

예측방법론

김지수



서울대학교 통계학과
2024-02-21

강의순서

- 회귀분석: 모형, 목적
- 단순선형회귀, 다중선형회귀: 추론
- 가변수(지시변수)
- 모형진단
- 변수(모형) 선택: 최적부분집합선택, 능선회귀, 라쏘
- 다항회귀, 비모수회귀, 일반화 가법모형

회귀분석: 모형, 목적

단순선형회귀, 다중선형회귀: 추론

가변수(지시변수)

모형진단

변수(모형) 선택: 최적부분집합선택, 능선회귀, 라쏘

다항회귀, 비모수회귀, 일반화 가법모형

광고자료와 회귀적합

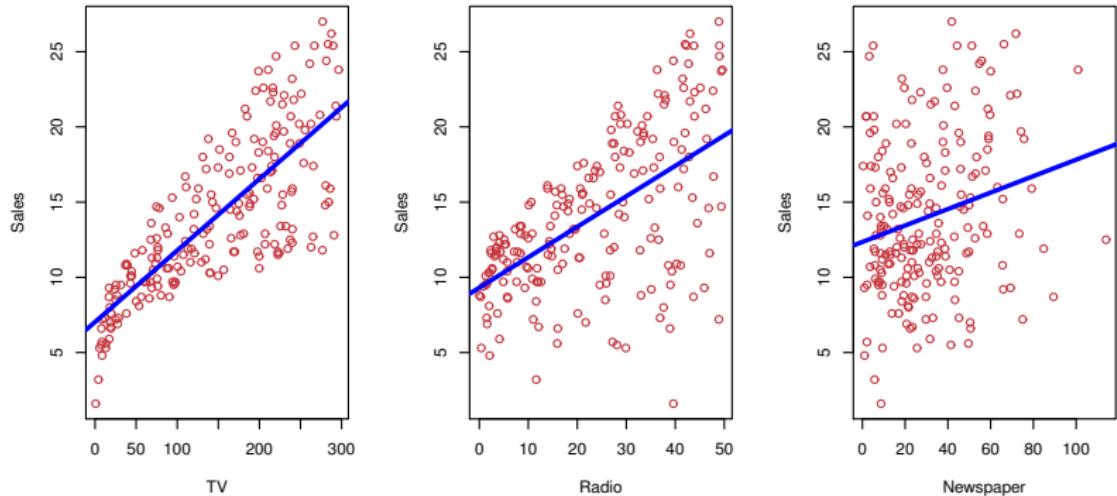


Figure: 마켓(200)에 관해 매출(단위 천개), TV, radio, newspaper 광고 지출액
(단위: 1000불)을 포함한 자료와 회귀적합결과.

회귀모형과 목적

- 변수간의 관계를 (확률)모형화하고 이를 이용하여 그 관계를 분석하는 통계적 방법(추론)
- 회귀모형

$$Y = f(X) + \epsilon$$

f : 함수 (모수, 비모수)

Y : 반응변수, 종속변수

$X = (X_1, \dots, X_p)$: 예측변수, 독립변수, 설명변수

ϵ : 평균이 0인 랜덤 오차항

- 목적

예측(prediction) : $\hat{Y} = \hat{f}(X)$ 로 주어진 X 에서 Y 값을 예측

추론(inference) : X 와 Y 의 관계를 이해

회귀분석: 모형, 목적

단순선형회귀, 다중선형회귀: 추론

가변수(지시변수)

모형진단

변수(모형) 선택: 최적부분집합선택, 능선회귀, 라쏘

다항회귀, 비모수회귀, 일반화 가법모형

단순선형회귀(Simple Linear Regression)

- 모형

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$f(X) = \beta_0 + \beta_1 X$$

β_0, β_1 : 회귀계수

ϵ : 평균이 0이고 분산이 σ^2 인 독립 오차항

- 예 :

$$\text{sales} \approx \beta_0 + \beta_1 \times \text{TV.}$$

- $E(Y|X) = \beta_0 + \beta_1 X$ (참 회귀선은 평균반응선과 일치)
 $Var(Y|X) = \sigma^2$

- 구체적인 목적

- ▶ 회귀계수 추정
- ▶ $H_0 : \beta_1 = 0$ 검정과 신뢰구간 구함

회귀계수의 추정 : 최소제곱법

- ▶ 최소제곱법(least squares method) :
잔차제곱합을 최소화하는 $\hat{\beta}_0, \hat{\beta}_1$ 을
구하는 방법
- ▶ 잔차제곱합(residual sum of squares,
RSS)

