는 매개변수가 3개
- $y = a_1x + a_2u + a_0$
- 일반적으로 설명 변수가 k 개이면 매개변수는 $k+1$ 개

7.6 다중 선형 회귀

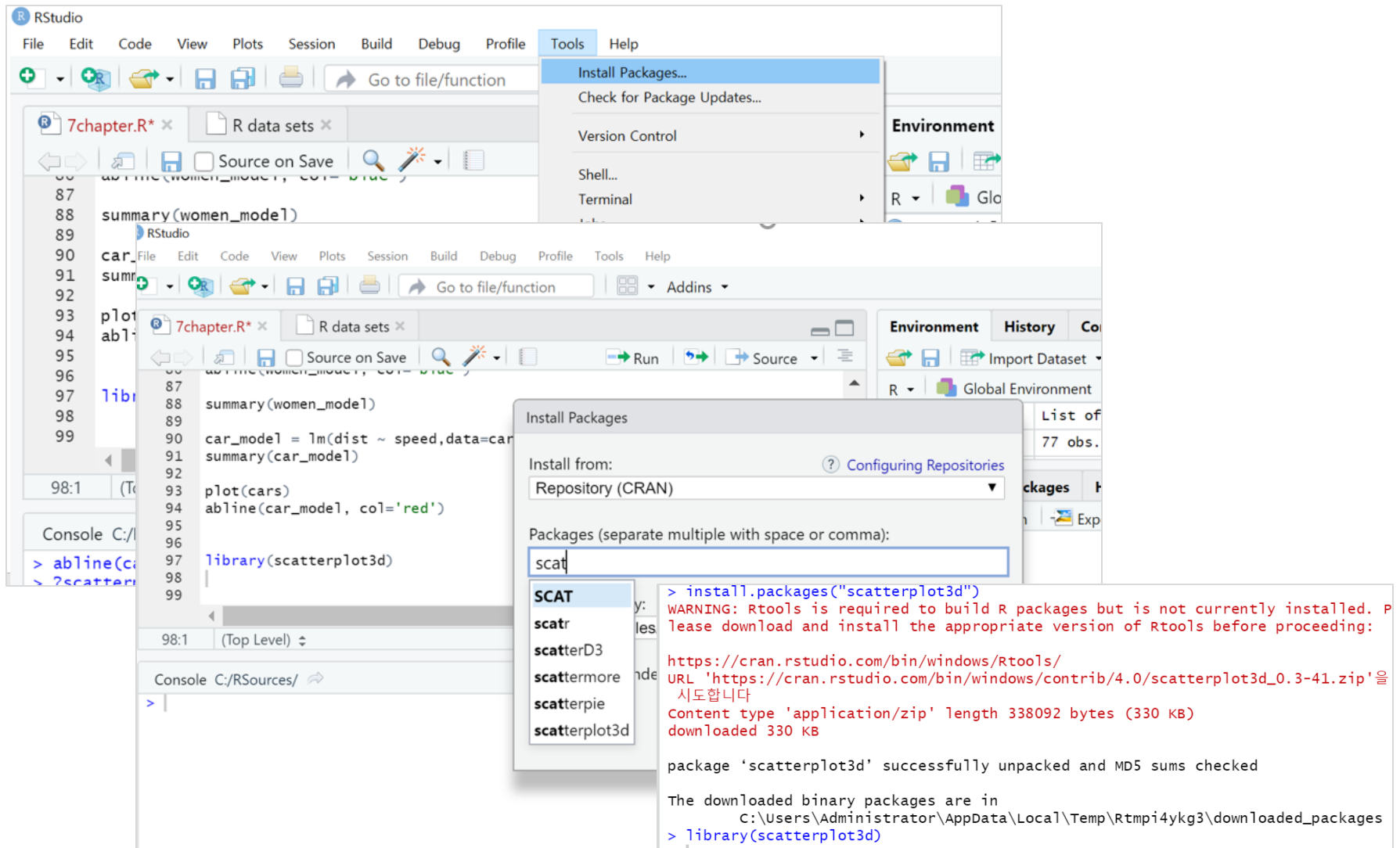
■ 사례 “영희의 물리 실험”

- 전기량과 물체의 무게에 따른 이동 거리를 측정하는 실험
- 전기량을 x , 무게를 u , 이동 거리를 y 로 표기하고 실험 데이터를 수집
- 설명 변수 : 전기량, 물체의 무게
- 반응 변수 : 이동 거리
- $X = \{3.0, 6.0, 3.0, 6.0\}$
- $U = \{10.0, 10.0, 20.0, 20.0\}$
- $Y = \{4.65, 5.9, 6.7, 8.02\}$



