



10주차: 일반화 선형 모델

ChulSoo Park

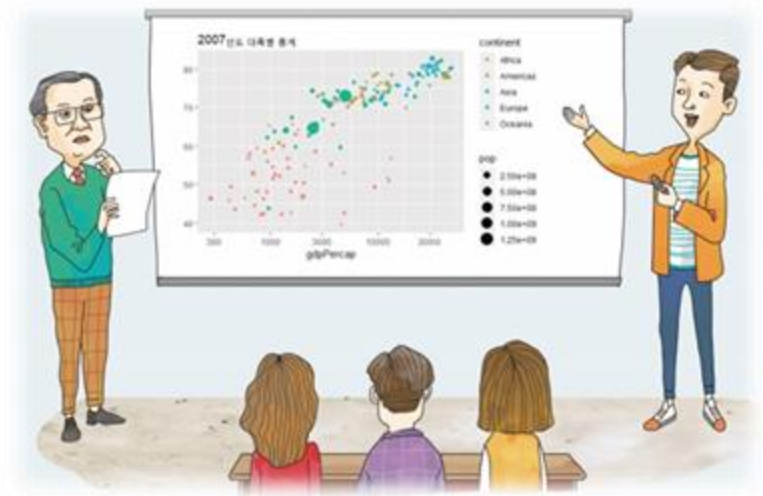
School of Computer Engineering & Information Technology
Korea National University of Transportation



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



8.2 일반화 선형 모델

- 머플러 판매 데이터에 일반 선형 모델인 glm 함수 적용
 - 이전과 달라진 것은 lm이 glm이 되고, family=binomial 옵션을 추가한 것
 - binomial 옵션은 반응 변수인 profit이 두 가지 값만 가진다고 glm에게 알려주는 역할

```
Console C:/RSources/    
> muffler=data.frame(discount=c(2.0, 4.0, 6.0, 8.0, 10.0),profit=c(0,0,0,1,1))  
> muffler  
  discount profit  
1         2      0  
2         4      0  
3         6      0  
4         8      1  
5        10      1  
> rest_glm=glm(profit~discount, data=muffler, family = binomial)  
경고메시지(들):  
glm.fit: 적합된 확률값들이 0 또는 1 입니다  
> coef(rest_glm)  
(Intercept)    discount  
-160.80782     22.98592  
> fitted(rest_glm)  
      1      2      3      4      5  
2.220446e-16 2.220446e-16 1.142877e-10 1.000000e+00 1.000000e+00  
> residuals(rest_glm)  
      1      2      3      4      5  
-2.107342e-08 -2.107342e-08 -1.511871e-05 1.376758e-05 2.107342e-08  
> deviance(rest_glm)  
[1] 4.181229e-10
```



8.2 일반화 선형 모델

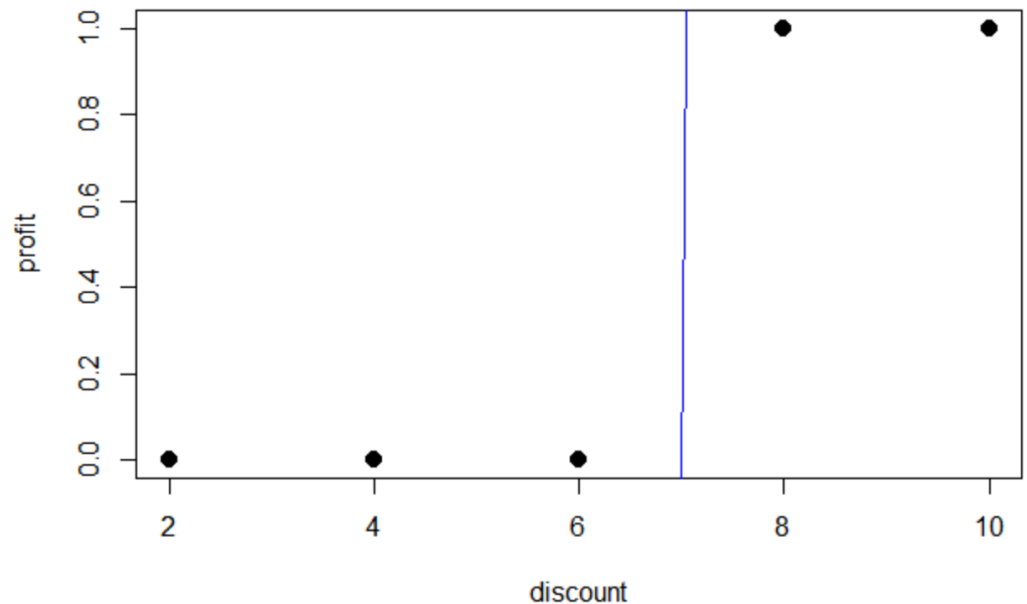
- 머플러 판매 데이터에 일반 선형 모델인 glm 함수 적용

```
> plot(muffler, pch=20, cex=2)
```

```
> abline(rest_gml, col='blue', lwd=1)
```

