Part 08

일반화된 선형 모델과 회귀 트리

2019. 07. 24

정 보 라

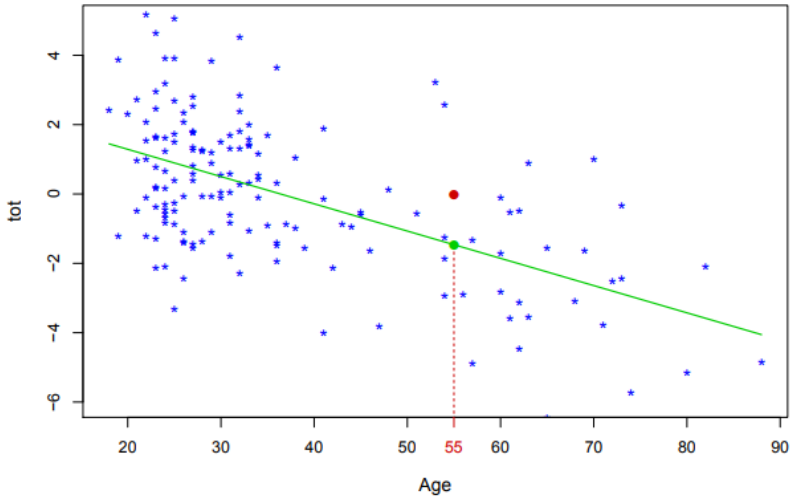
일반화된 선형 모델과 회귀 트리

INTRO

- 회귀모델은 다른 사람의 경험을 받아들이는 빈도주의 기법의 선택
- 실제 값 : 직접 증거 / 회귀선 : 간접 증거
- 현대의 통계적 관행의 특징

  : 기존의 회귀 기법을 점진적으로 공격적으로 사용함

새로운 기법을 위해 만든 용어이며,
타깃 데이터 집합의 크기에 관한 것



8.1 로지스틱 회귀

로지스틱 회귀란? **횟수나 비율 데이터** 의 회귀분석에 사용되는 특수한 기법이다.

1. 지라톤(항암 치료제) 주사량을 증가시켜가며 투약한 실험을 통해 소개 (10번씩, 11개의 실험 쥐 그룹)

- y_i : i 번째 그룹에서 죽은 쥐의 개체
- 죽은 비율

$$p_i = y_i/10, \qquad y_i \overset{\text{ind}}{\sim} \text{Bi}(n_i, \pi_i) \qquad \text{for } i = 1, 2, \ldots, N,$$

- 로짓 모수 : 투약량의 선형 함수

$$\lambda_i = \log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \alpha_0 + \alpha_1 x_i.$$

- 최대 우도 추정은 추정치$(\hat{\alpha}_0, \hat{\alpha}_1)$과의 적합화 곡선

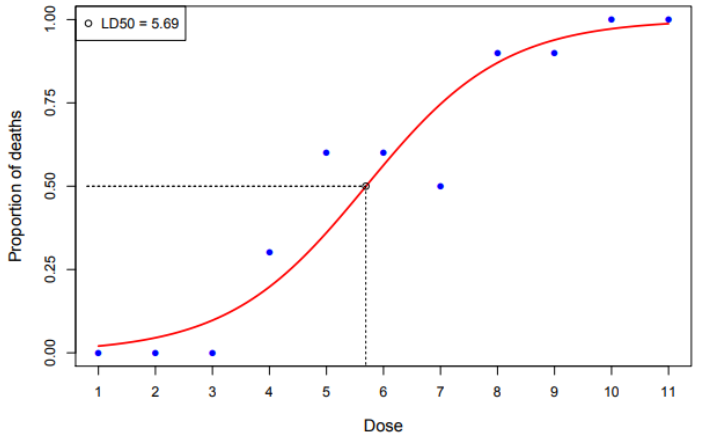
$$\hat{\lambda}(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x.$$

- 로짓 모수 λ의 역변환

$$\pi = \left(1 + e^{-\lambda}\right)^{-1}$$

→ 위 과정으로부터, 선형 로지스틱 회귀 곡선을 얻음

$$\hat{\pi}(x) = \left(1 + e^{-(\hat{\alpha}_0 + \hat{\alpha}_1 x)}\right)^{-1}$$



