



# 10주차: 일반화 선형 모델

**ChulSoo Park**

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

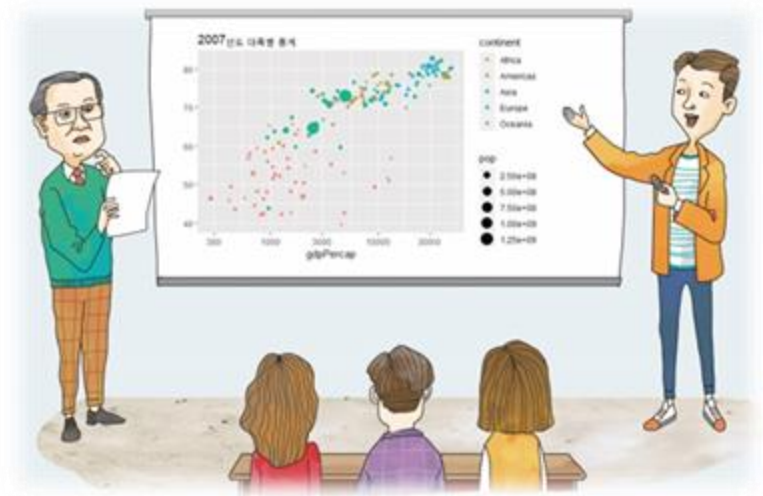
# 학습목표 (10주차)

- ❖ 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

## 8.4 로지스틱 회귀 : UCLA admission 데이터

The screenshot shows the UCLA Statistical Consulting website. The main navigation bar includes links for Prospective Students, Current Students, Faculty, Staff, Alumni, and Parents & Families. A search bar is present on the right. The page title is "UCLA Institute for Digital Research & Education Statistical Consulting". The main content area features a navigation menu with "HOME", "SOFTWARE", "RESOURCES", "SERVICES", and "ABOUT US". The "SOFTWARE" menu is open, showing options for "R", "Stata", "SAS", "SPSS", "Mplus", and "Other Packages". The "R" option is highlighted with a red box. On the right side, there is a section for "UPCOMING EVENTS" with three entries, each marked as "Remote consulting closed". The footer includes the copyright notice "© 2021 UC REGENTS" and links for "HOME" and "CONTACT".

데이터 과학의 가장 좋은 공부 방법은 비슷한 과정의 반복적 연습이다.

## 8.4 로지스틱 회귀 : UCLA admission 데이터

### ■ 로지스틱 회귀 실습 data 찾아보기

[HOME](#)
[SOFTWARE](#)
[RESOURCES](#)
[SERVICES](#)
[ABOUT US](#)

# UCLA

## DATA ANALYSIS EXAMPLES

### Examples

The pages packages. the output, description

**R**

The combi

- clients at w
- by our clie
- particular p
- our clients.

Stati

- 
- 
- 
- 

Impo

Models fo

- 
- 
-

<https://stats.idre.ucla.edu/r/>

Example 1. Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (response) variable is binary (0/1); win or lose. The predictor variables of interest are the amount of money spent on the campaign, the amount of time spent campaigning negatively and whether or not the candidate is an incumbent.

Example 2. A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

### Description of the data

For our data analysis below, we are going to expand on Example 2 about getting into graduate school. We have generated hypothetical data, which can be obtained from our website from within R. Note that *R requires forward slashes (/)* not back slashes () when specifying a file location even if the file is on your hard drive.

```
mydata <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
## view the first few rows of the data
head(mydata)
```

	admit	gre	gpa	rank
## 1	0	380	3.61	3
## 2	1	660	3.67	3
## 3	1	800	4.00	1
## 4	1	640	3.19	4
## 5	0	520	2.93	4
## 6	1	760	3.00	2

This dataset has a binary response (outcome, dependent) variable called **admit**. There are three predictor variables: **gre**, **gpa** and **rank**. We will treat the variables **gre** and **gpa** as continuous. The variable **rank** takes on the values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest. We can get basic descriptives for the entire data set by using **summary**. To get the standard deviations, we use **sapply** to apply the **sd** function to each variable in the dataset.

## 8.4 로지스틱 회귀 : UCLA admission 데이터

- UCLA admission이라는 데이터로 실습
  - 데이터를 읽고 확인하기

UCLA에 데이터 분석을 도와주는 IDRE(Institute for Digital Research and Education)가 있는데 IDRE가 제공하는 데이터 (대학원 입학 관련)

```
Console C:/RSources/
> ucla=read.csv('https://stats.idre.ucla.edu/stat/data/binary.csv')
> str(ucla)
'data.frame': 400 obs. of 4 variables:
 $ admit: int 0 1 1 1 0 1 1 0 1 0 ...
 $ gre  : int 380 660 800 640 520 760 560 400 540 700 ...
 $ gpa  : num 3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
 $ rank : int 3 3 1 4 4 2 1 2 3 2 ...
> head(ucla)
  admit gre  gpa rank
1     0 380 3.61    3
2     1 660 3.67    3
3     1 800 4.00    1
4     1 640 3.19    4
5     0 520 2.93    4
6     1 760 3.00    2
```

- admit : 불합격은 0, 합격은 1
- gre : 미국 대학원 수학능력시험인 gre의 점수
- gpa : 학부 성적(평균 학점)
- rank : 출신 대학 순위, {1, 2, 3, 4}의 4개 값

## 8.4 로지스틱 회귀 : UCLA admission 데이터

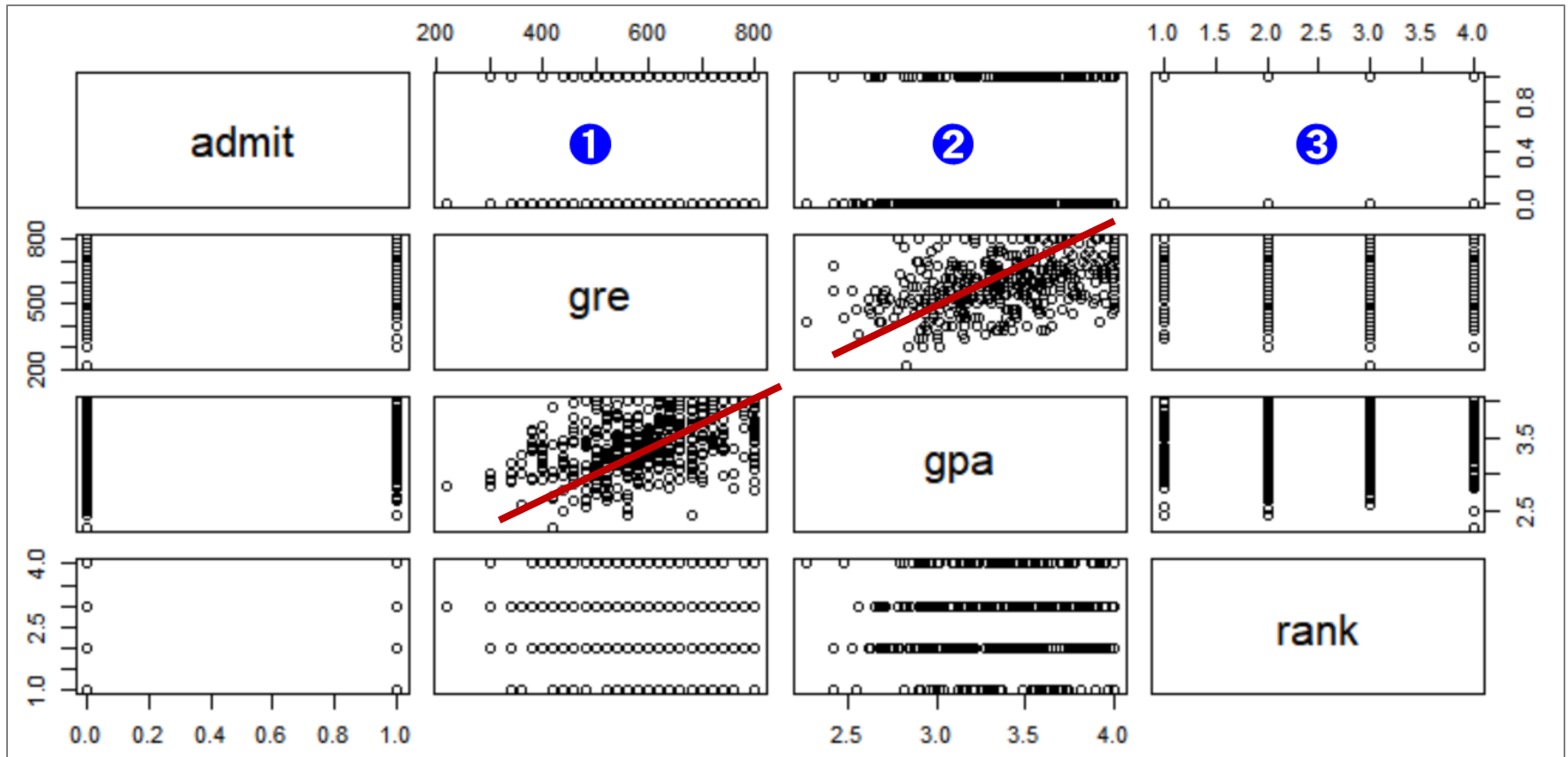
- UCLA admission이라는 데이터로 실습
  - 데이터를 읽고 확인하기
  - 4개의 변수 설명
    - ✓ admit : 0 : 불합격, 1: 합격
    - ✓ gre : 미국 대학원 수학능력 시험 점수
    - ✓ gpa : 학부 성적(평균 학점)
    - ✓ rank : 출신 대학 순위, {1, 2, 3, 4}의 4개 값(1 have the highest prestige)

```
> head(ucla)
  admit gre  gpa rank
1     0 380 3.61    3
2     1 660 3.67    3
3     1 800 4.00    1
4     1 640 3.19    4
5     0 520 2.93    4
6     1 760 3.00    2
```

