



10주차: 일반화 선형 모델

ChulSoo Park

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



학습목표 (10주차)

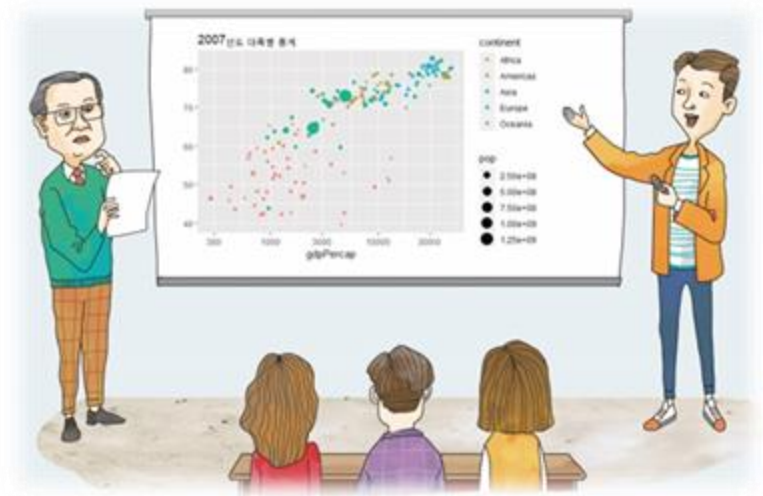
- ❖ 일반화 선형 모델의 이해
- ❖ 로지스틱 회귀의 이해
- ❖ 로지스틱 회귀 분석 및 모델링



08

CHAPTER

일반화 선형 모델



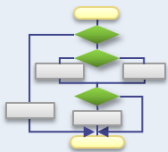

CONTENTS

- 8.1 일반화 선형 모델은 왜 필요한가?
- 8.2 일반화 선형 모델
- 8.3 로지스틱 회귀
- 8.4 로지스틱 회귀의 적용: UCLA admission 데이터
- 8.5 로지스틱 회귀의 적용 : colon 데이터
 - ※ 과잉적합
 - 요약



■ 모델링

- 현실 세계에서 일어나는 현상을 수학적식으로 표현하는 행위
- 모델링을 통해 모델을 알아내고 나면, 모델을 이용하여 새로운 사실을 예측(prediction)할 수 있음
- 데이터 분석 결과는 연구 대상에 대한 특징 설명 또는 어떤 값의 예측 형태로 나타남. 이를 위해 모델링 방식을 사용함
- 모델링이란 데이터를 발생시킨 원래 시스템을 설명하기 위해 설정한 구조
- 모델 표현 방법은 수식, 다이어그램, 알고리즘 등(데이터 사이언스 개론,p.171)

수식	다이어그램	알고리즘
가장 명쾌하나 현실적으로 많은 대상이 수식으로 설명하기 어려움	순서도 같은 형태의 처리 흐름도로 알고리즘에 비해 비교적 단순함	패턴 분석과 예측 모델 작성 등 보다 복잡한 논리적 흐름
$F = G \frac{m_1 m_2}{r^2}$ <p>중력 이론</p>		



■ 간단한 예로 알아보기(영업 사원의 월급)

- 자동차 판매회사의 신입 사원인 길동은 다음과 같이 계약

조건 : 100만원 기본급에 자동차 1대 팔 때마다 90만원을 추가로 받음

- 이 조건을 기반으로 모델링
 - ✓ 판매 대수를 x , 월급을 y 라 하고, x 를 독립 변수 y 를 종속 변수로 간주
 - ✓ 수식으로 표현하면

모델 : 월급(y) = 1,000,000(기본급) + 900,000 x (자동차 판매 대수)

- 위 수식을 **모델**이라 부름
- 길동이 급여 관련 변수를 찾고 변수 사이의 관계를 나타내는 수식을 구하는 과정을 **모델링**이라 부름
- 모델이 있으면 **예측**이 가능
 - ✓ 다음 달에 3대를 팔면 월급이 얼마일까? → 370만원
 - ✓ 더욱 분발하여 그 다음 달에 10대를 팔면? → 1000만원