8.1 로지스틱 회귀

1. 지라톤(항암 치료제) 주사량을 증가시켜가며 투약한 실험을 통해 소개 (10번씩, 11개의 실험 쥐 그룹)

- 실험 그룹 별, 이항표준편차 추정과의 편차 비교

x	1	2	3	4	5	6	7	8	9	10	11
sd $\hat{\pi}(x)$	.015	.027	.043	.061	.071	.072	.065	.050	.032	.019	.010
sd p_i	.045	.066	.094	.126	.152	.157	.138	.106	.076	.052	.035

: 오차는 50% 를 더 줄였지만, 모델이 잘못되었을 경우 편향을 가지게 될 가능성이 있음

- 로짓 변환의 장점
  1. λ 가 $[0,1]$ 에 한정되지 않음 → 금지된 영역에 절대로 가지 않는다.
  2. 지수 패밀리 성질의 이용과 관련이 있음

8.1 로지스틱 회귀

1. 지라톤(항암 치료제) 주사량을 증가시켜가며 투약한 실험을 통해 소개 (10번씩, 11개의 실험 쥐 그룹)

- $\mathrm{Bi}(n, \bar{y})$ 의 밀도함수 / 단일 모수 지수 패밀리(5.5절)에서 α로 불렸던 자연모수 λ가 있음

$$\binom{n}{y}\pi^y(1-\pi)^{n-y} = e^{\lambda y - n\psi(\lambda)}\binom{n}{y} \qquad \lambda = \log\left\{\frac{\pi}{1-\pi}\right\},$$

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ 가, N=11인 전체 데이터 집합을 나타낸다고 정의
- 위 식과 y의 독립성을 사용하면 y의 확률 밀도를 $(\hat{\alpha}_0, \hat{\alpha}_1)$ 의 함수로 나타낼 수 있음

$$f_{\alpha_0,\alpha_1}(\boldsymbol{y}) = \prod_{i=1}^{N} e^{\lambda_i y_i - n_i\psi(\lambda_i)}\binom{n_i}{y_i}$$

$$= e^{\alpha_0 S_0 + \alpha_1 S_1} \cdot e^{-\sum_1^N n_i\psi(\alpha_0+\alpha_1 x_i)} \cdot \prod_{i=1}^{N}\binom{n_i}{y_i},$$

$$S_0 = \sum_{i=1}^{N} y_i \quad \text{and} \quad S_1 = \sum_{i=1}^{N} x_i y_i.$$

- 위 식을 세가지 인자의 곱으로 표현

$$f_{\alpha_0,\alpha_1}(\boldsymbol{y}) = g_{\alpha_0,\alpha_1}(S_0, S_1)h(\alpha_0, \alpha_1)j(\boldsymbol{y}),$$

이 중, 첫 번째만 모수와 데이터에 모두 관여

→ (S_0, S_1) 이 충분통계량 임을 암시

→ N이 얼마나 크든지 두 숫자만 모든 실험정보를 가지고 있음

→ 로지스틱 모수화 에서만!

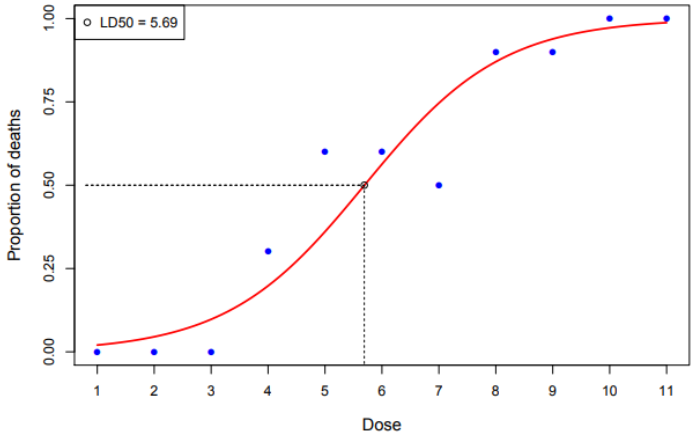
8.1 로지스틱 회귀

1. 지라톤(항암 치료제) 주사량을 증가시켜가며 투약한 실험을 통해 소개 (10번씩, 11개의 실험 쥐 그룹)