7.6 다중 선형 회귀

■ Scatterplot3d package Install



The screenshot shows the RStudio interface with the 'Tools' menu open and 'Install Packages...' selected. The 'Install Packages' dialog box is open, showing the 'Repository (CRAN)' dropdown and the 'Packages' field containing 'scatterplot3d'. The console shows the command `install.packages("scatterplot3d")` and the output indicating successful installation.

```
> install.packages("scatterplot3d")
WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:
https://cran.rstudio.com/bin/windows/Rtools/
URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/scatterplot3d_0.3-41.zip'을 시도합니다
Content type 'application/zip' length 338092 bytes (330 KB)
downloaded 330 KB

package 'scatterplot3d' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\Administrator\AppData\Local\Temp\Rtmpi4ykg3\downloaded_packages
> library(scatterplot3d)
```

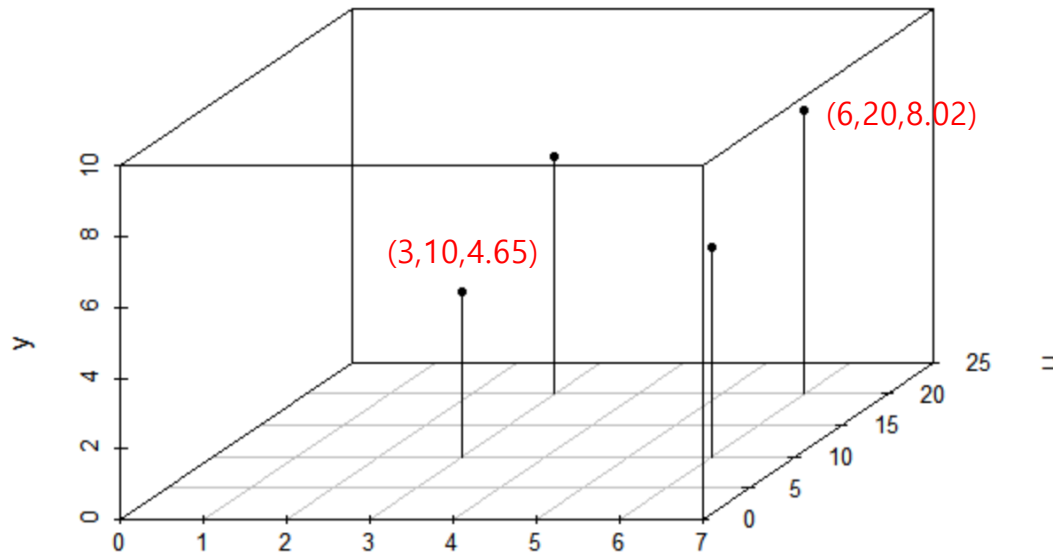
7.6 다중 선형 회귀

■ 사례 “영희의 물리 실험”

■ 가시화 하면

Console C:/RSources/ ↗

```
> x=c(3.0, 6.0, 3.0, 6.0)
> u=c(10.0, 10.0, 20.0, 20.0)
> y=c(4.65, 5.9, 6.7, 8.02)
> scatterplot3d(x,u,y, xlim=2:7, ylim=7:23, zlim=0:10, pch=20, type='h')
```



x Scatterplot3d 함수를 이용한 데이터 시각화

7.6 다중 선형 회귀

■ 사례 “영희의 물리 실험”

- 다중 선형 회귀 적용
- 단순 선형 회귀와 같이 lm함수 사용
- 설명 변수 : $x + u$ 적용
- 반응 변수 : y

```
Console C:/Rsources/
> m=lm(y~x+u)
> coef(m)
(Intercept)          x          u
  1.2625000    0.4283333    0.2085000
```

반응 변수(y)

설명 변수 (x + u)