■ 데이터 과학 세계의 모델링과 예측

- 데이터 사이언스 세계에서는 수집한 data로 모델링 작업
- 주어진 data을 일반적으로 훈련 데이터로부터 하나의 함수(모델)가 유추되고 나면 해당 함수에 대한 평가를 통해 파라미터를 최적화한다. 이러한 평가를 위해 교차 검증(Cross-Validation)이 이용되며 이를 위해 검증 집합을 다음의 세가지로 나눈다.(p,343 참조) 6:2:2

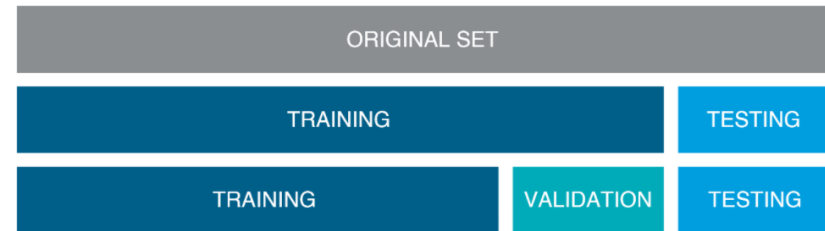
1. 훈련 집합 (A Training Set)
2. 검증 집합 (A Validation Set)
3. 테스트 집합 (A Test Set)



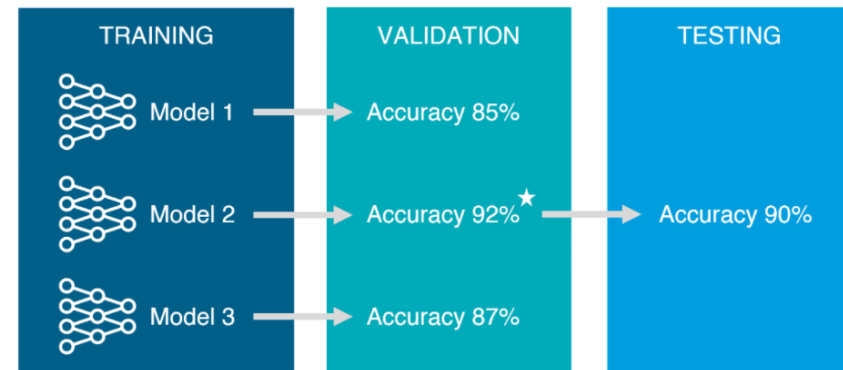
Training set의 목적 : Training set(훈련 데이터)은 모델을 학습하는데 사용된다.

Validation set의 목적 : Validation set(검정 데이터)은 training set으로 만들어진 모델의 성능을 측정하기 위해 사용된다.

Test set의 목적 : Test set(테스트 데이터)은 validation set으로 사용할 모델이 결정 된 후, 마지막으로 딱 한번 해당 모델의 예상되는 성능을 측정하기 위해 사용된다.