- 관측된 비중 p와 추정 파이 간의 편차인 $D(p, \pi)$ 에 종속

$$D\left(p_i, \hat{\pi}_i\right) = 2n_i\left[p_i \log\left(\frac{p_i}{\hat{\pi}_i}\right) + (1-p_i)\log\left(\frac{1-p_i}{1-\hat{\pi}_i}\right)\right].$$

- 로지스틱 회귀 MLE값 $(\hat{\alpha}_0, \hat{\alpha}_1)$, 은 N개의 점 p와 이에대한 추정 π 사이의 전체 편차를 최소화하는 $(\ddot{\alpha}_0, \ddot{\alpha}_1)$의 선택이 된다.



- 실선은 거리를 전체 편차로 계산했을 때,
  선형회귀곡선이 11점에 가장 가까워지게 되는 것이다.

→ 이 방법을 통해 200년이나 된 최소자승법 개념은 이항회귀로 일반화 된다.

8.1 로지스틱 회귀

좀 더 구조화된 로지스틱 회귀 데이터에 대한 분석

2 인간의 근육세포 군집에 쥐의 핵을 다섯가지 비율로 주입하고, 1~5일 동안 배양하여 잘 자라는지 관찰

- π_{ij} : 시간 주기 j 동안 비율 i 가 살아남을 참 확률
- λ_{ij} : 로짓 $\log\{\pi_{ij}/(1-\pi_{ij})\}$.

- 2-방향 가첨 로지스틱 회귀가 데이터에 대해 적합화 됨

$$\lambda_{ij} = \mu + \alpha_i + \beta_j, \qquad i = 1, 2, \ldots, 5, \;\; j = 1, 2, \ldots, 5.$$

→ 이전 실험때 단 두개였던 것과 달리, 모두 9개의 자유모수를 가지고 있다.

- 최대 우도 추정치 (표에서 초록색 수치)

$$\hat{\pi}_{ij} = 1 \Big/ \left[1 + e^{-\left(\hat{\mu}+\hat{\alpha}_i+\hat{\beta}_j\right)}\right]$$

Ratio	Time 1	2	3	4	5
1	5/31 .11	3/28 .25	20/45 .42	24/47 .54	29/35 .75
2	15/77 .24	36/78 .45	43/71 .64	56/71 .74	66/74 .88
3	48/126 .38	68/116 .62	145/171 .77	98/119 .85	114/129 .93
4	29/92 .32	35/52 .56	57/85 .73	38/50 .81	72/77 .92
5	11/53 .18	20/52 .37	20/48 .55	40/55 .67	52/61 .84

8.1 로지스틱 회귀

좀 더 구조화된 로지스틱 회귀 데이터에 대한 분석

3. 스팸 데이터에 적용된 57개 변수의 로지스틱 회귀에 대한 리포트

- $y_i = \begin{cases} 1 & \text{if email } i \text{ is spam} \\ 0 & \text{if email } i \text{ is ham} \end{cases}$

- x_{ij} : 이메일 i 에서의 키워드 j의 상대 빈도
- π_i : 이메일 i 가 스팸일 확률
- 로짓 변환하여, 가첨적 로지스틱 모델을 적합화

$$\lambda_i = \alpha_0 + \sum_{j=1}^{57} \alpha_j x_{ij}.$$

• 매우 큰 표준편차를 가진 매우 큰 추정치를 나타내는데

→ 고차원의 최대 우도 추정이 가지는 위험성을 볼 수 있다.

→ 축소 추정이 필요함(16장)