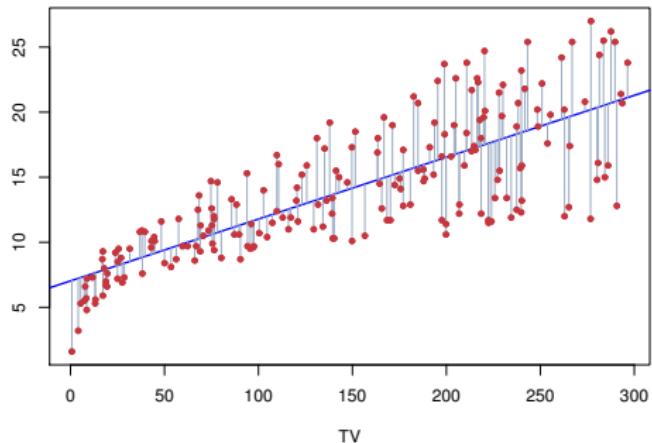
$$S(\beta_0, \beta_1) \Big|_{\beta_0 = \hat{\beta}_0, \beta_1 = \hat{\beta}_1} = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- ▶ 적합값(fitted value)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad i = 1, 2, \dots, n.$$

- ▶ 잔차(residual)

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n.$$



최소제곱합

추정치의 식

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

광고자료 예

$$\text{sales} = 7.03 + 0.0475 \times \text{TV}$$

불편성과 표준오차(분산)

불편성

분산(표준오차*표준오차)

$$\begin{aligned}\mathbb{E}(\hat{\beta}_0) &= \beta_0 \\ \mathbb{E}(\hat{\beta}_1) &= \beta_1\end{aligned}$$

$$\begin{aligned}SE(\hat{\beta}_0)^2 &= \text{Var}(\hat{\beta}_0) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]\end{aligned}$$

σ 의 추정

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-2}}$$

$$\begin{aligned}SE(\hat{\beta}_1)^2 &= \text{Var}(\hat{\beta}_1) \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

신뢰구간과 가설검정

95% 신뢰구간

$$\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$$

$$\hat{\beta}_0 \pm 1.96SE(\hat{\beta}_0)$$

p -값(p-value)이란

H_0 가 참일 때 현재의 관측치보다 더 H_1 에 가까운 혹은 H_1 을 지지할 자료를 관측할 확률이다.

$$p\text{-값} = \mathbb{P}_{H_0}(|T| \geq |t|)$$

여기서 t 는 관측된 t 값이다.

가설

H_0 : X 와 Y 사이에 관계가 없다. vs

H_0 : X 와 Y 사이에 관계가 있다.

$$\implies H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0.$$

유의수준 (significance level)

$100\alpha\%$ 일 때 기각역

$$|t| = \left| \frac{\hat{\beta} - 0}{SE(\hat{\beta})} \right| > t_{\alpha/2}(n - 2)$$