Training, validation and test set split



Training, validation and test set difference



■ 다중 선형 회귀 적용하기

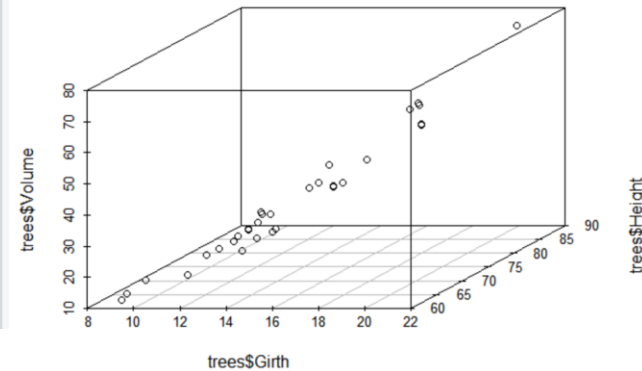
- 어떤 변수를 반응 변수로 하나? → 목재상이 알고자 하는 것은 나무의 상태에 따른 목재의 부피일 것이므로 Volume을 반응 변수로 삼음
- 반응 변수를 가로 축으로 하여 lm 함수를 사용해 다중 선형 회귀를 적용

```

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



부산 해수욕장 올해 여름 최대인파...77만명 몰려

송고시간 | 2018-07-14 17:23



김상현 기자
[기자 페이지](#)

(부산=연합뉴스) 김상현 기자 = 사흘째 폭염특보가 이어진 14일 부산지역 5대 해수욕장에 올해 여름 들어 가장 많은 77만명의 인파가 몰렸다.

부산지역은 이날 강렬한 햇볕 속에 낮 최고 33도를 기록하는 등 사흘째 찜통더위가 계속되면서 오전 일찍부터 많은 시민과 관광객들이 해운대와 광안리 등 주요 해수욕장을 찾았다.



■ 선형으로 표현하기 부적절한 데이터가 다수

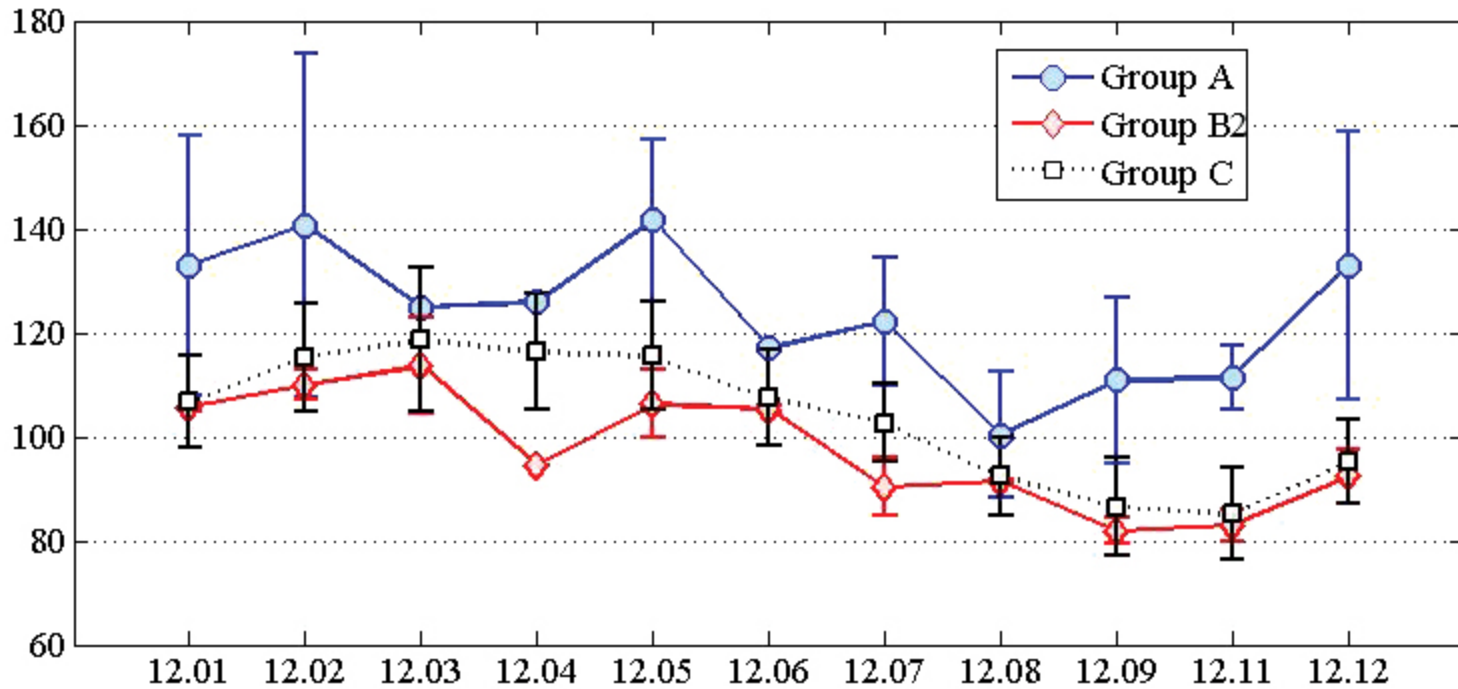
- 예) 해수욕장: 기온 x 와 방문객 수 y
 - 선형 모델 $y=1000x+200$ 을 사용하면,
 - x (기온)가 음수가 되면 방문객 수는 음수? (겨울 해수욕장 인파 ?)
 - 작은 해수욕장에는 적용 불가능 → 일반성이 매우 약한 모델
- 지수 관계가 더 적합
 - 기온이 10도 오를 때마다 방문객 수가 두 배로 증가한다고 모델링
 - 예를 들어, $y = 2000 * 2^{0.1x} + 200$

■ 8장에서 다룰 내용

- 일반화 선형 모델과 glm(generalized linear model) 함수
- 범주형 변수 다루기
- glm으로 구한 모델의 통계량 해석



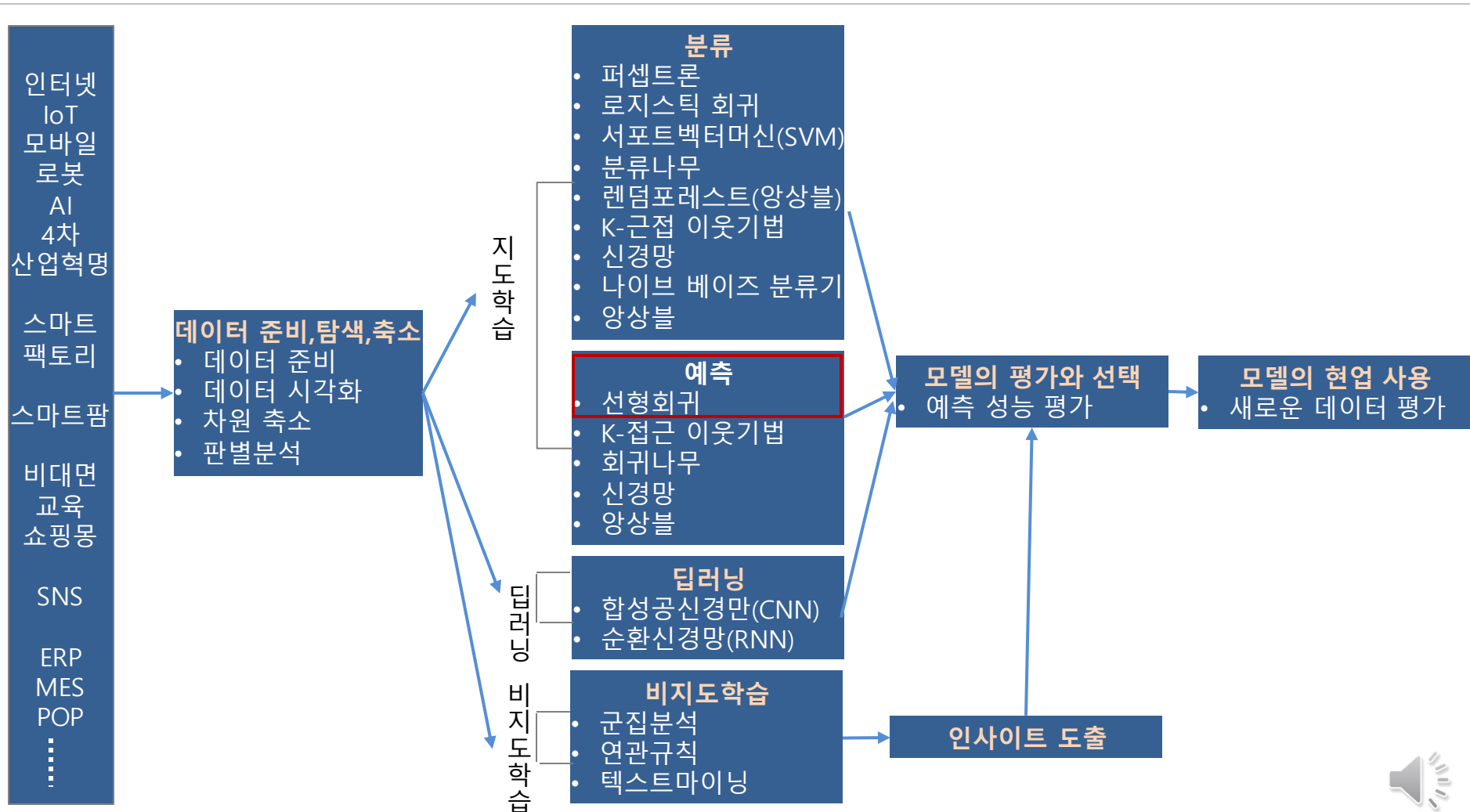
■ 선형으로 표현하기 부적절한 데이터가 다수