- 최적 모델 : $y = 1.2625 + 0.428333x + 0.2085 u$

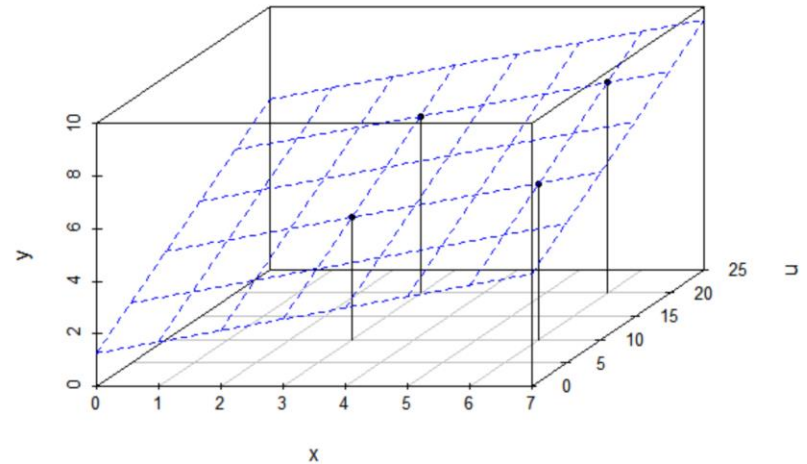
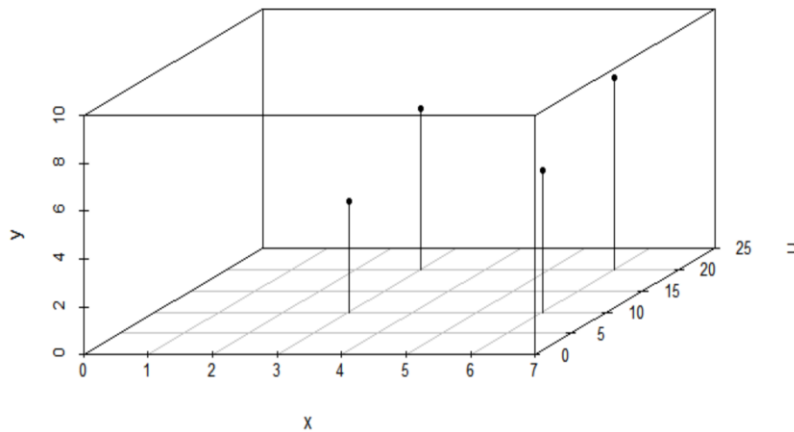
7.6 다중 선형 회귀

■ 사례 “영희의 물리 실험”

■ 최적 모델 가시화

Console C:/Rsources/ ↗

```
> s=scatterplot3d(x,u,y, xlim=0:7, ylim=0:25, zlim=0:10, pch=20, type='h')  
> s$plane3d(m, col='blue')
```



7.6 다중 선형 회귀

■ 사례 “영희의 물리 실험”

■ 오차를 분석

```

Console C:/Rsources/ ↗
> fitted(m)
      1      2      3      4
4.6325 5.9175 6.7175 8.0025
> residuals(m)                # 잔차(오차)
      1      2      3      4
0.0175 -0.0175 -0.0175 0.0175
> deviance(m)                  # 잔차 제곱합
[1] 0.001225
> deviance(m)/length(x)        # 평균 제곱 오차
[1] 0.00030625
  
```

(xi, ui)	(3.0, 10.0)	(6.0, 10.0)	(3.0, 20.0)	(6.0, 20.0)
예측값 $f(x_i, u_i)$	4.6325	5.9175	6.7175	8.0025
그라운드 투루스(y_i)	4.65	5.9	6.7	8.02
오차	0.0175	-0.0175	-0.0175	0.0175