## 8.4 로지스틱 회귀 : UCLA admission 데이터

### ■ plot(ucla) 실행

- ①은 gre(가로축)-admit(세로축), ②는 gpa-admit, ③은 rank-admit
- 400개 점이 겹쳐서 상관관계 분석이 어려움





## 8.4 로지스틱 회귀 : UCLA admission 데이터

- ggplot 라이브러리를 이용하여 gre-admit를 보다 정교하게 시각화
  - 점증적으로 생각하기 사례 (더 좋은 방향과 아이디어를 찾아가는 과정)

Console C:/RSources/ ↗

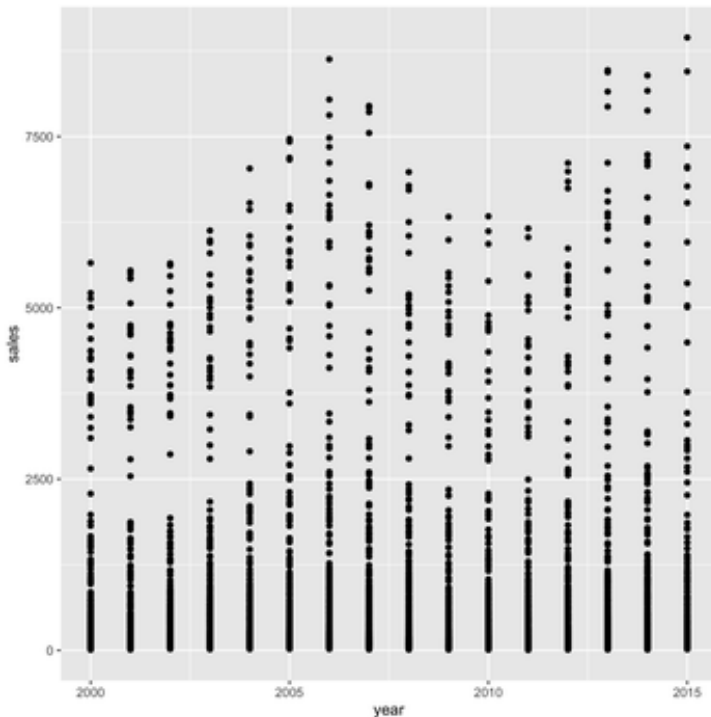
```
> library(dplyr)
> library(ggplot2)
① > ucla %>% ggplot(aes(gre, admit)) + geom_point()
>
② > ucla %>% ggplot(aes(gre, admit)) + geom_jitter()
>
③ > ucla %>% ggplot(aes(gre, admit)) + geom_jitter(aes(col=admit))
>
④ > ucla %>% ggplot(aes(gre, admit)) + geom_jitter(aes(col=factor(admit)))
>
⑤ > ucla %>% ggplot(aes(gre, admit)) + geom_jitter(aes(col=factor(admit)), height=0.1, width = 0.0)
```

## 8.4 로지스틱 회귀 : UCLA admission 데이터

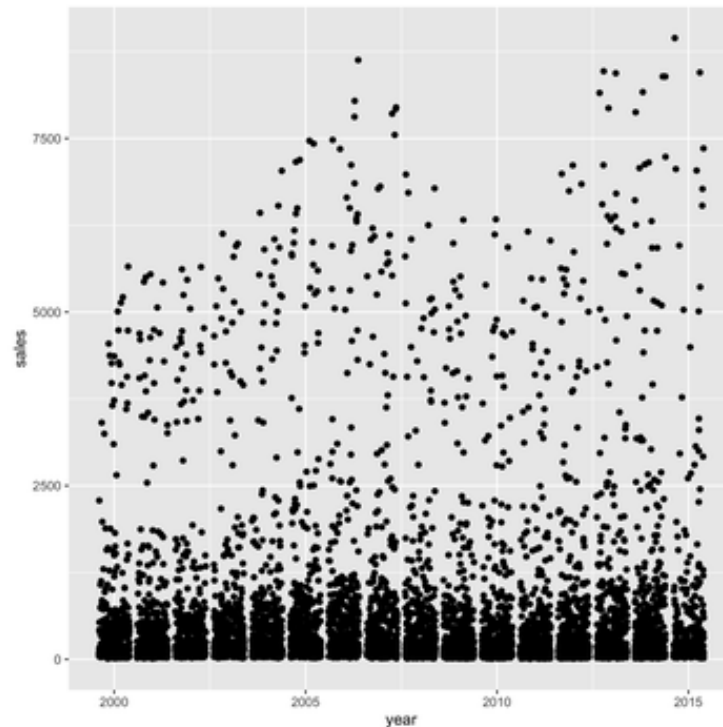
연도별로 세일즈 데이터의 분포를 확인해보려고 산점도를 그렸으나 정확하게 데이터가 몰려있는 것을 확인하기가 어렵다. 숫자형 데이터처럼 보이지만, 이렇게 이산형 데이터 형태의 경우 `geom_point()` 함수는 적합하지가 않다.

`geom_jitter()` : 그래프를 통해서는 데이터의 분포가 연도별로 어떻게 다른지 조금 더 확실하게 알 수 있다.