	Estimate	se	z-value		Estimate	se	z-value
intercept	−12.27	1.99	−6.16	lab	−1.48	.89	−1.66
make	−.12	.07	−1.68	labs	−.15	.14	−1.05
address	−.19	.09	−2.10	telnet	−.07	.19	−.35
all	.06	.06	1.03	857	.84	1.08	.78
3d	3.14	2.10	1.49	data	−.41	.17	−2.37
our	.38	.07	5.52	415	.22	.53	.42
over	.24	.07	3.53	85	−1.09	.42	−2.61
remove	.89	.13	6.85	technology	.37	.12	2.99
internet	.23	.07	3.39	1999	.02	.07	.26
order	.20	.08	2.58	parts	−.13	.09	−1.41
mail	.08	.05	1.75	pm	−.38	.17	−2.26
receive	−.05	.06	−.86	direct	−.11	.13	−.84
will	−.12	.06	−1.87	cs	−16.27	9.61	−1.69
people	−.02	.07	−.35	meeting	−2.06	.64	−3.21
report	.05	.05	1.06	original	−.28	.18	−1.55
addresses	.32	.19	1.70	project	−.98	.33	−2.97
free	.86	.12	7.13	re	−.80	.16	−5.09
business	.43	.10	4.26	edu	−1.33	.24	−5.43
email	.06	.06	1.03	table	−.18	.13	−1.40
you	.14	.06	2.32	conference	−1.15	.46	−2.49
credit	.53	.27	1.95	char;	−.31	.11	−2.92
your	.29	.06	4.62	char(	−.05	.07	−.75
font	.21	.17	1.24	char[	−.07	.09	−.78
000	.79	.16	4.76	char!	.28	.07	3.89
money	.19	.07	2.63	char$	1.31	.17	7.55
hp	−3.21	.52	−6.14	char#	1.03	.48	2.16
hpl	−.92	.39	−2.37	cap.ave	.38	.60	.64
george	−39.62	7.12	−5.57	cap.long	1.78	.49	3.62
650	.24	.11	2.24	cap.tot	.51	.14	3.75

8.2 일반화 선형 모델

로지스틱 회귀는 일반화 선형 모델(GLM)의 특수한 경우로, 알고리즘과 추론에 모두 영향을 끼친 1970년대의 핵심 기법

GLM은 최소 자승 곡선 적합화인 선형 회귀를 다양한 반응 변수로까지 확장시켜주는데, 이항, 포아송, 베타, 지수분포 형태까지도 포함된다.

- 단일 모수 지수 패밀리로 시작한다. (λ: 자연모수, y: 충분통계량)

$$\left\{ f_\lambda(y) = e^{\lambda y - \gamma(\lambda)} f_0(y),\ \lambda \in \Lambda \right\} \quad y_i \sim f_{\lambda_i}(\cdot) \text{ independently for } i = 1, 2, \ldots, N.$$

N-모수 모델을
p-모수 지수 패밀리로 축소
–> 고차원 추정의 어려움을 피함

- 핵심 GLM 전술은 λ를 선형 회귀 방정식으로 표현하는 것이다.
- **X**가 N X p 구조행렬이라고 하면, N-벡터 λ는 $\boldsymbol{\lambda} = \boldsymbol{X}\alpha.$ 로 나타낼 수 있다.
- 데이터 벡터 y의 확률 밀도 함수는 $f_\alpha(\boldsymbol{y}) = \prod_{i=1}^{N} f_{\lambda_i}(y_i) = e^{\sum_1^N (\lambda_i y_i - \gamma(\lambda_i))} \prod_{i=1}^{N} f_0(y_i)$ 이고, $f_\alpha(\boldsymbol{y}) = e^{\alpha' z - \psi(\alpha)} f_0(\boldsymbol{y})$ 로 쓸 수 있다.
- 위 식에서, $z = X'y \quad \text{and} \quad \psi(\alpha) = \sum_{i=1}^{N} \gamma(x_i'\alpha),$

  z: 충분통계량 벡터, α: 자연 모수 벡터

→ p-모수 GLM으로부터의 모든 정보는 p차원 벡터 z에 요약돼있고, N의 크기와 상관 없이 이해와 분석이 모두 더 쉬워진다.

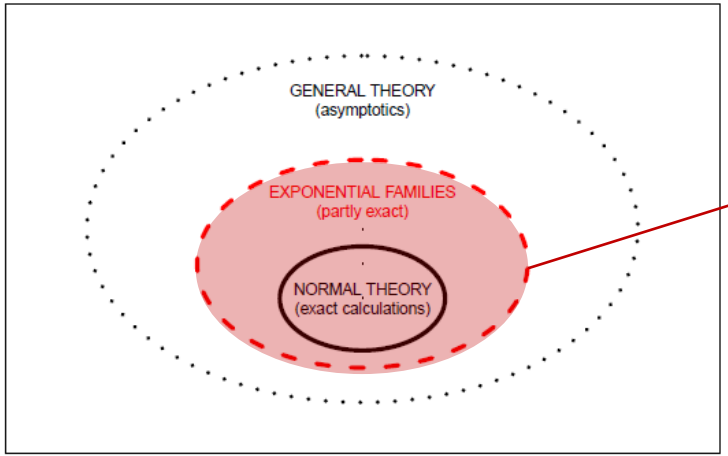
8.2 일반화 선형 모델

[회프딩의 보조정리]