출처 : 대형할인점 입점이 낙농 유제품 대리점의 성과에 미치는 영향 및 평가요인으로서의 중요도 분석(유통연구,2014년)



■ 데이터 분석 Process에서 이번주 교육 위치



8.1 일반화 선형 모델은 왜 필요한가?

■ 아이스크림 판매량 예측

- 예) 기온이 두 배 오르면 판매량이 두 배 오른다고 모델링하면, 20도에서 40도가 될 때 판매량이 두 배가 됨
- 현실에서는 쾌적한 27도에서 덥게 느껴지는 30도로 증가할 때 두 배로 늘 가능성이 큼 → 3도 오를 때마다 두 배로 증가한다는 가정이 더 어울림
- 기온과 판매량이 선형 관계라고 가정하면 여러 측면에서 모순 발생
- 프로모션에 따라 매출이 2배 이산 차이 발생 가능(예, 1+1 할인 행사)
- 기온 외에 다양한 영향 변수 존재(사례 논문 참조)



8.1 일반화 선형 모델은 왜 필요한가?

기상요인이 식음료업의 매출에 미치는 영향 분석(논문) 사례

연구자	기상요인						
	기온	강우량	강설량	습도	일조량	풍속	날씨(맑음, 구름, 비 등)
홍진환 외(2012)	○	○		○			
장은영 · 임병훈(2003)	○	○				○	
장은영 · 이선재(2002)	○	○			○	○	
이용기 외(2011)	○			○			○
Kyle B. Murray et al.(2010)	○		○	○	○		
Andrew G. Parsons(2001)	○	○					

부동산학연구 제23집 제1호, 2017, 3, pp. 61~72
http://dx.doi.org/10.19172/KREAA.23.1.5
Journal of the Korea Real Estate Analysts Association
Vol.23, No.1 2017, 3, pp. 61~72

기상요인이 식음료업의 매출에 미치는 영향 분석*

Analysis of the Effect of Meteorological Factors on the Sales of the Food and Beverage Services

성 은 영 (Seong, Eun Yeong)**
성 현 곤 (Sung, Hyun Gun)***
최 창 규 (Choi, Chang Gyu)****

< Abstract >