line width


Discount (%)	Profit(0 or 1)
2	0
4	0
6	0
8	1
10	1



8.2 일반화 선형 모델

■ 머플러 판매 데이터에 glm 함수 적용

- 새로운 데이터(할인율 1%, 5%, 12%, 20%, 30%)에 대한 예측
- 적용 모델 : $\text{profit} = 22.98592 \times \text{discount} - 160.80782$
- 예측 결과 : $\approx 0, 0, 1, 1, 1$

Console C:/RSources/ 

```
> newdisc=data.frame(discount=c(1,5,12,20,30)) # 5개의 새로운 할인율
> pred=predict(rest_glm, newdisc, type='response')
> pred
```

1	2	3	4	5
2.220446e-16	2.220446e-16	1.000000e+00	1.000000e+00	1.000000e+00

```
> |
```

`type=c("link", "response", "terms")` : 예측 결과의 유형을 지정한다. 기본값은 "link"이다.

- ✓ link : 선형 독립 변수들의 연산 결과의 크기로 값을 반환한다.
- ✓ response : 반응변수의 크기로 값을 반환하며 로지스틱 회귀의 경우 확률이다.
- ✓ terms : 행렬에 모델 포물리의 각 변수에 대한 적합된 값을 선형 예측 변수의 크기로 반환한다.