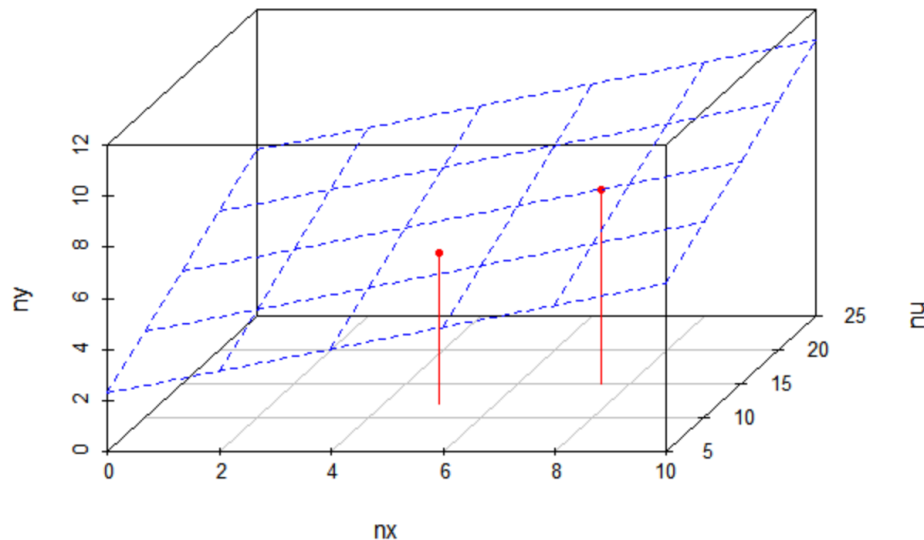
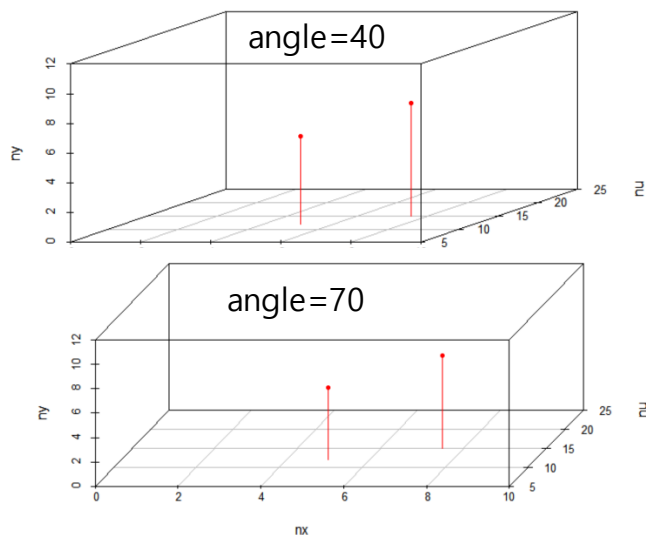
7.6 다중 선형 회귀

■ 사례 “영희의 물리 실험”

- 최적 모델을 이용하여 새로운 데이터 (7.5,15.0)과 (5.0,12.0)에 대한 예측

Console C:/RSources/ ↗

```
> nx=c(7.5, 5.0)
> nu=c(15.0, 12.0)
> new_data=data.frame(x=nx, u=nu)
> ny=predict(m, new_data)
> ny
      1      2
7.602500 5.906167
> s=scatterplot3d(nx,nu,ny, xlim=0:10, ylim=7:25, zlim=0:12, pch=20, type='h',color
='red',angle=60)
> s$plane3d(m, col='blue')
```



7.7 다중 선형 회귀의 적용(trees data)

■ trees 데이터 확인하기

```
Console C:/RSources/
> str(trees)
'data.frame': 31 obs. of 3 variables:
 $ Girth : num 8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
 $ Height: num 70 65 63 72 81 83 66 75 80 75 ...
 $ Volume: num 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
> head(trees,5)
  Girth Height volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
> summary(trees)
      Girth      Height      Volume
Min.   : 8.30   Min.   :63   Min.   :10.20
1st Qu.:11.05   1st Qu.:72   1st Qu.:19.40
Median :12.90   Median :76   Median :24.20
Mean   :13.25   Mean   :76   Mean   :30.17
3rd Qu.:15.25   3rd Qu.:80   3rd Qu.:37.30
Max.   :20.60   Max.   :87   Max.   :77.00
```

7.7 다중 선형 회귀의 적용(trees data)

R: Diameter, Height and Volume for Black Cherry Trees ▾

Find in Topic

trees {datasets}

R Documentation

Diameter, Height and Volume for Black Cherry Trees

Description

This data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. Note that the diameter (in inches) is erroneously labelled Girth in the data. It is measured at 4 ft 6 in above the ground.