`geom_point()`



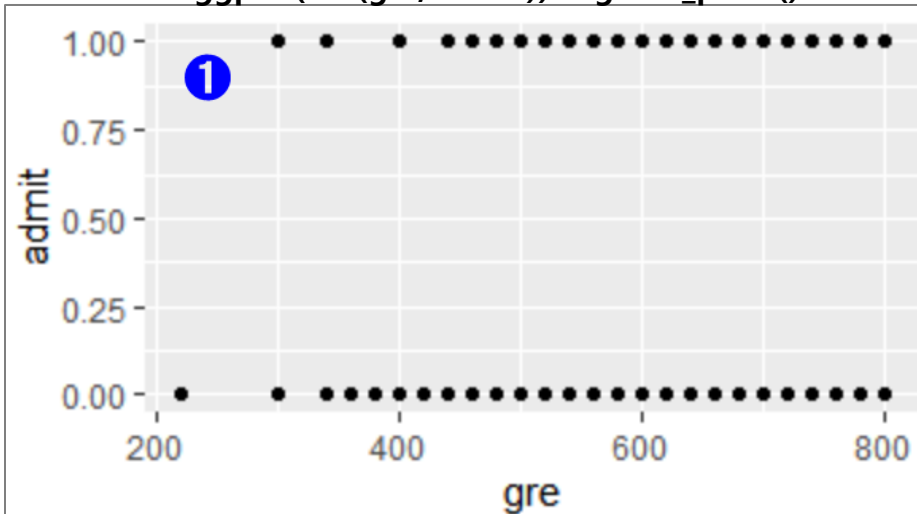
`geom_jitter()`



## 8.4 로지스틱 회귀 : UCLA admission 데이터

■ ggplot 라이브러리를 이용하여 gre-admit를 보다 정교하게 시각화

`ucla %>% ggplot(aes(gre, admit)) + geom_point()`

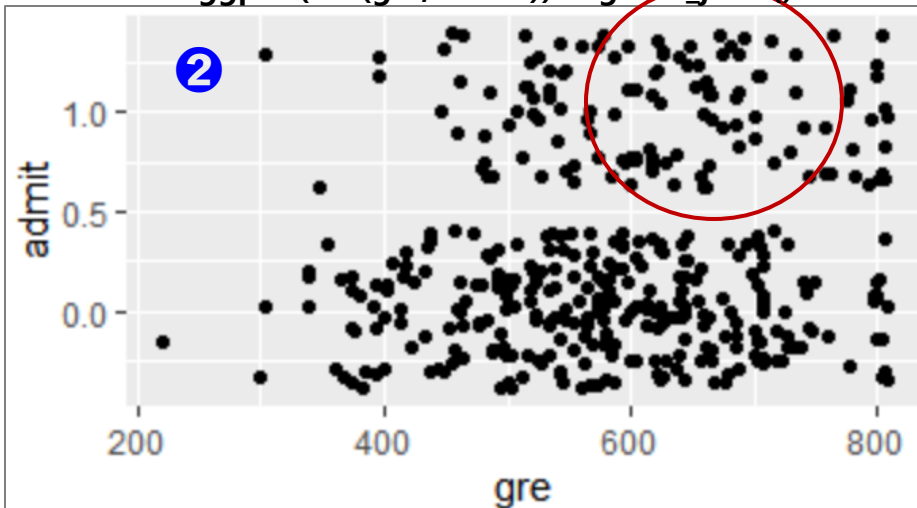


① 의미 분석에 별다른 좋아진 점이 없어 보임

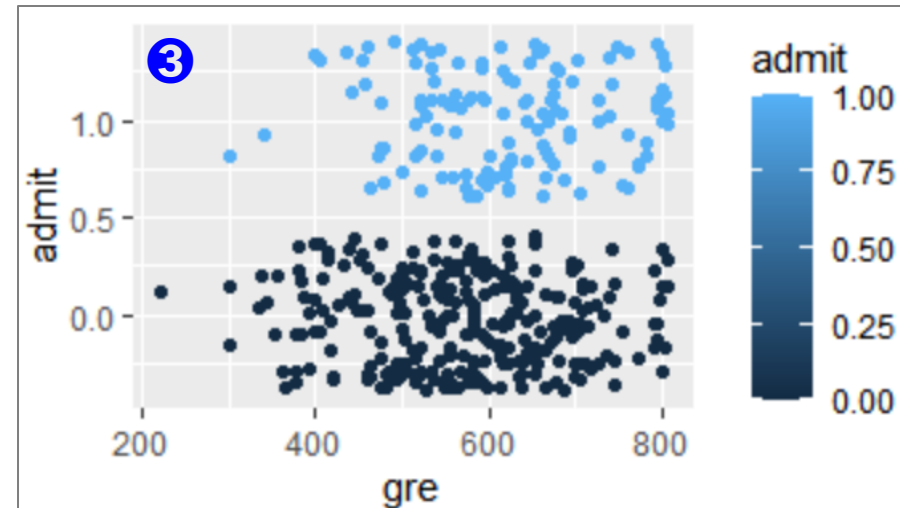
② 합격(상단)의 gre 점수가 높게 분포

③ 색상이 연속적인 것으로 보임

`ucla %>% ggplot(aes(gre, admit)) + geom_jitter()`



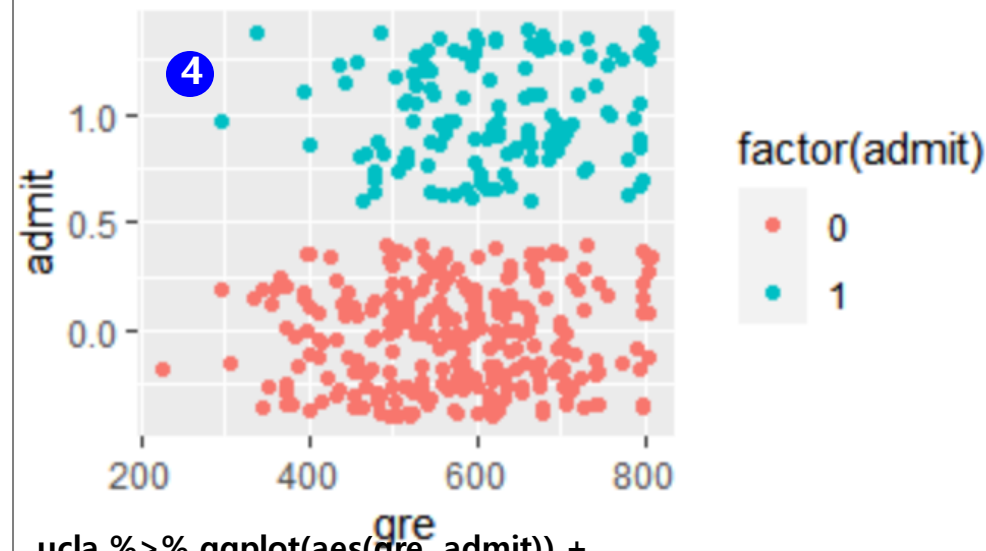
`ucla %>% ggplot(aes(gre, admit)) + geom_jitter(aes(col=admit))`



## 8.4 로지스틱 회귀 : UCLA admission 데이터

- ④ 색상을 범주형으로 나타냄
- ⑤ 잡음을 기본 40% → 10%
- ⑥ 잡음을 기본 40% → 20%

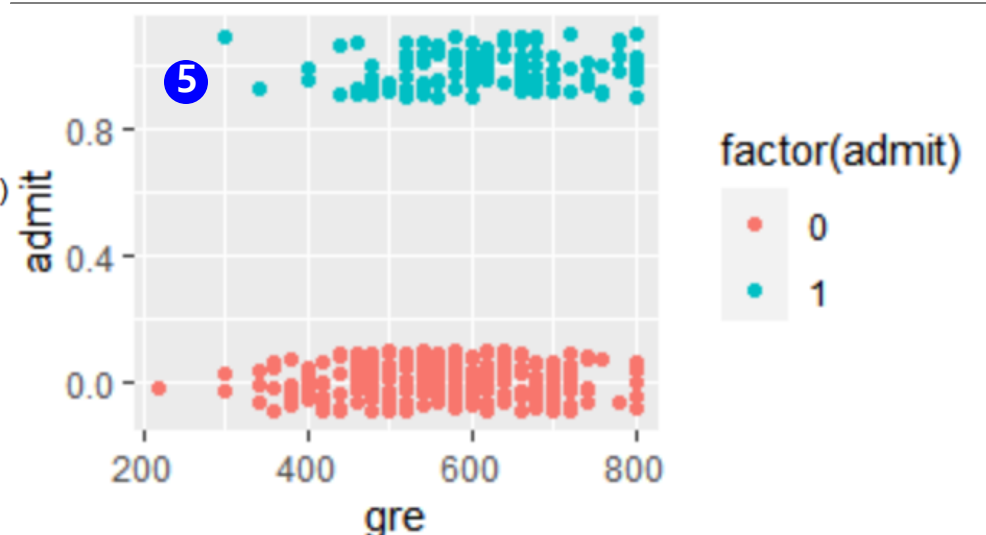
```
ucla %>% ggplot(aes(gre, admit)) + geom_jitter(aes(col=factor(admit)))
```



```
ucla %>% ggplot(aes(gre, admit)) +  
geom_jitter(aes(col=factor(admit)), height=0.2, width = 0.0)
```



```
ucla %>% ggplot(aes(gre, admit)) +  
geom_jitter(aes(col=factor(admit)), height=0.1, width = 0.0)
```