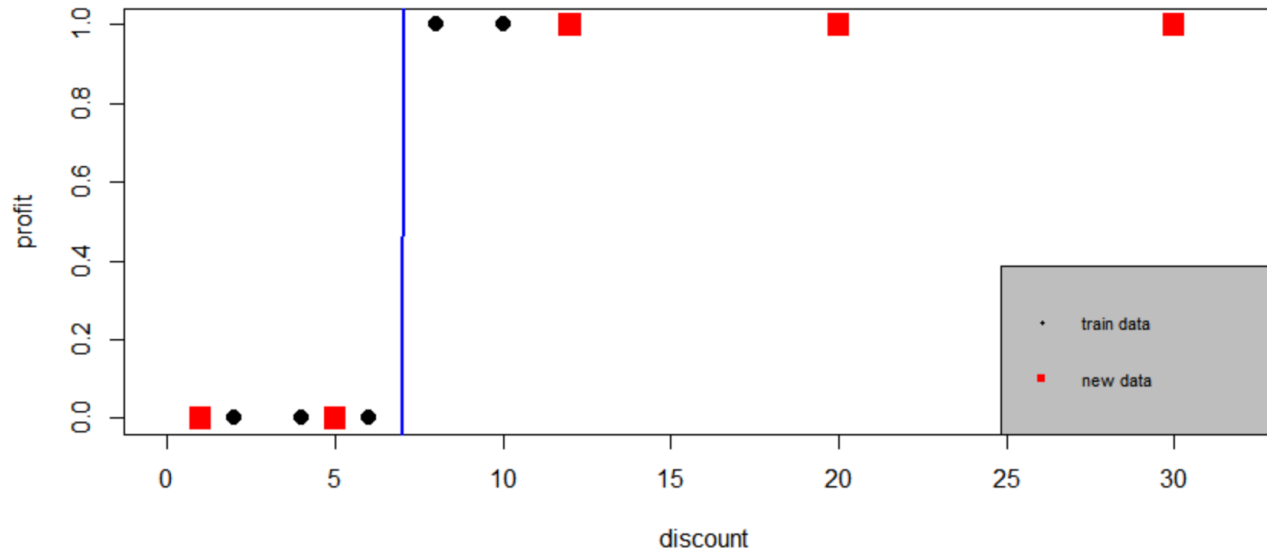
8.2 일반화 선형 모델

■ 머플러 판매 데이터에 glm 함수 적용

- 예측 결과를 기준으로 그래프를 그리면
- ```

> plot(muffler, pch=20, cex=2, xlim=c(0,32))
> abline(rest_gml, col='blue', lwd=2)
> res = data.frame(discount=newdisc, profit=pred)
> points(res, pch=15, cex=2, col='red')
> legend("bottomright", legend = c("train data", "new data"), pch=c(20,15),
 cex=0.7,col=c("black","red"),bg="gray")

```



## 8.2 일반화 선형 모델

- Haberman survival 읽어 들이고 확인하기
  - UCI 리퍼지토리에서 제공
  - URL은 웹에서 이름으로 검색
  - 네 개의 변수
    - 설명 변수: 수술 받을 당시 나이, 수술 연도, 양성 림프샘 개수
    - 반응 변수: 수술 후 생존 연수(5년 이상은 1, 5년 이내는 2)

← → ↻ archive.ics.uci.edu/ml/datasets/haberman's+survival



### Haberman's Survival Data Set

Download: [Data Folder](#), [Data Set Description](#)

**Abstract:** Dataset contains cases from study conducted on the survival of patients who had undergone surgery for breast cancer

|                                   |                |                              |     |                            |            |
|-----------------------------------|----------------|------------------------------|-----|----------------------------|------------|
| <b>Data Set Characteristics:</b>  | Multivariate   | <b>Number of Instances:</b>  | 306 | <b>Area:</b>               | Life       |
| <b>Attribute Characteristics:</b> | Integer        | <b>Number of Attributes:</b> | 3   | <b>Date Donated</b>        | 1999-03-04 |
| <b>Associated Tasks:</b>          | Classification | <b>Missing Values?</b>       | No  | <b>Number of Web Hits:</b> | 229607     |



## 8.2 일반화 선형 모델

### ■ Haberman survival Data Set Description

1. The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.
2. Number of Instances: 306
3. Number of Attributes: 4 (including the class attribute)
4. Attribute Information:
  - ① Age of patient at time of operation (numerical)
  - ② Patient's year of operation (year - 1900, numerical)
  - ③ Number of positive axillary nodes detected (numerical)
  - ④ Survival status (class attribute)
    - 1 = the patient survived 5 years or longer
    - 2 = the patient died within 5 year
5. Missing Attribute Values: None





## 8.2 일반화 선형 모델

### ■ Haberman survival 읽어 들이고 확인하기

- 먼저 data를 받은 후 메모장이나 엑셀로 데이터 확인
- header 가 없는 data 확인

Console C:/RSources/ ➔

```
> haberman=read.csv("c:/rdata/haberman.csv",header=FALSE)
> names(haberman)=c('age','op_year','no_nodes','survival')
> str(haberman)
'data.frame': 306 obs. of 4 variables:
 $ age : int 30 30 30 31 31 33 33 34 34 34 ...
 $ op_year : int 64 62 65 59 65 58 60 59 66 58 ...
 $ no_nodes : int 1 3 0 2 4 10 0 0 9 30 ...
 $ survival : int 1 1 1 1 1 1 1 2 2 1 ...
> head(haberman)
 age op_year no_nodes survival
1 30 64 1 1
2 30 62 3 1
3 30 65 0 1
4 31 59 2 1
5 31 65 4 1
6 33 58 10 1
```