Usage

trees

Format

A data frame with 31 observations on 3 variables.

```
> head(trees, 5)
```

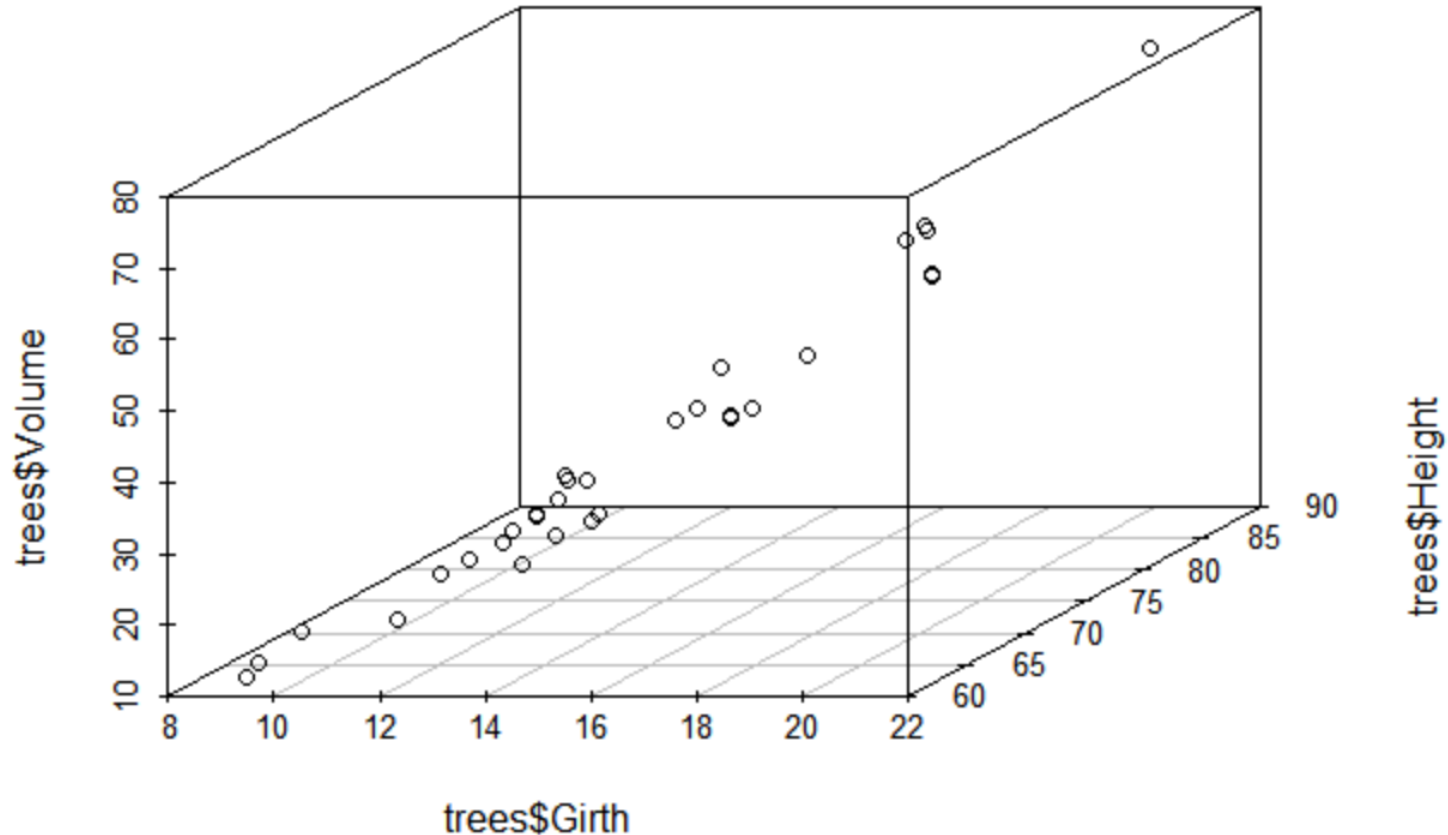
	Girth	Height	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8

[1] Girth numeric Tree diameter (rather than girth, actually) in inches

7.7 다중 선형 회귀의 적용(trees data)

■ Trees data를 가시화 하면



- `scatterplot3d(trees$Girth, trees$Height, trees$Volume)`



7.7 다중 선형 회귀의 적용(trees data)

■ 다중 선형 회귀 적용하기

- 어떤 변수를 반응 변수로 하나? → 목재상이 알고자 하는 것은 나무의 상태에 따른 목재의 부피일 것이므로 Volume을 반응 변수로 삼음
- 반응 변수를 가로 축으로 하여 lm 함수를 사용해 다중 선형 회귀를 적용

```
Console C:/RSources/    
> m=lm(Volume ~ Girth + Height, data = trees)  
> m  
  
Call:  
lm(formula = volume ~ Girth + Height, data = trees)  
  
Coefficients:  
(Intercept)      Girth      Height  
   -57.9877    4.7082    0.3393
```