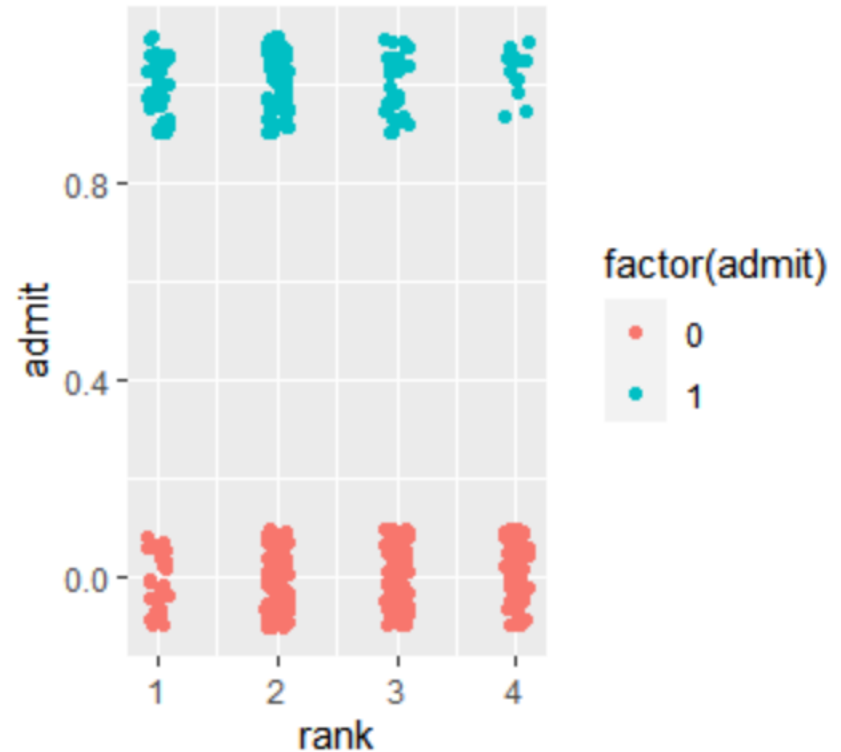
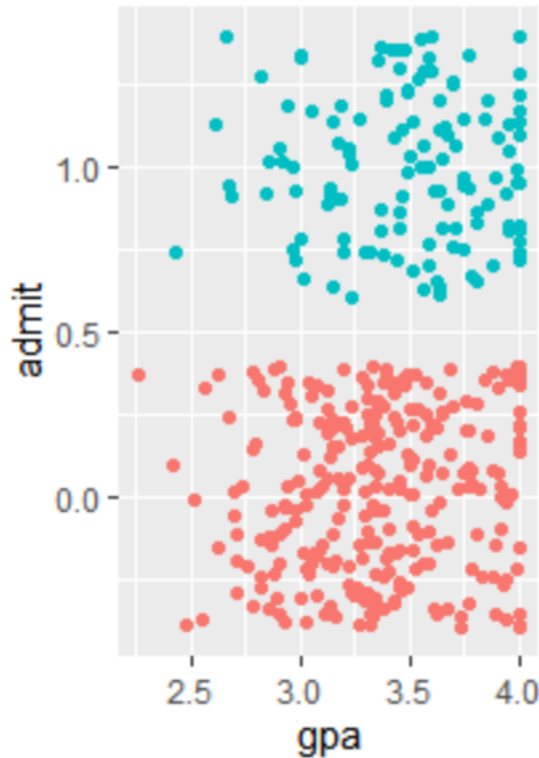


## 8.4 로지스틱 회귀 : UCLA admission 데이터

■ gpa-admit와 rank-admit를 정교하게 시각화

Console C:/RSources/

```
> library(gridExtra)
> p1 = ucla%>%ggplot(aes(gpa, admit)) + geom_jitter(aes(col = factor(admit)), height = 0.1, width = 0.0)
> p1 = ucla%>%ggplot(aes(gpa, admit)) + geom_jitter(aes(col = factor(admit)))
> grid.arrange(p1, p2, ncol = 2)
```



## 8.4 로지스틱 회귀 : UCLA admission 데이터

### ■ glm 적용하기

```
Console C:/RSources/ ↗
> m = glm(admit~., data = ucla, family = binomial)
> coef(m)
(Intercept)          gre          gpa          rank
-3.44954840  0.00229396  0.77701357 -0.56003139
> deviance(m,type='response')
[1] 459.4418
> summary(ucla)
      admit          gre          gpa          rank
Min.   :0.0000  Min.   :220.0  Min.   :2.260  Min.   :1.000
1st Qu.:0.0000  1st Qu.:520.0  1st Qu.:3.130  1st Qu.:2.000
Median :0.0000  Median :580.0  Median :3.395  Median :2.000
Mean   :0.3175  Mean   :587.7  Mean   :3.390  Mean   :2.485
3rd Qu.:1.0000  3rd Qu.:660.0  3rd Qu.:3.670  3rd Qu.:3.000
Max.   :1.0000  Max.   :800.0  Max.   :4.000  Max.   :4.000
```

### ■ 모델 살펴보기

- gre 계수가 gpa 계수보다 훨씬 작은 이유: gre 값의 범위가 훨씬 크기 때문
- rank 계수가 음수인 이유: 값이 작을수록 좋은 대학이기 때문

```
> head(ucla)
  admit gre  gpa rank
1     0 380 3.61   3
2     1 660 3.67   3
3     1 800 4.00   1
4     1 640 3.19   4
5     0 520 2.93   4
6     1 760 3.00   2
```

## 8.4 로지스틱 회귀 : UCLA admission 데이터

### ■ 예측


- predict 함수로 수행
- 예를 들어, gre=376, gpa=3.6, rank=3인 새로운 학생이 발생하면,
  - 합격 확률은 18.7%

Console C:/RSources/ ↗

```
> s = data.frame(gre = c(376), gpa = c(3.6), rank = c(3))  
> predict(m, newdata = s, type = 'response')  
1  
0.1869631  
>
```

## 8.5 로지스틱 회귀의 적용

- survival 라이브러리가 제공하는 colon cancer(결장암) 데이터
  - 결측치가 존재하는 일반적인 colon data 로지스틱 회귀 적용 실습

```
Console C:/RSources/   
> library(survival)  
> str(colon)  
'data.frame': 1858 obs. of 16 variables:  
 $ id      : num  1 1 2 2 3 3 4 4 5 5 ...  
 $ study   : num  1 1 1 1 1 1 1 1 1 1 ...  
 $ rx      : Factor w/ 3 levels "Obs","Lev","Lev+5FU": 3 3 3 3 1 1 3 3 1 1 ...  
 $ sex     : num  1 1 1 1 0 0 0 0 1 1 ...  
 $ age     : num  43 43 63 63 71 71 66 66 69 69 ...  
 $ obstruct: num  0 0 0 0 0 0 1 1 0 0 ...  
 $ perfor  : num  0 0 0 0 0 0 0 0 0 0 ...  
 $ adhere  : num  0 0 0 0 1 1 0 0 0 0 ...  
 $ nodes   : num  5 5 1 1 7 7 6 6 22 22 ...  
 $ status  : num  1 1 0 0 1 1 1 1 1 1 ...  
 $ differ  : num  2 2 2 2 2 2 2 2 2 2 ...  
 $ extent  : num  3 3 3 3 2 2 3 3 3 3 ...  
 $ surg    : num  0 0 0 0 0 0 1 1 1 1 ...  
 $ node4   : num  1 1 0 0 1 1 1 1 1 1 ...  
 $ time    : num  1521 968 3087 3087 963 ...  
 $ etype   : num  2 1 2 1 2 1 2 1 2 1 ...
```



## 8.5 로지스틱 회귀의 적용

## 8.5 로지스틱 회귀의 적용

- data 정제 작업(이상치, 결측 값 등 확인)

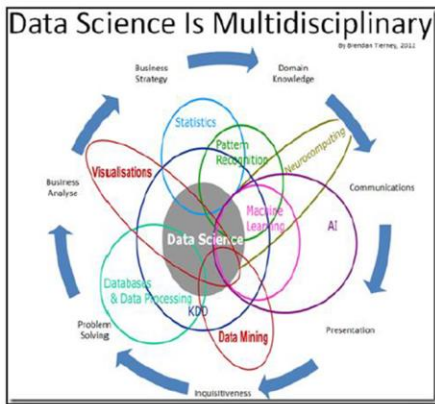
```
write.csv(colon, file="c:/rdata/08_colon.csv",quote=F)
```