데이터를 다운로드한 후  
아래 엑셀 시트를  
확인(바로 연결 사용 가능)

| 파일   | 홈    | 삽입 | 페이지 레이아웃 | 수식   | 데이터 |
|------|------|----|----------|------|-----|
| 잘라내기 | 붙여넣기 | 복사 | 서식 복사    | 클립보드 | 글꼴  |
| F11  |      |    |          |      |     |
|      | A    | B  | C        | D    |     |
| 1    | 30   | 64 | 1        | 1    |     |
| 2    | 30   | 62 | 3        | 1    |     |
| 3    | 30   | 65 | 0        | 1    |     |
| 4    | 31   | 59 | 2        | 1    |     |
| 5    | 31   | 65 | 4        | 1    |     |
| 6    | 33   | 58 | 10       | 1    |     |
| 7    | 33   | 60 | 0        | 1    |     |

haberman.csv - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

30,64,1,1  
30,62,3,1  
30,65,0,1  
31,59,2,1  
31,65,4,1  
33,58,10,1  
33,60,0,1  
34,59,0,2  
34,66,9,2  
34,58,30,1  
34,60,1,1



## 8.2 일반화 선형 모델

### ■ 변수의 유형

#### 1. 기능에 따른 분류

- ① 독립 변수
- ② 종속 변수(반응 변수)
- ③ 매개 변수
- ④ 조절 변수와 통제 변수

**survival 변수는 ?**

#### 2. 측정 수준에 따른 분류

- ① 명목 변수(성별, 종교, 직업 등)
- ② 서열 변수(범주 간의 순위를 매겨질 수 있는 변수)
- ③ 등간 변수(①, ②의 속성을 가지며 변수 값 간의 간격을 알 수 있는 변수(온도, 지능지수, 학년 등))
- ④ 비율 변수((①, ②, ③의 속성을 다 가지면서 절대 영점의 의미 추가, 모든 수학적 조작 가능))



## 8.2 일반화 선형 모델

### ■ Haberman survival 읽어 들이고 확인하기

- 먼저 data를 받은 후 메모장이나 엑셀로 데이터 확인
- 년도 : 끝 두 자리 사용
- 변수 4개 전체가 정수형 임.
- survival data는 정수 아니라 0과 1의 범주형 변수임

Console C:/RSources/ ↗

```
> head(haberman)
 age op_year no_nodes survival
1 30 64 1 1
2 30 62 3 1
3 30 65 0 1
4 31 59 2 1
5 31 65 4 1
6 33 58 10 1

> str(haberman)
'data.frame': 306 obs. of 4 variables:
 $ age : int 30 30 30 31 31 33 33 34 34 34 ...
 $ op_year : int 64 62 65 59 65 58 60 59 66 58 ...
 $ no_nodes : int 1 3 0 2 4 10 0 0 9 30 ...
 $ survival : int 1 1 1 1 1 1 1 2 2 1 ...
```



## 8.2 일반화 선형 모델




- Haberman survival 읽어 들이고 확인 및 변경
  - survival data는 범주형이나 0과 1의 범주형으로 변경

```
Console C:/RSources/
> haberman=read.csv("c:/rdata/haberman.csv",header=FALSE)
> names(haberman)=c('age','op_year','no_nodes','survival')
> head(haberman)
 age op_year no_nodes survival
1 30 64 1 1
2 30 62 3 1
3 30 65 0 1
4 31 59 2 1
5 31 65 4 1
6 33 58 10 1
> str(haberman)
'data.frame': 306 obs. of 4 variables:
 $ age : int 30 30 30 31 31 33 33 34 34 34 ...
 $ op_year : int 64 62 65 59 65 58 60 59 66 58 ...
 $ no_nodes : int 1 3 0 2 4 10 0 0 9 30 ...
 $ survival: int 1 1 1 1 1 1 1 2 2 1 ...
> haberman$survival=factor(haberman$survival)
> str(haberman)
'data.frame': 306 obs. of 4 variables:
 $ age : int 30 30 30 31 31 33 33 34 34 34 ...
 $ op_year : int 64 62 65 59 65 58 60 59 66 58 ...
 $ no_nodes : int 1 3 0 2 4 10 0 0 9 30 ...
 $ survival: Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 2 2 1 ...
```