주어진 y에 대해 최대 우도 추정 μ 는 y 자신이고, 로그 우도 $\log f_\mu(y)$는 최대 $\log f_y(y)$ 로부터 편차 $D(y,\mu)$에 따른 크기만큼 감소한다.

$$f_\mu(y) = f_y(y)e^{-D(y,\mu)/2}.$$

★ MLE $\hat{\alpha}$ 는 전체 편차 $\sum_1^N D(y_i, \mu_i(\alpha))$ 를 최소화하는 α 를 고르는 것이다.



내부 원 : 정규 이론 – 정확 추론(t 검정, F 분포, 다변량 분석)

원 바깥 : 테일러 전개식, 중심극한 이론에 근거한 점근 근사

지수 패밀리 이론 : 정규이론에 이상적인 대응이 되는 볼록 우도 표면, 최소 편차 회귀 등의 일부 정확함을 가지면서 동시에 일부 근사가 필요함

로지스틱 회귀는 추정 효율과 계산에서의 이점, 그리고 200년 된 최소 자승법과도 훨씬 더 유사하기 때문에 그 이전방법을 완전히 몰아냈으며,

GLM은 대부분 빈도주의지만 우도 기반 추론과 피셔의 '귀납적 추론의 논리' 로 부터 얻은 힌트를 혼합한 것

8.3 포아송 회귀

• 포아송 회귀는, 최소자승과 로지스틱 회귀에 이어 세번째로 빈번하게 사용되는 것

• **하늘의 일부에서 측정한 은하계 계수 데이터** (표의 일부는 잘렸음)
486개의 은하계에 대해 적색편이 r 과 시등급 m을 측정한 것

redshift (farther) ⟶

magnitude (dimmer) ↑	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
18	1	6	6	3	1	4	6	8	8	20	10	7	16	9	4
17	3	2	3	4	0	5	7	6	6	7	5	7	6	8	5
16	3	2	3	3	3	2	9	9	6	3	5	4	5	2	1
15	1	1	4	3	4	3	2	3	8	9	4	3	4	1	1
14	1	3	2	3	3	4	5	7	6	7	3	4	0	0	1
13	3	2	4	5	3	6	4	3	2	2	5	1	0	0	0
12	2	0	2	4	5	4	2	3	3	0	1	2	0	0	1
11	4	1	1	4	7	3	3	1	2	0	1	1	0	0	0
10	1	0	0	2	2	2	1	2	0	0	0	1	2	0	0
9	1	1	0	2	2	2	0	0	0	0	1	0	0	0	0
8	1	0	0	0	1	1	0	0	0	0	1	1	0	0	0
7	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
6	0	0	3	1	1	0	0	0	0	0	0	0	0	0	0
5	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0
3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

• r : 역의 r 값을 나열하는 270-벡터

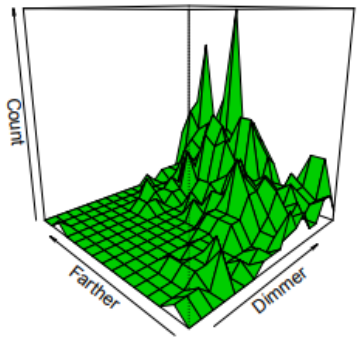
• m : 270 m 값

• 270 X 5 행렬 X를 정의 $\boldsymbol{X} = [\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{r}^2, \boldsymbol{rm}, \boldsymbol{m}^2]$,

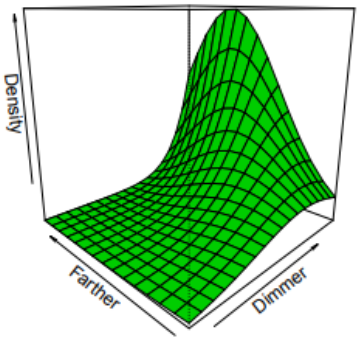
• (r,m)에서의 이변량 정규분포의 로그밀도는
$\alpha_1 r + \alpha_2 m + \alpha_3 r^2 + \alpha_4 rm + \alpha_5 m^2$ 이며, $\log \mu_i = x_i' \alpha$ 와 일치

=> i 번째 칸의 개수를 y로 하는 포아송 GLM을 사용해 잘려나간 영역에 대한 가상의 이변량 정규분포 부분의 추정을 한다.