1	id	study	rx	sex	age	obstruct	perfor	adhere	nodes	status	differ	extent	surg	node4	time	etype
451	254	1	Obs	0	62	0	0	0	5	1	3	3	0	1	221	1
452	482	1	Lev+5FU	0	58	1	0	0	5	1	3	3	0	1	79	2
453	482	1	Lev+5FU	0	58	1	0	0	5	1	3	3	0	1	40	1
454	584	1	Lev+5FU	1	55	1	0	0	5	1	3	3	0	1	34	2
455	584	1	Lev+5FU	1	55	1	0	0	5	1	3	3	0	1	9	1
456	816	1	Obs	1	62	0	0	0	5	1	3	3	0	1	587	2
457	816	1	Obs	1	62	0	0	0	5	1	3	3	0	1	286	1
458	917	1	Obs	0	71	0	0	0	5	1	3	3	0	1	259	2
459	917	1	Obs	0	71	0	0	0	5	1	3	3	0	1	122	1
460	89	1	Lev+5FU	0	80	0	0	0	4	0	1	3	0	0	2724	2
461	89	1	Lev+5FU	0	80	0	0	0	4	0	1	3	0	0	2724	1
462	170	1	Lev+5FU	0	48	0	0	0	4	0	1	3	0	0	2631	2
463	170	1	Lev+5FU	0	48	0	0	0	4	0	1	3	0	0	2631	1
464	207	1	Lev+5FU	0	53	0	0	0	4	0	1	1	0	0	2835	2
465	207	1	Lev+5FU	0	53	0	0	0	4	1	1	1	0	0	1037	1
466	414	1	Obs	1	59	1	0	0	4	1	1	3	0	0	537	2
467	414	1	Obs	1	59	1	0	0	4	1	1	3	0	0	238	1
468	451	1	Lev	0	55	1	0	0	4	0	1	3	0	0	2544	2

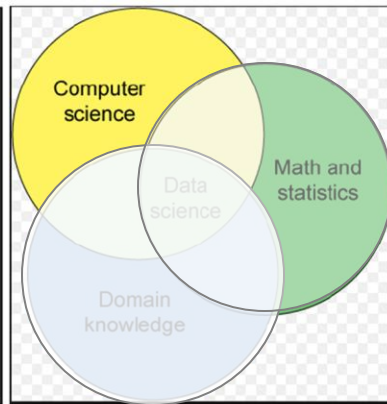
## 8.5 로지스틱 회귀의 적용

### ■ colon cancer 데이터 살펴보기 : Domain knowledge

- 결측치가 존재하는 일반적인 colon cancer data 로지스틱 회귀 적용 실습



출처 : www.oralalytics.com



데이터 사이언스 학과(대학원)

<p><b>T(Tumor, 종양)인자</b></p> <p>T0 : 종양의 근거가 없음</p> <p>T1 : 점막층과 점막하층에 국한된 대장암</p> <p>T2 : 고유근층까지 침습한 대장암</p> <p>T3 : 장막층을 침습한 결장암 또는 직장간막층을 침습한</p> <p>T4 : 인접한 다른 장기까지 침습한 대장암</p>	<p><b>Dukes (Astler-Coller 개정) 분류법</b></p> <p>A기 : 점막에 국한된 대장암</p> <p>B기</p> <p>B1기 : 고유근층까지 침습한 대장암</p> <p>B2기 : 대장암이 장막층을 뚫고 나간 상태</p> <p>B3기 : 대장암이 인접한 장기에 유착되거나 침습한 상태</p>
<p><b>N(Node, 림프절)인자</b></p> <p>N0 : 림프절 전이 없음</p> <p>N1 : 1~3개의 국소 림프절 전이</p> <p>N2 : 4개 이상의 국소 림프절 전이</p> <p>N3 : 비 전형적인 림프절 또는 큰 혈관 주위의 림프절 전이</p>	<p><b>C기</b></p> <p>C1기 : B1 + 국소 림프절 전이</p> <p>C2기 : B2 + 국소 림프절 전이</p> <p>C3기 : B3 + 국소 림프절 전이</p>
<p><b>M(Metastasis, 원격전이)인자</b></p> <p>M0 : 원격전이 없음</p> <p>M1 : 원격전이 있음</p>	<p>D기 : 원격전이</p>

- 대장암(colorectal cancer) = 결장암(colon cancer) + 직장암(rectal cancer)
- 대장암 진행 정도 표기법
  - ✓ AJCC(American Joint Committee on Cancer) : A,B,C,D로 표현
  - ✓ 국제표준 TNM 분류법 : 로마자로 I, II, III,IV로 표현
- 등급 판정 기준 : T인자, N인자,M인자 종합
- 5년 생존율 : A기(90%), B기(60%), C기(40%), D기(5% 미만)
- 최근 수술후 생존률을 유의하게 높일 수 있는 5-FU 효과 인정

출처 : 현대의학, 자연과학 그리고 의용공학의 세계 (<https://blog.daum.net/inbio880/16096552>)

## 8.5 로지스틱 회귀의 적용

### ■ colon cancer 데이터 살펴보기 : Description

- 결측치가 존재하는 일반적인 colon cancer data 로지스틱 회귀 적용 실습

R: Chemotherapy for Stage B/C colon cancer ▾
Find in Topic

colon {survival}
R Documentation

# Chemotherapy for Stage B/C colon cancer

B / C 기 결장암에 대한 화학 요법 기술

**Description**

These are data from chemotherapy for cc compound previousl FU is a moderately t There are two recor death

이들은 대장 암에 대한 보조 화학 요법의 첫 번째 성공적인 시험 중 하나에서 얻은 데이터입니다. Levamisole은 이전에 동물의 벌레 감염을 치료하는 데 사용 된 저독성 화합물입니다. 5-FU는 중등도의 독성 (이런 일들이 진행됨에 따라) 화학 요법 제입니다. 한 사람당 두 개의 기록이 있습니다. 하나는 재발에 대한 것이고 다른 하나는 죽음에 대한 것입니다

## 8.5 로지스틱 회귀의 적용

### ■ colon cancer 데이터 살펴보기 : Description of 16개 변수

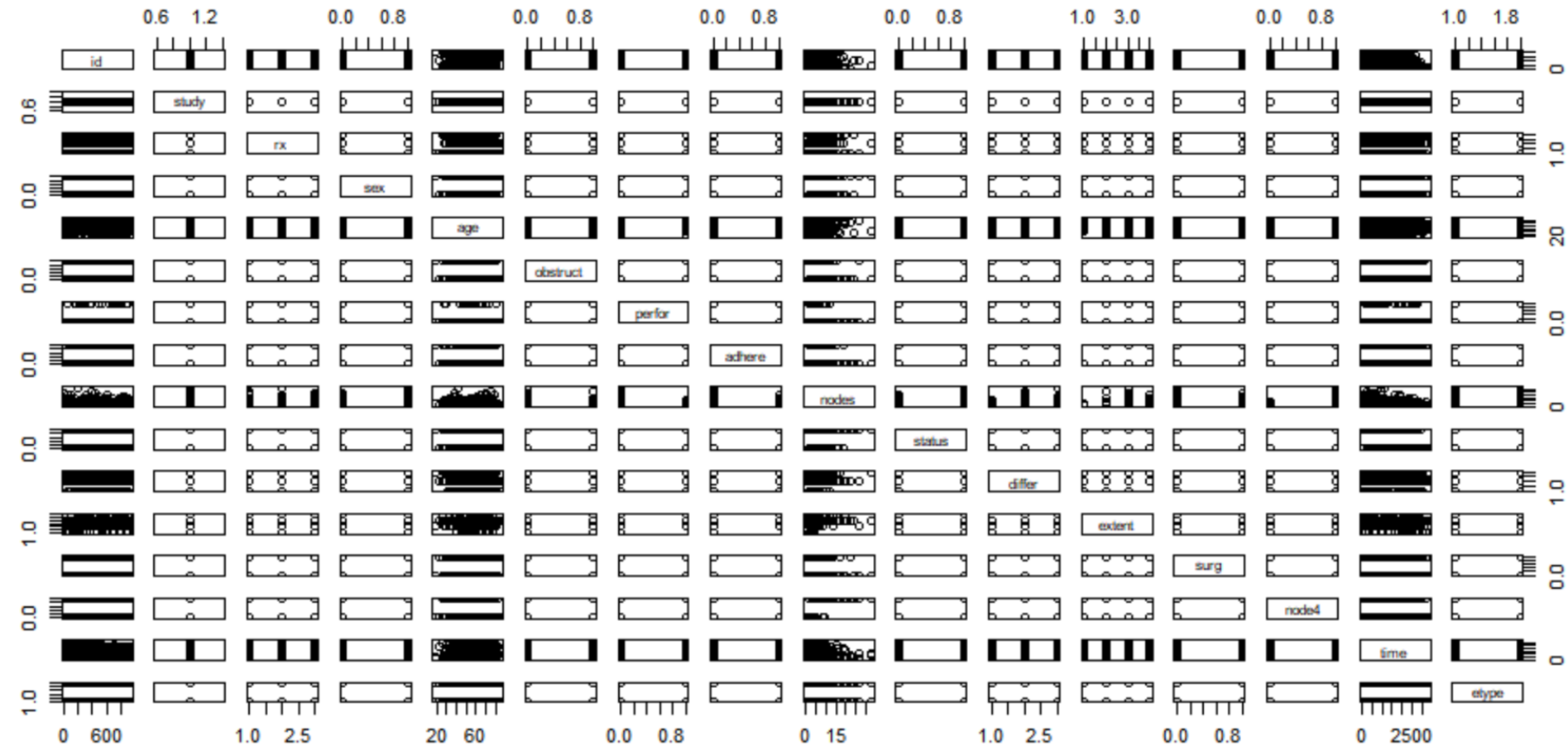
- id : 환자 번호
- study : 모든 샘플이 1(모두 조사에 참여)
- rx : 치료 방법(Observation, Levamisole, Levamisole+5-FU)
- sex : 성별(여성 : 0, 남성 : 1)
- age : 나이
- obstruct : 결장의 폐쇄 여부(폐쇄 안 됨 : 0, 폐쇄 : 1)
- perfor : 결장의 구멍 여부(구멍 없음 : 0, 구멍 있음 : 1)
- adhere : 인접 장기와 붙었는지 여부(붙지 않음 : 0, 붙음 : 1)
- nodes : 암세포가 있는 림프절의 수
- **status : 재발/사망 여부 (완치 : 0, 재발 또는 사망 :1)**
- differ : 암세포의 조직학적 분화 정도(well : 1, moderate : 2, poor :3)
- extent : 암세포가 침습한 깊이(submucosa:1, muscle:2, serosa: , 인접 장기:4)
- surg : 수술 후 등록기까지의 기간 (short : 0, long :1)
- node4 : 양성 림프절 수가 4개 이상인지 여부( 4개 미만 :0, 4개 이상 :1)
- time : etype까지의 일수
- etype : 재발 또는 사망(재발 : 1, 사망 :2)