Weather and meteorological condition have direct and indirect effects on consumers' emotion and consumption pattern. Traditionally, physical location characteristics have been the major interest in the urban planning and real estate fields, and there has been insufficient research on the effect of meteorological characteristics on sales. Location characteristics and meteorological characteristics could affect sales individually and in combination. In this study, the effects of location characteristics and meteorological factors on sales were empirically analyzed using the five-month sales of 8 stores in the food and beverage services in Seoul as the dependent variable. The results of the analysis are as follows. First, the Two-way random effect model was found to be more sophisticated than the general regression model, when analyzing the sales of the food and beverage services using the panel data. Second, the daily mean temperature, rain and the humidity were the meteorological variables that had a significant effect on the sales of the food and beverage services. The

구분	변수		구축방법	단위	자료출처
종속변수	일매출		일매출	ln원	POS 데이터
독립변수	기상특성	일평균기온	관측일 평균 기온	℃	기상청
		강우량	관측일 일강우량	mm	기상청
		일평균풍속	관측일 평균풍속	m/s	기상청
		일평균습도	관측일 평균 습도	%	기상청
		미세먼지	PM10 기준 30μg/m ³ 미만=0, 30μg/m ³ 이상=1		기상청
	입지특성	버스정류장까지 거리	매장에서 버스정류장까지 거리	m	직접조사
		지하철역까지 거리	매장에서 지하철역까지 거리	m	직접조사
		전면도로 너비	전면 도로의 차선 수	차선	직접조사
		매장면적	매장면적	m ²	건축물대장
		업종특성	알콜 및 음료업=0, 음식업=1		직접조사

출처 : 부동산학연구 제23집 제1호, 2017,성은영 외2명

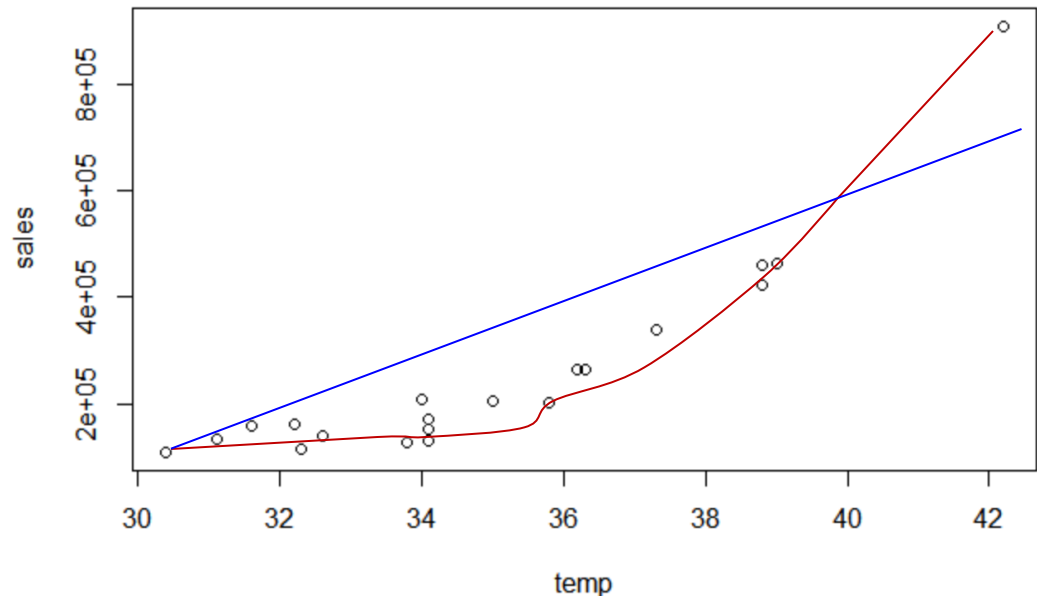


8.1 일반화 선형 모델은 왜 필요한가?

■ 예제) 영희의 아이스크림 장사

- 해수욕장에서 아이스크림을 팔면서 10일 동안 기온과 판매량을 기록하여 데이터 수집
- 선형에서 크게 벗어남
- 선형 모델 `lm(linear model)` 대신 일반화 선형 모델인 `glm(generalized linear model)` 적용
- 반응 변수가 두 가지 값만 가지는 경우(예, 상품과 하품, 환자와 정상인)

temp	sales
38.8	423000
34.0	207900
39.0	464600
38.8	460000
36.2	264500
30.4	107500
32.2	161600
34.1	131200
35.0	206000
42.2	910400
37.3	338600
32.6	138300
31.6	157400
34.1	172100
34.1	153000
33.8	127200
35.8	200600
32.3	116100
36.3	265200
31.1	132500



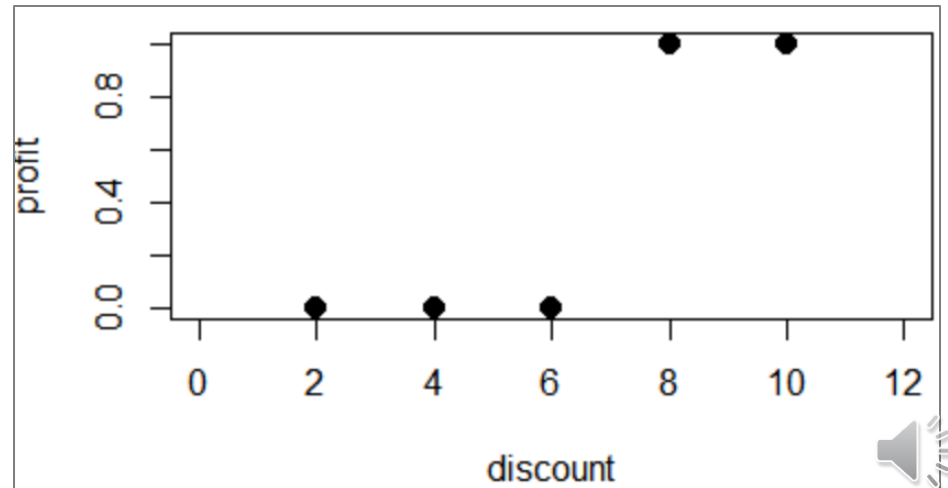
8.1 일반화 선형 모델은 왜 필요한가?

■ 예제) 할인율에 따른 이익 “머플러 판매 데이터”

- 가판에서 목도리를 팔면서 데이터 수집, 순이익은 5만원 미만이면 0, 넘으면 1로 기록