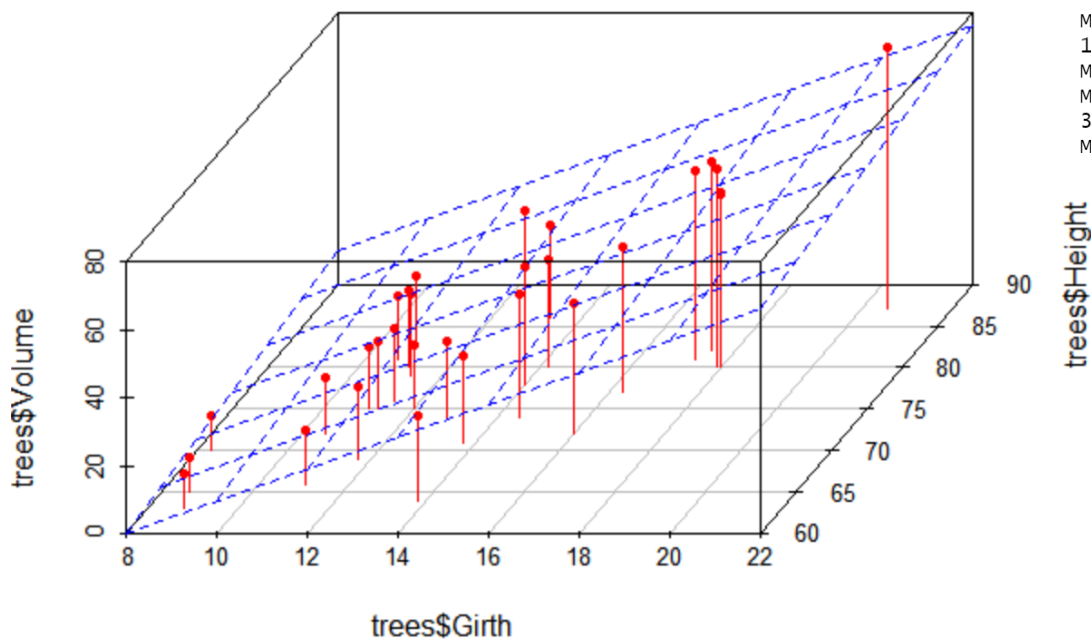
- 최적 모델 : $\text{Volume} = -57.9877 + 4.7082 \times \text{Girth} + 0.3393 \times \text{Height}$

7.7 다중 선형 회귀의 적용(trees data)

■ 모델을 가시화 하면

Console C:/RSources/

```
> s=scatterplot3d(trees$Girth, trees$Height, trees$Volume, xlim
=8:21, ylim=60:90, zlim=8:80, pch=20, type='h', color='red', an
gle=55)
> s$plane3d(m, col='blue')
```



```
> summary(trees)
```

Girth		Height		Volume	
Min.	: 8.30	Min.	: 63	Min.	: 10.20
1st Qu.	: 11.05	1st Qu.	: 72	1st Qu.	: 19.40
Median	: 12.90	Median	: 76	Median	: 24.20
Mean	: 13.25	Mean	: 76	Mean	: 30.17
3rd Qu.	: 15.25	3rd Qu.	: 80	3rd Qu.	: 37.30
Max.	: 20.60	Max.	: 87	Max.	: 77.00

7.7 다중 선형 회귀의 적용(trees data)

■ 예측

- 이제 목재상은 벗나무의 지름과 키를 측정하면 목재의 부피를 예측할 수 있음
- 자르기로 마음먹은 나무의 지름과 키를 재어 새로운 데이터 수집하고 예측

Console C:/RSources/ ↗

```
> ndata = data.frame(Girth=c(8.5, 13.0, 19.0), Height=c(72,86,85))  
> predict(m, newdata=ndata)  
          1          2          3  
6.457794 32.394034 60.303746
```

- 목재상은 세 그루를 자르면 총 99세제곱 피트 가량의 목재를 얻을 수 있음을 알게 됨
- 이 예측을 바탕으로 집을 짓는데 목재가 충분할지 더 잘라야 할지 판단