$sales = \beta_0 + \beta_1 \times TV$ 적합했을 때의 R 결과

```
advertising <-  
  read.csv("https://www.statlearning.com/s/Advertising.csv")[, -1]  
lm.fit <- lm(sales ~ TV, data = advertising)  
summary(lm.fit)  
  
##  
## Call:  
## lm(formula = sales ~ TV, data = advertising)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -8.3860 -1.9545 -0.1913  2.0671  7.2124  
##  
## Coefficients:  
##                 Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 7.032594    0.457843   15.36   <2e-16 ***  
## TV          0.047537    0.002691   17.67   <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 3.259 on 198 degrees of freedom  
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099  
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

$sales = \beta_0 + \beta_1 \times TV$ 를 적합했을 때의 R 결과

	회귀계수	표준오차	t-통계량	p값
intercept	7.0325	0.4578	15.36	≤ 0.0001
TV	0.0475	0.0027	17.67	≤ 0.0001

모형 적합성

모형 적합성의 정도를
나타내는 척도

1. 잔차표준오차 :

3.26

2. R^2 : 0.612

3. F 값 : 312.1

잔차표준오차

residual standard error, RSE
root mse

$$\begin{aligned} RSE &= \sqrt{\frac{1}{n-2} RSS} \\ &= \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \end{aligned}$$

R^2 결정계수

정의

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 값은 X 와 Y 의 선형관계의 정도를 측정한다.

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\begin{aligned} r &= \text{corr}(X, Y) \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \end{aligned}$$

$$R^2 = r^2$$

R^2 는 y 의 전체 변동량 중 x 를 이용해 선형모형으로 설명하는 비율이다.

중회귀모형

모형

예

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

$$\begin{aligned} sales &= \beta_0 + \beta_1 \times TV \\ &\quad + \beta_2 \times radio + \beta_3 \times newspaper + \epsilon \end{aligned}$$

X_j : j 번째 예측변수

β_j : X_j 의 회귀계수

$$f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

ϵ : 평균이 0이고 분산이 σ^2 인
독립 오차항

예측식

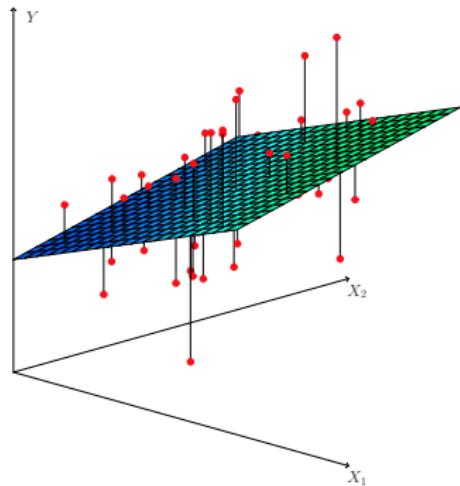
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$$

회귀계수 추정

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}))^2$$

를 최소화하는 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 를
추정량으로 한다.



$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times newspaper$$
 적합

```
lm.fit <- lm(sales ~ TV + radio + newspaper, data=advertising)
summary(lm.fit)

## 
## Call:
## lm(formula = sales ~ TV + radio + newspaper, data = advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.938889  0.311908   9.422 <2e-16 ***
## TV          0.045765  0.001395  32.809 <2e-16 ***
## radio       0.188530  0.008611  21.893 <2e-16 ***
## newspaper   -0.001037  0.005871  -0.177    0.86  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956 
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

변수들간의 상관계수

```
cor(advertising)

##                      TV      radio newspaper      sales
## TV            1.00000000 0.05480866 0.05664787 0.7822244
## radio         0.05480866 1.00000000 0.35410375 0.5762226
## newspaper    0.05664787 0.35410375 1.00000000 0.2282990
## sales         0.78222442 0.57622257 0.22829903 1.0000000
```

반응변수와 예측변수들 사이에 관계가 있는가?

가설

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$ vs

H_1 : 최소한 한 개의 β_j 는 0이 아니다.

F-통계량

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

q 개의 회귀계수가 0 이라는 검정

통계량

$$H_0 : \beta_{p-q+1} = \dots = \beta_p = 0$$

vs $H_1 : \beta_{p-q+1}, \dots, \beta_p$ 중 하나는 0 이 아니다.

F-통계량

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)}$$

여기서 RSS_0 는 q 개의 변수를 제외하고 적합한 회귀모형의 잔차제곱합이다.

회귀분석: 모형, 목적

단순선형회귀, 다중선형회귀: 추론

가변수(지시변수)

모형진단

변수(모형) 선택: 최적부분집합선택, 능선회귀, 라쏘

다항회귀, 비모수회귀, 일반화 가법모형

신용카드자료

balance : 신용카드 대출의 크기

age : 나이(햇수)

cards : 신용카드의 개수

education : 교육받은 햇수

income : 수입(천불)

limit : 신용카드 한도

rating : 신용평가

own : 집 유무

student : 학생신분 유무

status : 결혼 상태

region : 출신 지역 (동, 서, 남)