반응 변수

## 8.5 로지스틱 회귀의 적용

### ■ 시각화 >plot(colon)

분석 대상이 되는 변수가 너무 많아 상관관계 파악 어려움



## 8.5 로지스틱 회귀의 적용

### ■ ggplot으로 상세하게 시각화

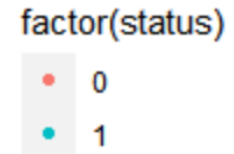
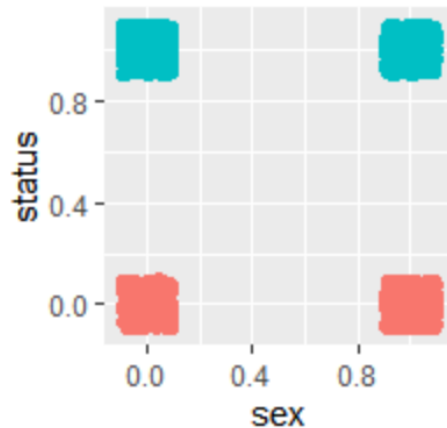
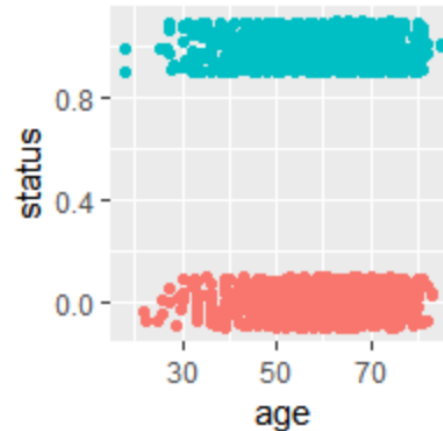
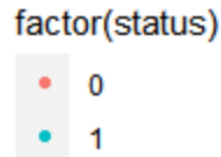
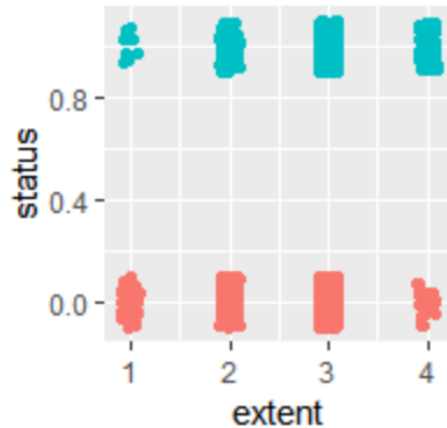
- extent-status, age-status, sex-status, nodes-status 상관관계
- node4-status, differ-status의 상관관계

```
> p1 = colon %>% ggplot(aes(extent, status)) + geom_jitter(aes(col =  
  factor(status)), height = 0.1, width = 0.1)  
> p2 = colon %>% ggplot(aes(age, status)) + geom_jitter(aes(col =  
  factor(status)), height = 0.1, width = 0.1)  
> p3 = colon %>% ggplot(aes(sex, status)) + geom_jitter(aes(col =  
  factor(status)), height = 0.1, width = 0.1)  
> p4 = colon %>% ggplot(aes(nodes, status)) + geom_jitter(aes(col =  
  factor(status)), height = 0.1, width = 0.1)  
> grid.arrange(p1,p2,p3,p4, ncol=2,nrow=2)  
  
> p5 = colon %>% ggplot(aes(node4, status)) + geom_jitter(aes(col =  
  factor(status)), height = 0.1, width = 0.1)  
> p6 = colon %>% ggplot(aes(differ, status)) + geom_jitter(aes(col =  
  factor(status)), height = 0.1, width = 0.1)  
> grid.arrange(p5,p6, ncol=2,nrow=1)
```

## 8.5 로지스틱 회귀의 적용

### ■ 시각화 결과 해석

- extent(침습의 깊이)가 클수록 1(재발 또는 사망)인 샘플이 많음
- age가 클수록 1인 샘플이 많지만 age가 적은 샘플도 1인 경우 적지 않음 → 결장암에 걸리면 나이가 적더라도 재발 또는 사망 확률이 크다는 사실을 알 수 있음

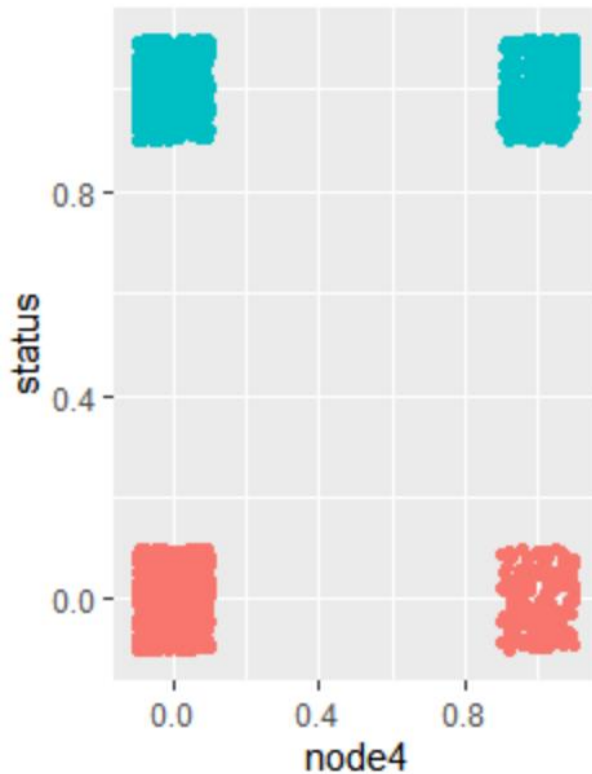




## 8.5 로지스틱 회귀의 적용

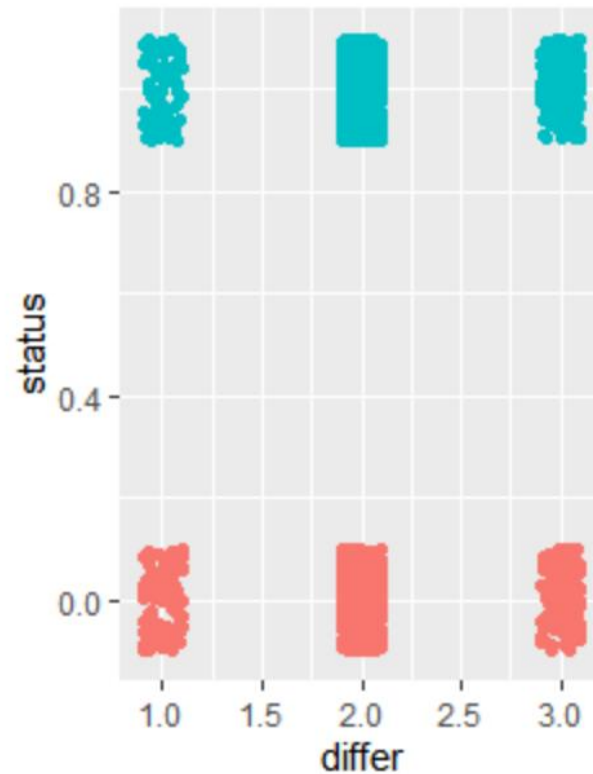
### ■ 시각화 결과 해석

- node4의 값이 1일때 1(재발 또는 사망)인 샘플이 많음
- 암세포 분화 정도는 별차이를 알 수 없음



factor(status)