8.3 포아송 회귀

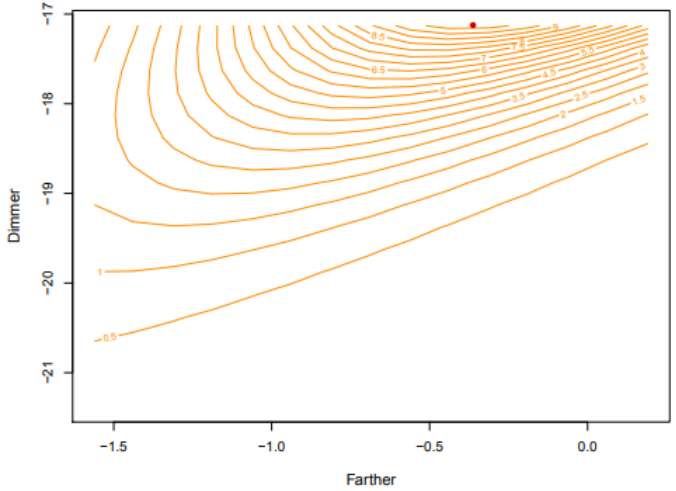


은하계 데이터, 각 칸의 개수



포아송 GLM 밀도 추정

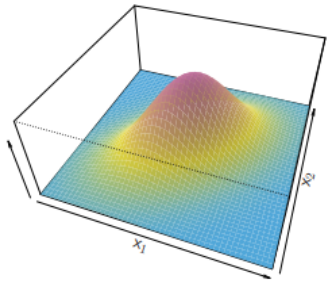
적합화된 로그 밀도의 동일값의 등고선

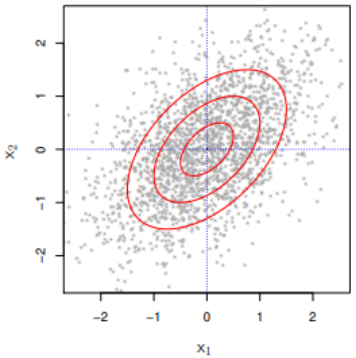


•관측된 개수 y 와 적합화 값 $\hat{\mu}$ 사이의 **포아송 편차 잔차**

$$Z = \text{sign}(y - \hat{\mu})\, D(y, \hat{\mu})^{1/2},$$

• **포아송 GLM은** 밀도 추정을 익숙하고 유연한 추론 기술인 회귀 모델 적합화로 축소시켜 준다.





8.4 회귀 트리

- 포아송 GLM 같은 회귀 알고리즘은 입력에 대해 규칙을 생성하는데, 이 규칙에는 3가지 주요 용도가 있다.

1) 예측 : 주어진 새로운 관측치 x에 대해 상응하는 y 값이 알려져 있지 않은 경우
(스팸 예제에서 스팸여부를 예측)

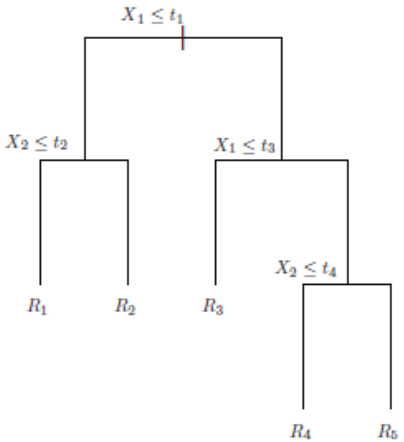
2) 추정 : χ 에 대한 회귀표면 $\hat{s}$ 를 기술함, S를 정교하게 알아내고자 함
(은하계 밀도에서 잘려나간 부분 추정)

3) 설명 : 회귀 표면이 어떻게 형성되었는지
(당뇨병 데이터의 열개 예측변수 중 당뇨병 진행의 원인 변수를 선택 – 서로 다른 예측변수들의 상대적 공헌도)