가변수(dummy variable) : 두 개의 값을 갖는 변수

가변수

$$x_i = \begin{cases} 1, & i\text{번째 사람이 집이 있음} \\ 0, & i\text{번째 사람이 집이 없음} \end{cases}$$

회귀모형에서 가변수의 해석

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & i\text{번째 사람이 집이 있음} \\ \beta_0 + \epsilon_i, & i\text{번째 사람이 집이 없음} \end{cases} \end{aligned}$$

회귀모형 추정치

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	<0.0001
Own[Yes]	19.73	46.05	0.429	0.6690

$$Balance = \beta_0 + \beta_1 \times OwnYes$$

```
credit <- read.csv("https://www.statlearning.com/s/Credit.csv")
lm.fit <- lm(Balance ~ Own, data = credit)
summary(lm.fit)

##
## Call:
## lm(formula = Balance ~ Own, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -529.54 -455.35 -60.17  334.71 1489.20 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 509.80     33.13   15.389   <2e-16 ***
## OwnYes       19.73     46.05    0.429    0.669    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 460.2 on 398 degrees of freedom
## Multiple R-squared:  0.0004611, Adjusted R-squared:  -0.00205 
## F-statistic: 0.1836 on 1 and 398 DF,  p-value: 0.6685
```

가변수(dummy variable) : 세 개 이상의 값을 갖는 변수

가변수의 정의

$$x_{i1} = \begin{cases} 1, & i\text{번째 사람이 남쪽 출신} \\ 0, & i\text{번째 사람이 남쪽 출신 아님} \end{cases}, \quad x_{i2} = \begin{cases} 1, & i\text{번째 사람이 서쪽 출신} \\ 0, & i\text{번째 사람이 서쪽 출신 아님} \end{cases}.$$

회귀모형에서 가변수의 해석

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & i\text{번째 사람이 남쪽 출신} \\ \beta_0 + \beta_2 + \epsilon_i, & i\text{번째 사람이 서쪽 출신} \\ \beta_0 + \epsilon_i, & i\text{번째 사람이 동쪽 출신} \end{cases}. \end{aligned}$$

회귀모형 추정치

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	<0.0001
Region[South]	-12.50	56.68	-0.221	0.8260
Region[West]	-18.69	65.02	-0.287	0.7740

$$Balance = \beta_0 + \beta_1 \times RegionSouth + \beta_2 \times RegionWest$$

```
lm.fit <- lm(Balance ~ Region, data = credit)
summary(lm.fit)

##
## Call:
## lm(formula = Balance ~ Region, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -531.00 -457.08 -63.25  339.25 1480.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 531.00     46.32 11.464   <2e-16 ***
## RegionSouth -12.50     56.68 -0.221    0.826
## RegionWest  -18.69     65.02 -0.287    0.774
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 460.9 on 397 degrees of freedom
## Multiple R-squared:  0.0002188, Adjusted R-squared:  -0.004818
## F-statistic: 0.04344 on 2 and 397 DF,  p-value: 0.9575
```

회귀분석: 모형, 목적

단순선형회귀, 다중선형회귀: 추론

가변수(지시변수)

모형진단

변수(모형) 선택: 최적부분집합선택, 능선회귀, 라쏘

다항회귀, 비모수회귀, 일반화 가법모형

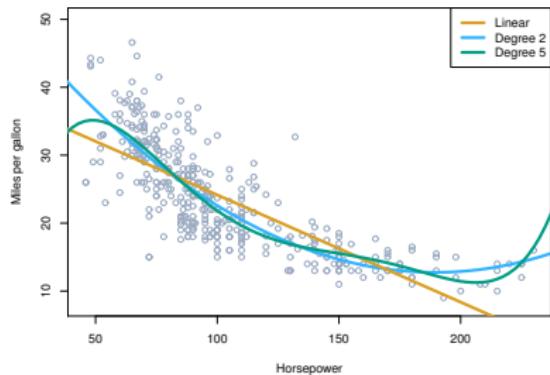
중회귀모형

- 모형

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- ϵ : 평균이 0이고 분산이 σ^2 인 독립 오차항, 즉
 - $E(Y|X) = f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$ (참 회귀선은 평균반응선과 일치)
 - $Var(Y|X) = \sigma^2$