- 0
- 1



factor(status)

- 0
- 1

## 8.5 로지스틱 회귀의 적용

### ■ 결측값 확인 및 제거

```

Console C:/RSources/
> table(is.na(colon))

FALSE TRUE
29646   82
> cl_colon=na.omit(colon)
> str(colon)
'data.frame':   1858 obs. of  16 variables:
 $ id      : num  1 1 2 2 3 3 4 4 5 5 ...
 $ study   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ rx      : Factor w/ 3 levels "Obs","Lev","Lev+5FU": 3 3 3 3 1 1 3 3 :
 $ sex     : num  1 1 1 1 0 0 0 0 1 1 ...
 $ age     : num  43 43 63 63 71 71 66 66 69 69 ...
 $ obstruct: num  0 0 0 0 0 0 1 1 0 0 ...
 $ perfor  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ adhere  : num  0 0 0 0 1 1 0 0 0 0 ...
 $ nodes   : num  5 5 1 1 7 7 6 6 22 22 ...
 $ status  : num  1 1 0 0 1 1 1 1 1 1 ...
 $ differ  : num  2 2 2 2 2 2 2 2 2 2 ...
 $ extent  : num  3 3 3 3 2 2 3 3 3 3 ...
 $ surg    : num  0 0 0 0 0 0 1 1 1 1 ...
 $ node4   : num  1 1 0 0 1 1 1 1 1 1 ...
 $ time    : num  1521 968 3087 3087 963 ...
 $ etype   : num  2 1 2 1 2 1 2 1 2 1 ...
> str(cl_colon)
'data.frame':   1776 obs. of  16 variables:
 $ id      : num  1 1 2 2 3 3 4 4 5 5 ...
 $ study   : num  1 1 1 1 1 1 1 1 1 1 ...
 $ rx      : Factor w/ 3 levels "Obs","Lev","Lev+5FU": 3 3 3 3 1 1 3 3 :
 $ sex     : num  1 1 1 1 0 0 0 0 1 1 ...
 $ age     : num  43 43 63 63 71 71 66 66 69 69 ...

```

82개 row data 제거

## 8.5 로지스틱 회귀의 적용

### ■ summary(cl\_colon)

- 결장암은 성에는 별 상관 없음 (sex mean : 0.518)
- 치료 방법 3가지 고루 적용
- 결장 폐쇄 : 19.26 %
- 결장 구멍 : 3.04%

```
> summary(cl_colon)
```

id		study	rx	sex	age	obstruct
Min. : 1.0	Min. :1	obs	:610	Min. :0.000	Min. :18.00	Min. :0.0000
1st Qu.:234.8	1st Qu.:1	Lev	:588	1st Qu.:0.000	1st Qu.:53.00	1st Qu.:0.0000
Median :466.5	Median :1	Lev+5FU	:578	Median :1.000	Median :61.00	Median :0.0000
Mean :466.5	Mean :1			Mean :0.518	Mean :59.81	Mean :0.1926
3rd Qu.:700.2	3rd Qu.:1			3rd Qu.:1.000	3rd Qu.:69.00	3rd Qu.:0.0000
Max. :929.0	Max. :1			Max. :1.000	Max. :85.00	Max. :1.0000

perfor		adhere	nodes	status	differ
Min. :0.00000	Min. :0.0000	Min. : 0.000	Min. :0.0000	Min. :0.0000	Min. :1.000
1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.: 1.000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:2.000
Median :0.00000	Median :0.0000	Median : 2.000	Median :0.0000	Median :0.0000	Median :2.000
Mean :0.03041	Mean :0.1441	Mean : 3.663	Mean :0.4932	Mean :2.062	
3rd Qu.:0.00000	3rd Qu.:0.0000	3rd Qu.: 5.000	3rd Qu.:1.0000	3rd Qu.:2.000	
Max. :1.00000	Max. :1.0000	Max. :33.000	Max. :1.0000	Max. :3.000	

extent		surg	node4	time	etype
Min. :1.000	Min. :0.000	Min. :0.0000	Min. : 8	Min. :1.0	
1st Qu.:3.000	1st Qu.:0.000	1st Qu.:0.0000	1st Qu.: 573	1st Qu.:1.0	
Median :3.000	Median :0.000	Median :0.0000	Median :1856	Median :1.5	
Mean :2.884	Mean :0.268	Mean :0.2646	Mean :1543	Mean :1.5	
3rd Qu.:3.000	3rd Qu.:1.000	3rd Qu.:1.0000	3rd Qu.:2331	3rd Qu.:2.0	
Max. :4.000	Max. :1.000	Max. :1.0000	Max. :3329	Max. :2.0	

## 8.5 로지스틱 회귀의 적용

- status를 반응 변수로 하고 glm 적용하면

```

Console C:/RSources/
> m = glm(status~., data = cl_colon, family = binomial)
> m

call: glm(formula = status ~ ., family = binomial, data = cl_colon)

Coefficients:
(Intercept)      id      study      rxLev      rxLev+5FU      sex      age
 8.112834    -0.003895      NA    0.091332    -0.439472    0.040958   -0.009759
obstruct      perfor      adhere      nodes      differ      extent      surg
-0.012119    0.352175    0.355135    0.101352   -0.463883    0.047188    0.433553
node4      time      etype
-0.270756   -0.004346    1.371260

Degrees of Freedom: 1775 Total (i.e. Null), 1760 Residual
Null Deviance:      2462
Residual Deviance: 666.3      AIC: 698.3
  
```

모두 "1"로 모델 수립에  
아무런 영향을 미치지  
못함 따라서 제거 시킴

시간은 별 의미 없음

glm의 na.action 옵션의 기본값이 결측치를 자동으로 제거 시킴

etype도 2가지로 1가지 제외 시킴

(data 확인하는 이유?)




## 8.5 로지스틱 회귀의 적용

### ■ 모델 해석

- study 변수의 계수가 NA : 변수 설명을 보면 study 변수는 모든 샘플이 1 값을 가짐. 반응 변수 status에 아무 영향을 미치지 못함. 즉 study 변수는 분별력이 전혀 없음
- id : 영향이 없는 data
- time : 영향이 없는 data
- etype : 동일 환장에 대해 샘플 1, 샘플 2가 있음 1가지만 선택하고 etype 변수 삭제
- 전체 16개 변수 중 status는 반응변수, 4개는 삭제, 11개 변수를 설명 변수로 사용

## 8.5 로지스틱 회귀의 적용

- 영향이 없는 데이터 제외 후 glm 재 작업
  - 같은 환자에 대해 etype이 2인 샘플과 1인 샘플 두개 존재 → 홀수 번째만 취하고 etype과 time 변수 제외. id는 환자 번호에 불과하므로 제외
  - 15개 설명 변수 중에서 study, time, etype, id를 제외(특징 선택)하고 glm 적용하면

```
Console C:/RSources/   
```

```
> cl_colon = cl_colon[c(TRUE, FALSE), ] # 동일 환자 홀 번째 data 만 사용
> m= glm(status~rx + sex + age + obstruct + perfor + adhere + nodes + differ + extent + surg
+ node4, data = cl_colon, family = binomial)
> m
```

Call: glm(formula = status ~ rx + sex + age + obstruct + perfor + adhere + nodes + differ + extent + surg + node4, family = binomial, data = cl\_colon)