## 8.2 일반화 선형 모델

- 일반화 선형 모델 적용(glm)
  - survival data는 범주형이나 0과 1의 범주형으로 변경

```
Console C:/RSources/   
> resh=glm(survival~age+op_year+no_nodes, data=haberman,family=binomial)
> coef(resh) # 계수를 계산하는 함수
(Intercept) age op_year no_nodes
-1.86162525 0.01989935 -0.00978386 0.08844244
> resh=glm(survival~., data=haberman,family=binomial)
> coef(resh) # 계수를 계산하는 함수
(Intercept) age op_year no_nodes
-1.86162525 0.01989935 -0.00978386 0.08844244
> deviance(resh) # 잔차제곱
[1] 328.2564
```

모델을 구했으니 새로운 환자가 오면 생존 여부를 예측할 수 있다.



## 8.2 일반화 선형 모델

### ■ 일반화 선형 모델 적용(glm)

- 모델을 가지고 새로운 환자 생존율 예측
  - ✓ 환자 1 : 나이 37, 수술 연도 1958, 림프샘 개수 5
  - ✓ 환자 2 : 나이 66, 수술 연도 1960, 림프샘 개수 32
- predict 함수에 type='response' option에 주목

Console C:/RSources/ ↗

```
> new_patients1=data.frame(age=c(37),op_year=c(58),no_nodes=c(5))
> predict(resh,newdata=new_patients1,type='response')
 1
0.2225961
> new_patients2=data.frame(age=c(66),op_year=c(60),no_nodes=c(32))
> predict(resh,newdata=new_patients2,type='response')
 1
0.844862
```

type=c("link", "response", "terms")

# 예측 결과의 유형을 지정한다. 기본값은 "link"이다.

# link : 선형 독립 변수들의 연산 결과의 크기로 값을 반환한다.

# response : 반응변수의 크기로 값을 반환하며 로지스틱 회귀의 경우 확률이다.

# terms : 행렬에 모델 포물러의 각 변수에 대한 적합된 값을 선형 예측 변수의 크기로 반환한다.

## 8.2 일반화 선형 모델

### ■ 일반화 선형 모델 적용(glm) 결과 해석

| 환자   | 나이 | 수술 연도 | 림프샘 개수 |
|------|----|-------|--------|
| 환자 1 | 33 | 1958  | 5      |
| 환자 2 | 66 | 1960  | 32     |

```

Console C:/RSources/ ↗
> predict(resh,newdata=new_patients1,type='response')
 1
0.2225961
> predict(resh,newdata=new_patients2,type='response')
 1
0.844862
1 = the patient survived 5 years or longer
2 = the patient died within 5 year

```

### ■ 결과 해석

- 예측 결과 [0, 1] 사이의 확률로 5년 내에 사망할 확률
- 즉 환자1은 5년 이내에 사망할 확률 22% 1년 이상 생존 확률 78%
- 환자 2는 5년 이내에 사망할 확률 84%이고 5년 이상 생존할 확률은 16%이다



## 8.2 일반화 선형 모델

### ■ 일반화 선형 모델 특징 선택 (feature selection)

- 어떤 기준에 따라 일부 설명 변수만 선택하는 작업을 특징 선택이라 함
- 설명 변수의 중요도를 계산하고 중요도가 높은 변수를 자동으로 선택
- 예제) Haberman survival의 경우 수술 연도는 생존에 영향을 미치지 않는다고 생각하고 수작업으로 제외
- 분석 결과 : 3개의 변수의 잔차 328.2564와 2개 변수의 잔차 328.3107로 3개의 변수를 선택하는 것이 유리하지 않음을 알 수 있다.

```
Console C:/RSources/ ↗
> resh=glm(survival~age+no_nodes, data=haberman,family=binomial)
> coef(resh) # 계수를 계산하는 함수
(Intercept) age no_nodes
-2.46289804 0.01965028 0.08832453
> deviance(resh) # 잔차제공
[1] 328.3107
>
> new_patients=data.frame(age=c(37,66),no_nodes=c(5,32))
> predict(resh,newdata=new_patients,type='response')
 1 2
0.2151402 0.8402924
```





## 8.2 일반화 선형 모델

### ■ 일반화 선형 모델 특징 선택 (feature selection) 방법

Predict the selling price of Toyota corolla...