생각의 시간

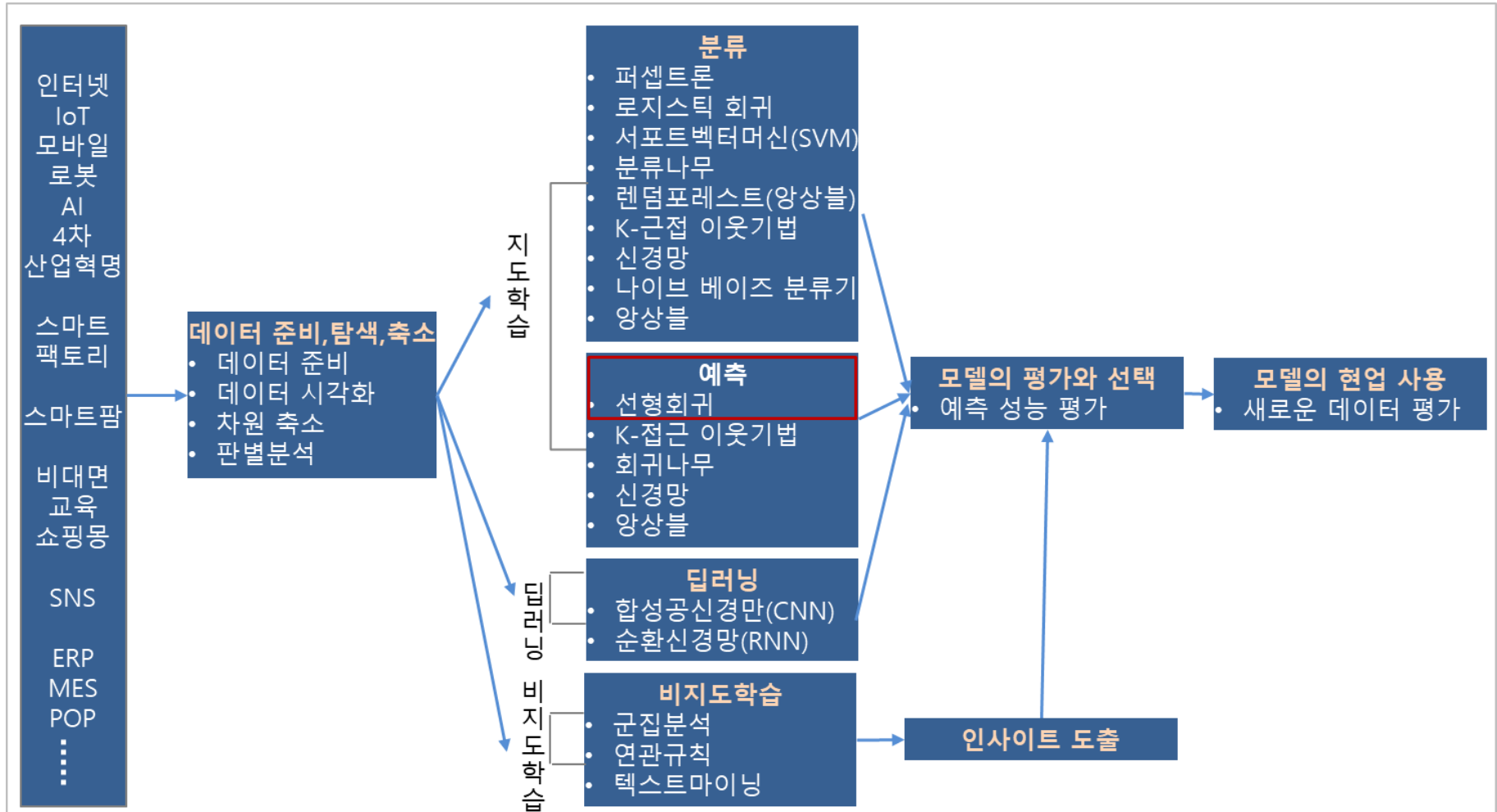
Ground-truth는 기상학에서 유래된 용어로 어느 장소에서 수집된 정보를 의미합니다.

Ground-truth는 보통 '**지상 실측 정보**'로 해석되며 인공위성과 같이 지구에서 멀리 떨어져서 지구를 관찰하였을 때 지구의 전체적인 관점을 보는 것에는 넓은 시야를 가질 수 있지만 실제 지면의 구조를 세밀하게 보는 것은 빛이 구름이나 대기를 통과하게 되면서 실제 모습이 왜곡되어 제대로 파악하는 것은 어렵습니다.

이러한 상황에서 지상 정보를 직접 측정한다면 보다 정확한 정보를 얻을 수 있는 것입니다. 이러한 정보에 인공위성에서 관측된 데이터를 참조하여 사용한다면 좀 더 정확한 데이터를 얻을 수 있습니다.



요약



Thank you