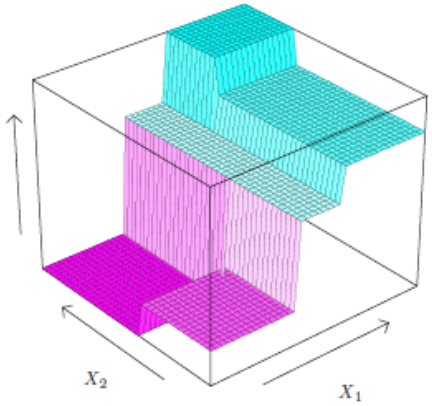
8.4 회귀 트리

- 회귀 트리는 회귀 표면을 추정하기 위해, **재귀적 분할**을 사용함

두 예측변수 X1, X2인 경우의 가상 상황



좌측에 해당하는 회귀 표면 그림



8.4 회귀 트리

- 어떤 변수를 분할할 것인지, 어떤 분할 값을 트리 구성에 사용할 것인지 결정함

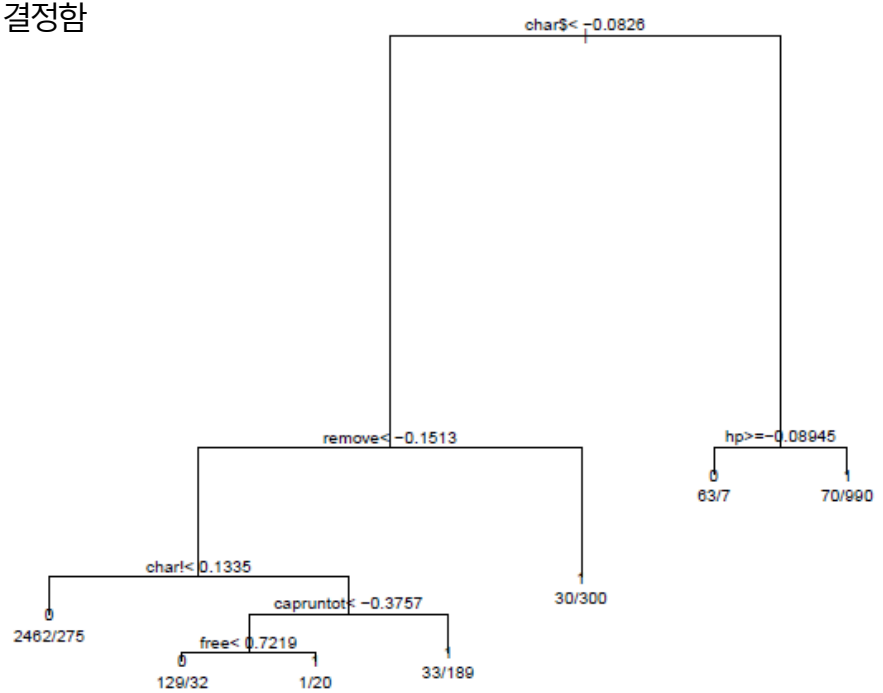
- 알고리즘 단계 k에서 N_k개의 경우 수를 가진 group_k 가 분할 대상
- 이 경우, 아래와 같은 평균과 제곱합을 가짐

$$m_k = \sum_{i \in \text{group}_k} y_i / N_k \quad \text{and} \quad s_k^2 = \sum_{i \in \text{group}_k} (y_i - m_k)^2.$$

- group_k 분할
  - $\text{group}_{k,\text{left}}$, $\text{group}_{k,\text{right}}$
  - $m_{k,\text{left}}$, $m_{k,\text{right}}$ / $s_{k,\text{left}}^2$, $s_{k,\text{right}}^2$.

- 알고리즘은 다음을 최소화 하는 문턱 값과 분할 변수를 선택하며 진행

  $s_{k,\text{left}}^2 + s_{k,\text{right}}^2$. : 두 그룹이 가능한 한 서로 완전히 다르게 분할



[스팸 데이터에 대한 회귀 트리]

- 과해석이 높아 불안정하며 추정으로의 사용은 불가능, "예측" 알고리즘의 핵심부분으로 쓰임
- 회귀트리는 직관적인 전통적 기법과의 단절을 나타냄 → 완전히 비모수적