선형성 가정의 완화

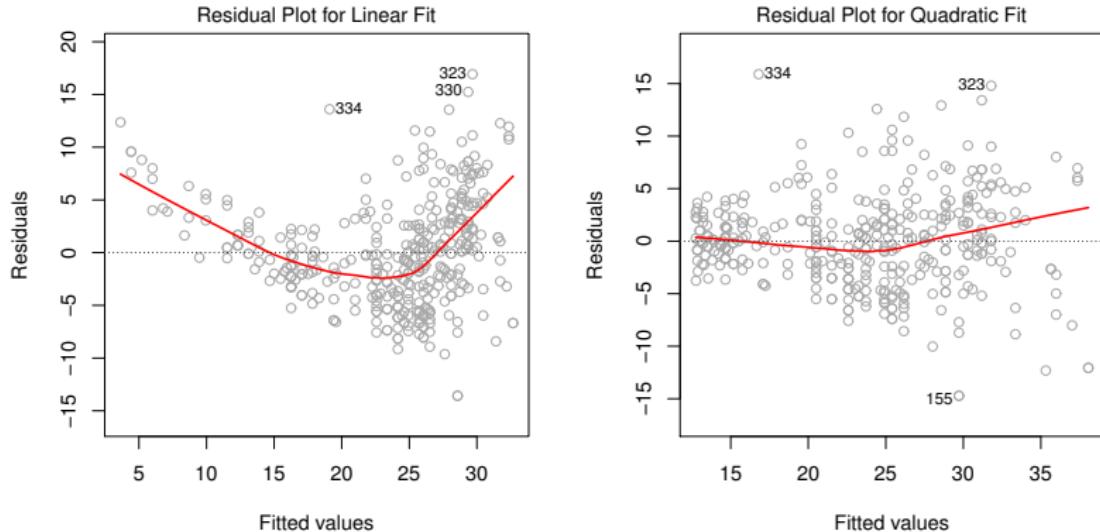


	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower ²	0.0012	0.0001	10.1	< 0.0001

2차선형회귀모형

$$mpg = \beta_0 + \beta_1 \times horsepower + \beta_2 \times horsepower^2 + \epsilon$$

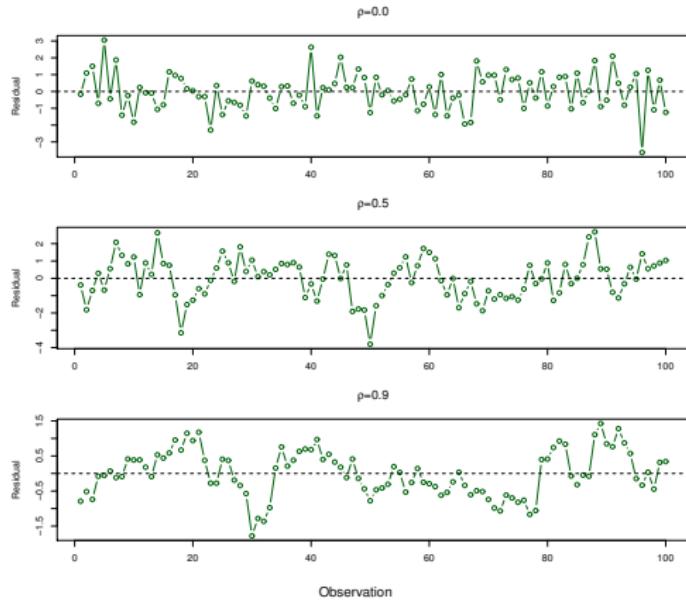
설명변수-반응변수 관계의 비선형성



해결책

간단한 방법으로 문제가 되는 반응변수를 $\log X$, \sqrt{X} , X^2 으로 변환하거나, 이 변수들을 포함하는 것이다. 다른 고급 방법들도 많이 존재한다.