```
muffler=data.frame(discount=c(2.0, 4.0, 6.0, 8.0, 10.0),profit=c(0,0,0,1,1))
plot(muffler,pch=20, cex=2, xlim=c(0,12))
```

```
> muffler=data.frame(discount=c(2.0, 4.0, 6.0, 8.0, 10.0),profit=c(0,0,0,1,1))
> plot(muffler,pch=20, cex=2, xlim=c(0,12))
> head(muffler)
  discount profit
1         2      0
2         4      0
3         6      0
4         8      1
5        10      1
> |
```



8.1 일반화 선형 모델은 왜 필요한가?

■ (부적절한) 모델 적용

- lm을 적용하고 모델의 적정성을 확인한다.

```
Console C:/RSources/ ↗  
> head(muffler)  
  discount profit  
1         2      0  
2         4      0  
3         6      0  
4         8      1  
5        10      1  
> m=lm(profit~discount, data=muffler) # 최적 모델 계산(선형 회귀)  
> coef(m) # 계수를 계산하는 함수  
(Intercept) discount  
      -0.50      0.15  
> fitted(m) # 훈련 집합에 있는 sample에 대한 예측 결과  
      1      2      3      4      5  
-0.2  0.1  0.4  0.7  1.0  
> residuals(m) # 잔차  
      1      2      3      4      5  
2.000000e-01 -1.000000e-01 -4.000000e-01  3.000000e-01  1.249001e-16  
> deviance(m) # 잔차제곱  
[1] 0.3
```



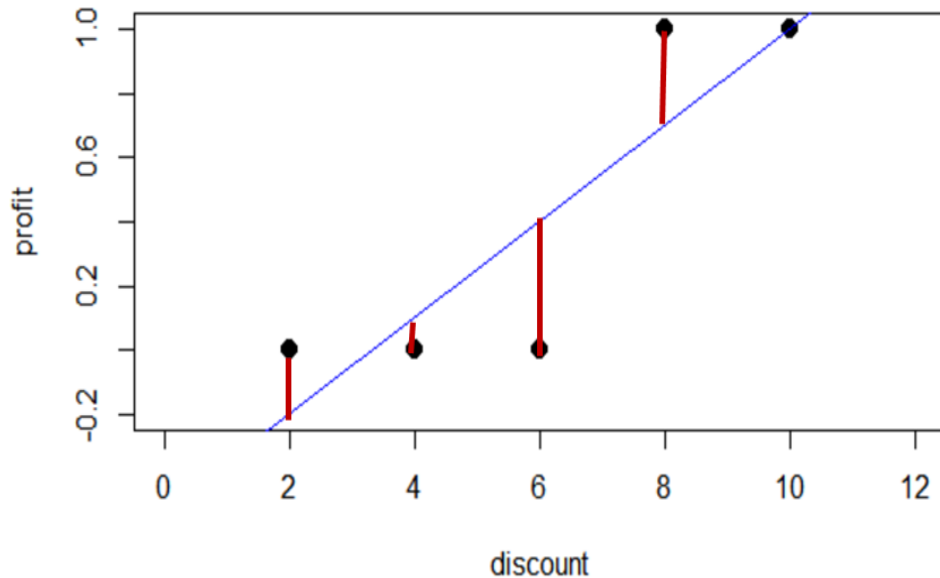
8.1 일반화 선형 모델은 왜 필요한가?

■ (부적절한) 모델 적용

- 모델 적정성 확인 (오차 분석)
- 최적 모델 : $\text{profit} = 0.15 \times \text{discount} - 0.5$