Coefficients:

(Intercept)	rxLev	rxLev+5FU	sex	age	obstruct
-2.983344	-0.117773	-0.501094	-0.003632	0.010041	0.295354
perfor	adhere	nodes	differ	extent	surg
0.002631	0.364115	0.124319	0.030283	0.569443	0.385432
node4					
0.627999					

Degrees of Freedom: 887 Total (i.e. Null); 875 Residual  
Null Deviance: 1230  
Residual Deviance: 1111 AIC: 1137

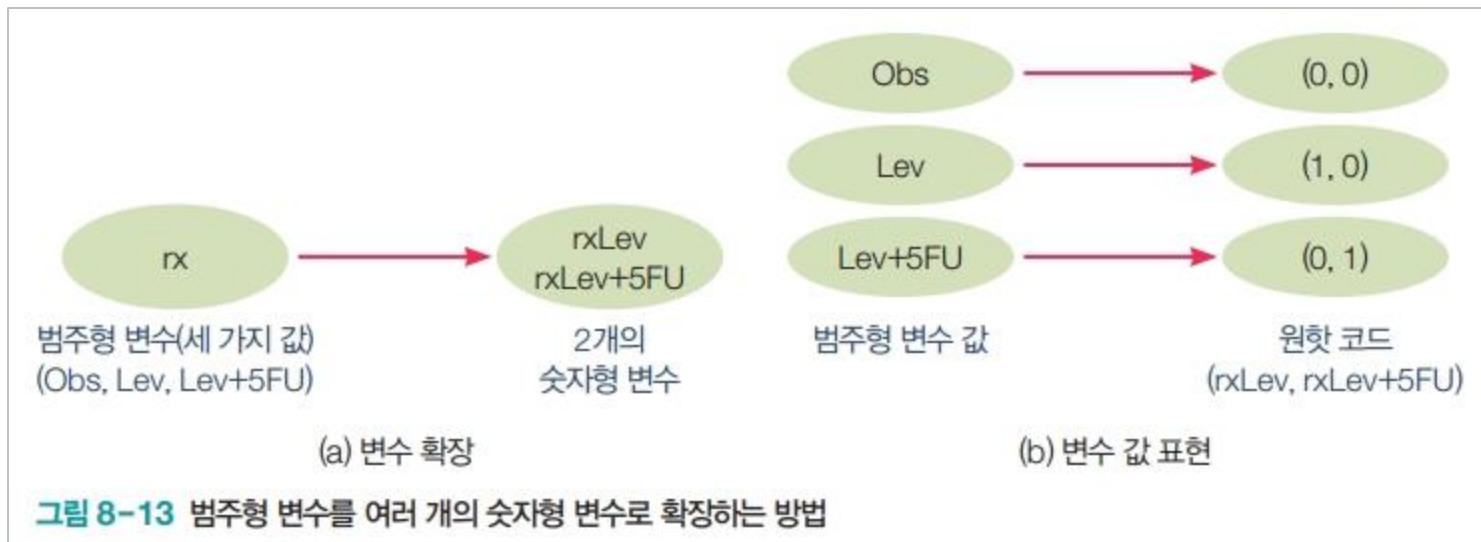
## 8.5 로지스틱 회귀의 적용

### ■ 범주형 (factor 형) 변수

- 범주형은 순서값(ordinal value)과 명칭값(nominal value)으로 구분
- 순서값은 거리 개념이 있음 → 숫자를 부여하면 모델링에 그대로 참여 가능
  - ✓ A, B, C, D, F의 학점 (1, 2, 3, 4, 5로 표현하면 거리 계산 가능)
  - ✓ UCLA admission 데이터의 rank 변수도 순서 값
- 명칭값은 거리 개념이 없음 → 정수를 부여해도 거리 개념이 없어 그대로 모델링에 참여 불가능
  - ✓ {A, B, O, AB}의 혈액형
  - ✓ {전북, 전남, 경북, 경남,...}의 지역

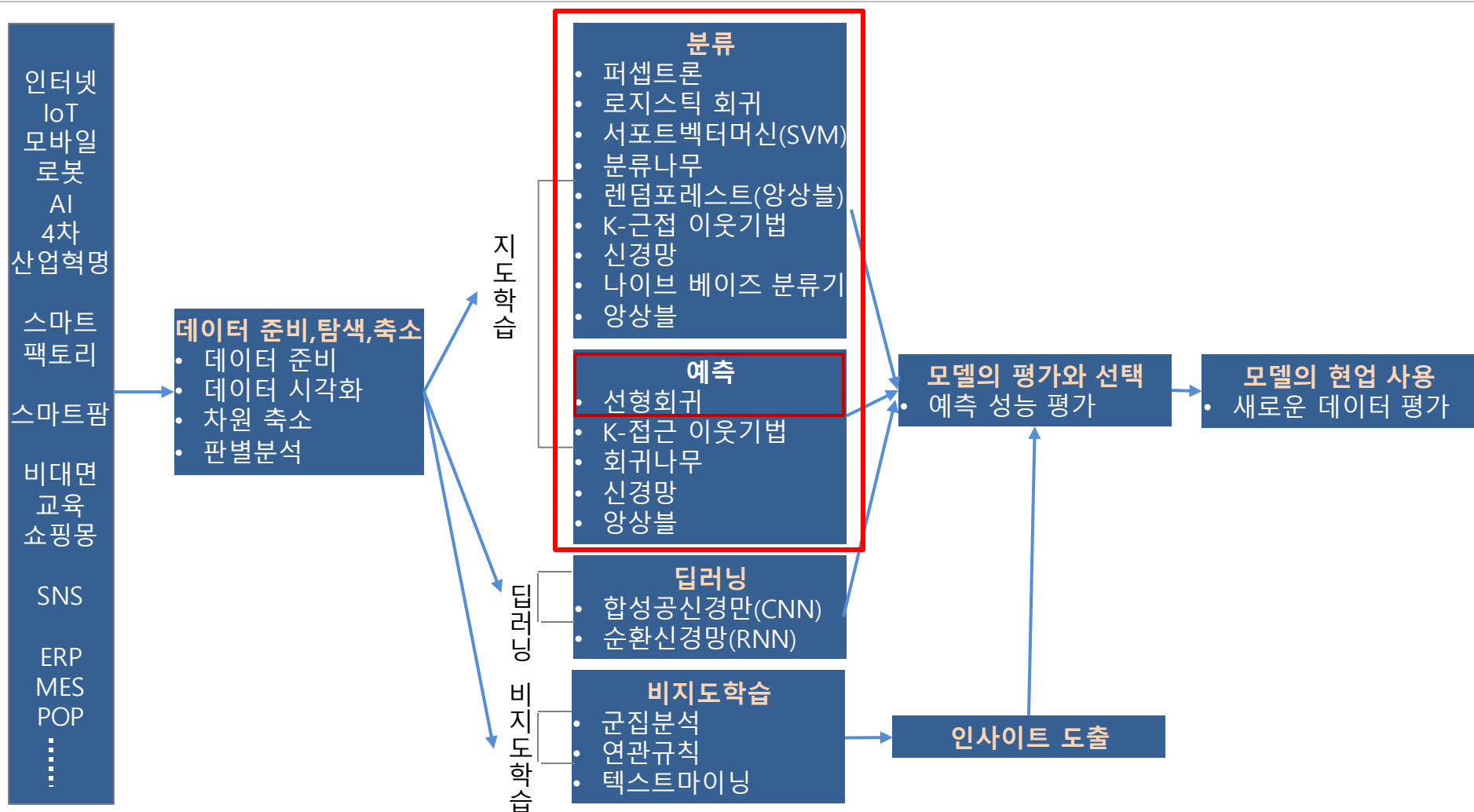
## 8.5 로지스틱 회귀의 적용

- colon 데이터의 범주형 변수
  - rx : 치료 방법(Observation, Levamisole, Levamisole+5-FU)
  - rx 변수는 범주형(factor)에 해당
  - 모델을 잘 살펴보면, rx가 rxLev와 rxLev+5FU라는 두 개의 변수로 바뀌었음
    - rx는 Obs, Lev, Lev+5FU의 세 가지 값을 가지는데, rx에 두 번째와 세 번째 값을 붙여서 rxLev와 rxLev+5FU를 만들
  - 원핫 코드(one-hot code)를 사용하여 값을 표현함





## 요약



# Thank you