Dependent variable  
(target)

Independent variables  
(attributes, features)

| Variable      | Description                          |
|---------------|--------------------------------------|
| Price         | Offer Price in EUROS                 |
| Age_08_04     | Age in months as in August 2004      |
| KM            | Accumulated Kilometers on odometer   |
| Fuel_Type     | Fuel Type (Petrol, Diesel, CNG)      |
| HP            | Horse Power                          |
| Met_Color     | Metallic Color? (Yes=1, No=0)        |
| Automatic     | Automatic (Yes=1, No=0)              |
| CC            | Cylinder Volume in cubic centimeters |
| Doors         | Number of doors                      |
| Quarterly_Tax | Quarterly road tax in EUROS          |
| Weight        | Weight in Kilograms                  |



## 8.2 일반화 선형 모델

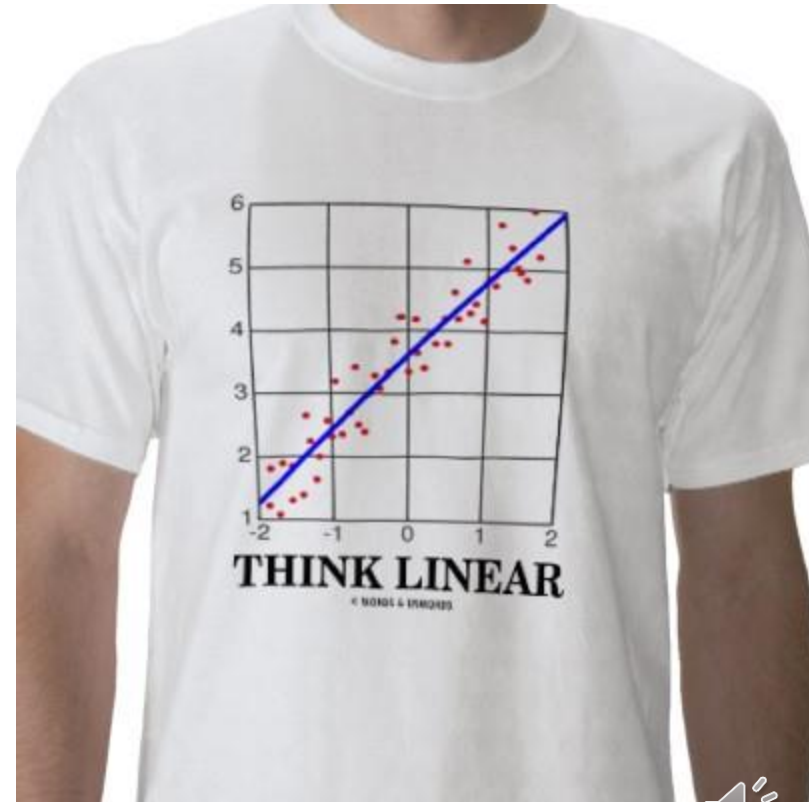
### ■ 일반화 선형 모델 특징 선택 (feature selection) 방법

#### Goal

- Fit a linear relationship between a quantitative dependent variable  $Y$  and a set of predictors  $X_1, X_2, \dots, X_p$ .