```
plot(muffler, pch=20, cex=2, xlim=c(0,12), ylim=c(-0.2,1.0))
abline(m, col="blue")
```

```
> head(muffler)
  discount profit
1         2      0
2         4      0
3         6      0
4         8      1
5        10      1
```



discount	2	4	6	8	10
Profit(예측)	-0.2	0.1	0.4	0.7	1.0
Profit(실측)	0	0	0	1	1
오차	0.2	-0.1	-0.4	0.3	0



8.1 일반화 선형 모델은 왜 필요한가?

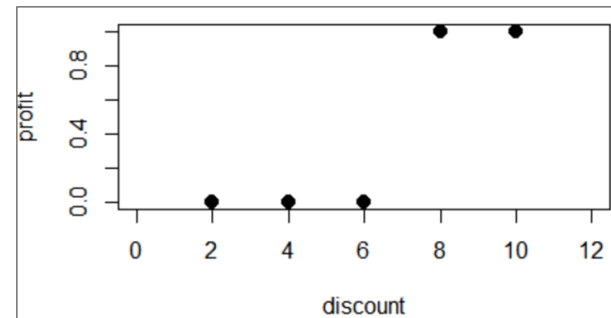
■ (부적절한) 모델 적용

- 새로운 할인율에 대해 예측 실행 (할인율 : 1, 5, 12, 20, 30%)
- 오차를 확인
- 오차가 크다는 것을 확인 : 실행 결과 -0.35, 0.25, 1.30, 2.50, 4.0을 확인 할 수 있으며 이는 순이익 0(5만원 미만)과 1(5만원 이상)로 표시한 것과 큰 차이 발생

Console C:/RSources/ ➡

```
> newdisc=data.frame(discount=c(1,5,12,20,30)) # 5개의 새로운 할인율
> p=predict(m, newdisc)
> p
```

	1	2	3	4	5
	-0.35	0.25	1.30	2.50	4.00

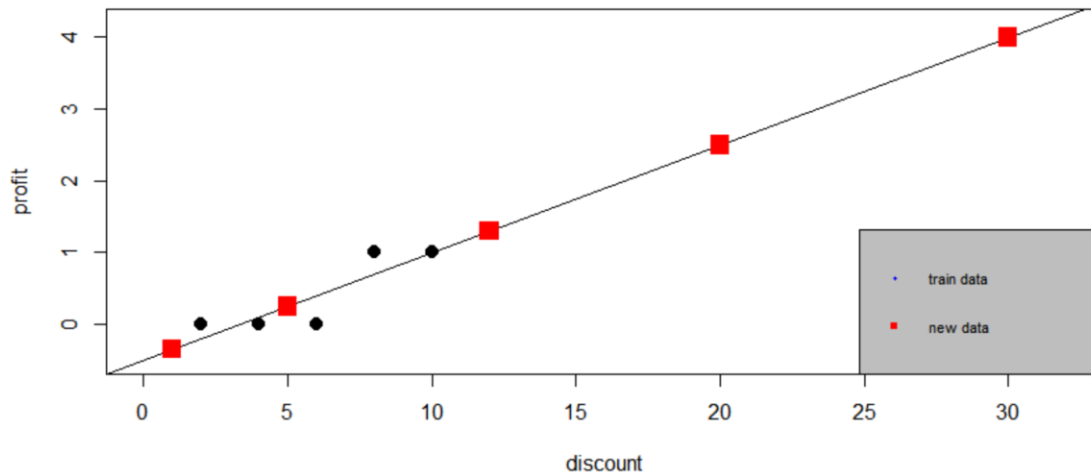


8.1 일반화 선형 모델은 왜 필요한가?

■ (부적절한) 모델 적용

- 지금까지 분석한 결과를 정리하면 다음과 같다.

```
plot(muffler,pch=20, cex=2, xlim=c(0,32), ylim=c(-0.5,4.2))
abline(m)
rest=data.frame(discount=newdisc, profit=pred)
points(rest, pch=15, cex=2, col='red')
legend("bottomright", legend=c("train data","new data"), pch=c(20,15),
cex=0.7, col=c("blue","red"),bg="gray")
```



Thank you