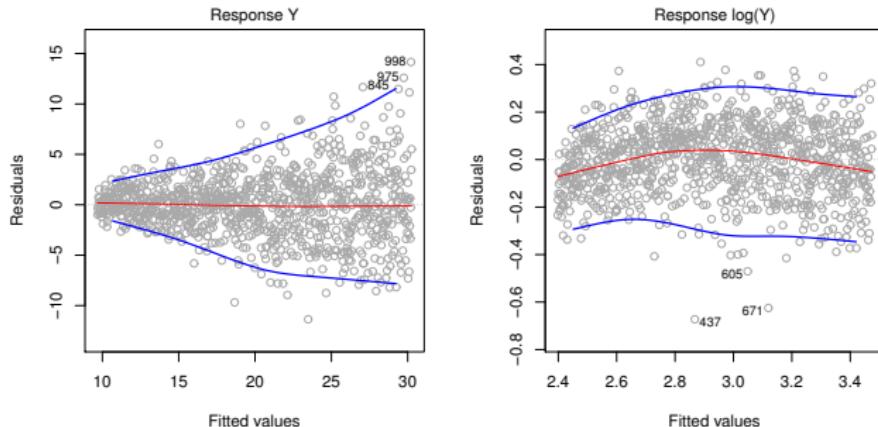
오차의 자기 상관성



해결책

시계열 모형을 적용한다. 혹은 선형모형의 분산을 바꾼다.

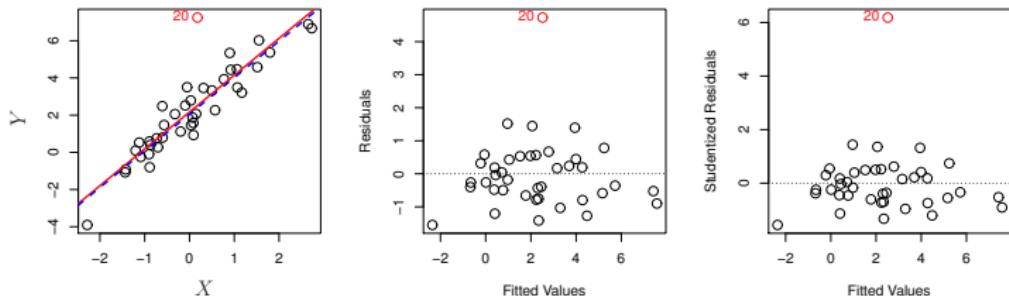
오차의 이분산성



해결책

1. Y 를 $\log Y$ 나 \sqrt{Y} 로 변환한다.
2. 가중최소제곱법(weighted least squares method)을 이용하여 추정한다.

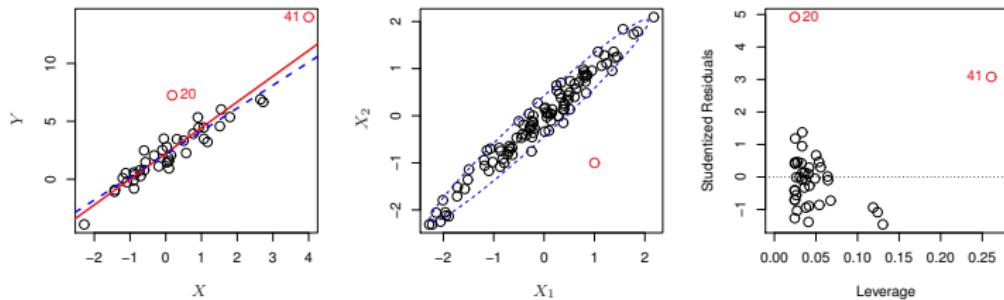
이상점



해결책

1. 이상점이 자료수집시 실수 때문에 발생한 것이면 제거가 가능하다. 그렇지 않을 경우는 중요한 예측변수가 포함이 안되었거나, 모형이 적합하지 않을 수 있을 수 있다. 이상점을 제거하는 것에 신중을 기해야한다.