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

coefficients
unexplained



## 8.2 일반화 선형 모델

Example: predict the selling price of Toyota corolla

Y

X

| Price | Age_08_04 | KM    | Fuel_Type | HP  | Met_Color | Automatic | cc   | Doors | Quarterly_Tax | Weight |
|-------|-----------|-------|-----------|-----|-----------|-----------|------|-------|---------------|--------|
| 13500 | 23        | 46986 | Diesel    | 90  | 1         | 0         | 2000 | 3     | 210           | 1165   |
| 13750 | 23        | 72937 | Diesel    | 90  | 1         | 0         | 2000 | 3     | 210           | 1165   |
| 13950 | 24        | 41711 | Diesel    | 90  | 1         | 0         | 2000 | 3     | 210           | 1165   |
| 14950 | 26        | 48000 | Diesel    | 90  | 0         | 0         | 2000 | 3     | 210           | 1165   |
| 13750 | 30        | 38500 | Diesel    | 90  | 0         | 0         | 2000 | 3     | 210           | 1170   |
| 12950 | 32        | 61000 | Diesel    | 90  | 0         | 0         | 2000 | 3     | 210           | 1170   |
| 16900 | 27        | 94612 | Diesel    | 90  | 1         | 0         | 2000 | 3     | 210           | 1245   |
| 18600 | 30        | 75889 | Diesel    | 90  | 1         | 0         | 2000 | 3     | 210           | 1245   |
| 21500 | 27        | 19700 | Petrol    | 192 | 0         | 0         | 1800 | 3     | 100           | 1185   |
| 12950 | 23        | 71138 | Diesel    | 69  | 0         | 0         | 1900 | 3     | 185           | 1105   |
| 20950 | 25        | 31461 | Petrol    | 192 | 0         | 0         | 1800 | 3     | 100           | 1185   |
| 19950 | 22        | 43610 | Petrol    | 192 | 0         | 0         | 1800 | 3     | 100           | 1185   |
| 19600 | 25        | 32189 | Petrol    | 192 | 0         | 0         | 1800 | 3     | 100           | 1185   |
| 21500 | 31        | 23000 | Petrol    | 192 | 1         | 0         | 1800 | 3     | 100           | 1185   |
| 22500 | 32        | 34131 | Petrol    | 192 | 1         | 0         | 1800 | 3     | 100           | 1185   |
| 22000 | 28        | 18739 | Petrol    | 192 | 0         | 0         | 1800 | 3     | 100           | 1185   |
| 22750 | 30        | 34000 | Petrol    | 192 | 1         | 0         | 1800 | 3     | 100           | 1185   |
| 17950 | 24        | 21716 | Petrol    | 110 | 1         | 0         | 1600 | 3     | 85            | 1105   |
| 16750 | 24        | 25563 | Petrol    | 110 | 0         | 0         | 1600 | 3     | 19            | 1065   |



## 8.2 일반화 선형 모델

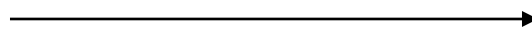
### With six variables

#### The Regression Model

| Input variables  | Coefficient  | Std. Error  | p-value    | SS          |
|------------------|--------------|-------------|------------|-------------|
| Constant term    | -3874.492188 | 1415.003052 | 0.00640071 | 97276411904 |
| Age_08_04        | -123.4366303 | 3.33806777  | 0          | 8033339392  |
| KM               | -0.01749926  | 0.00173714  | 0          | 251574528   |
| Fuel_Type_Petrol | 2409.154297  | 319.5795288 | 0          | 5049567     |
| HP               | 19.70204735  | 4.22180223  | 0.00000394 | 291336576   |
| Quarterly_Tax    | 16.88731384  | 2.08484554  | 0          | 192390864   |
| Weight           | 15.91809368  | 1.26474357  | 0          | 281026176   |

#### Training Data scoring - Summary Report

Model Fit



| Total sum of squared errors | RMS Error   | Average Error |
|-----------------------------|-------------|---------------|
| 1516825972                  | 1326.521353 | -0.000143957  |

#### Validation Data scoring - Summary Report

Predictive performance



(compare to 12-predictor model!)

| Total sum of squared errors | RMS Error   | Average Error |
|-----------------------------|-------------|---------------|
| 1021510219                  | 1334.029433 | 118.4483556   |



## 8.3 로지스틱 회귀

### ■ 로지스틱 회귀

- 목적 : 일반적인 회귀 분석의 목표와 동일하게 종속 변수와 독립 변수간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용
- 반응 변수가 두 가지 값만 가지는 경우의 회귀  
(참/거짓, 성공/실패, 환자/정상, 사망/생존, 승리/패배 등)
- 로지스틱 분석은 독립 변수의 선형 결합으로 종속 변수를 설명한다는 관점에서는 선형 회귀 분석과 유사하다. 하지만 로지스틱 회귀는 선형 회귀 분석과는 다르게 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나뉘기 때문에 일종의 분류 (classification) 기법으로도 볼 수 있다.



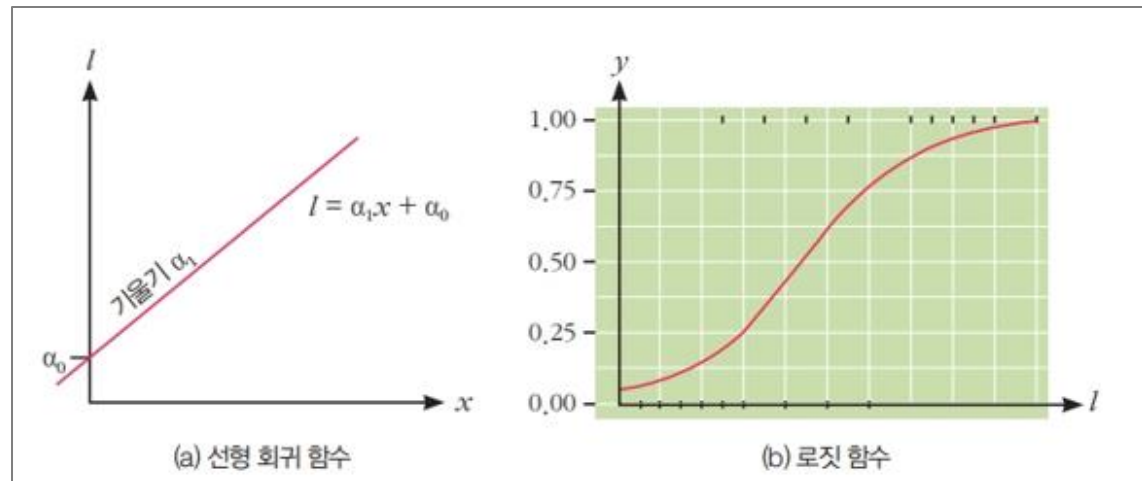
## 8.3 로지스틱 회귀

### ■ 원리

- 설명 변수를  $x$ , 반응 변수를  $l$ 로 표기 [그림 (a)]에서 가로축은  $x$ , 세로축은  $l$ 을 나타냄  

$$l = a_1 x + a_0 \quad \text{식(1)}$$
- 반응 변수  $l$ 의 범위는  $[-\infty, \infty]$ 이므로 로지스틱 회귀를 모델링할 수 없음
- 해결책: 로짓 함수(logit function)라 부르는 식 (2)를 추가로 사용  $\rightarrow$  범위를  $[0,1]$ 로 축소
- [그림(b)]에서 가로축은  $l$ , 세로축은  $y$ 를 나타내며  $y$ 는  $[0,1]$  사이로 축소되었음

$$y = \frac{1}{1+e^{-l}} \quad \text{식(2)}$$



선형 회귀 함수와 로짓 함수



## 8.3 로지스틱 회귀

### ■ 원리

- 여기서  $l$ 을 잠복(latent) 변수 또는 은닉(hidden) 변수라 부름

### ■ 일반화 선형 회귀는 두 단계 변환

- 일반화 선형 회귀에는 로지스틱 회귀뿐 아니라 지수 회귀, 포와송 회귀 등이 있음
- 로지스틱 회귀는 식 (1)과 식 (2)를 사용
- 식 (2)와 같은 함수를 링크 함수라 부름 (로지스틱 회귀는 링크 함수로 로짓 함수를 사용)

$$l = a_1 x + a_0 \quad \text{식(1)}$$

$$y = \frac{1}{1 + e^{-l}} \quad \text{식(2)}$$



# Thank you