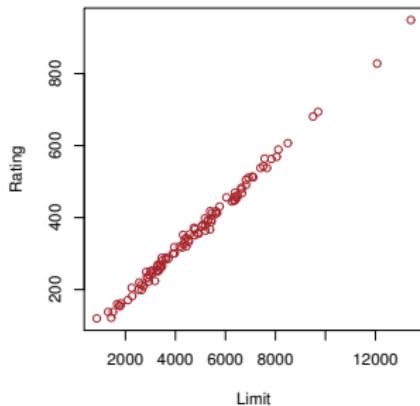
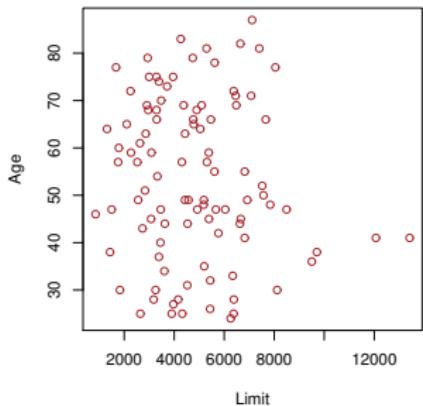
지렛대점



1. 보통의 범위에서 벗어난 x_i 값 때문에 추정식에 큰 영향을 미치는 점을 말한다.
2. 지렛대 통계량(leverage statistic)으로 체크한다. 단순회귀일 때는 다음과 같은 식을 갖는다.

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}$$

다중공선성



1. 두개 이상의 설명변수가 상관성이 큰 경우이다.
2. 표준오차가 매우 커진다.
3. 통계적 가설 $H_0 : \beta = 0$ 의 검정력이 떨어진다.

다중공선성

감지(detect)하는 법

1. 예측 변수간의 상관계수를 본다.
2. 분산팽창인수(VIF, variance inflation factor)를 본다.

해결책

1. 문제가 되는 변수를 제거한다.
2. 문제가 되는 여러 개의 변수를 한 개의 변수로 합친다.

회귀분석: 모형, 목적

단순선형회귀, 다중선형회귀: 추론

가변수(지시변수)

모형진단

변수(모형) 선택: 최적부분집합선택, 능선회귀, 라쏘

다항회귀, 비모수회귀, 일반화 가법모형

변수선택

모든 모형 비교 방법

1. Mallow's C_p , AIC, BIC, adjusted R^2 등이 있다.
2. $p = 30$ 만 되어도 이 방법을 쓰기가 어렵다.
 $2^{30} = 1,073,741,824$.

자동선택 방법

1. 전진선택법(forward selection)
2. 후진선택법(backward selection)
3. 단계적선택법(stepwise or mixed selection)

최적부분집합선택(Best Subset Selection)

알고리듬

1. M_0 을 영모형(null model)이라 하자. 영모형은 예측변수를 하나도 포함하지 않은 모형을 말한다.
2. $k = 1, 2, \dots, p$
 - 2.1 예측변수가 k 개인 $\binom{p}{k}$ 개의 모형을 고려한다.
 - 2.2 이 중 RSS가 가장 작거나 R^2 가 가장 큰 모형을 M_k 라고 한다.
3. M_0, \dots, M_p 중 C_p , AIC, BIC, adjusted R^2 중 하나의 기준을 이용하여 가장 좋은 모형을 선택한다.

전진선택법(Forward Stepwise Selection)

알고리듬

1. M_0 을 영모형(null model)이라 하자.
2. $k = 0, 1, 2, \dots, p - 1$
 - 2.1 M_k 에 포함이 안된 $p - k$ 개의 예측변수를 하나씩 M_k 에 추가한 $p - k$ 개의 모형을 고려한다.
 - 2.2 이 중 RSS가 가장 작거나 R^2 가 가장 큰 모형을 M_{k+1} 라고 한다.
3. M_0, \dots, M_p 중 C_p , AIC, BIC, adjusted R^2 중 하나의 기준을 이용하여 가장 좋은 모형을 선택한다.

후진선택법(Backward Stepwise Selection)

알고리듬

1. M_p 을 full model이라 하자. 예측변수를 모두 포함하는 모형을 말한다.
2. $k = p, p - 1, \dots, 1$
 - 2.1 M_k 에서 한 개의 변수를 제거한 k 개의 모형을 고려한다.
 - 2.2 RSS가 가장 작거나 R^2 가 가장 큰 모형을 M_{k-1} 라고 한다.
3. M_0, \dots, M_p 중 C_p , AIC, BIC, adjusted R^2 중 하나의 기준을 이용하여 가장 좋은 모형을 선택한다.

혼합방법

알고리듬

1. M_0 을 영모형(null model)이라 하자.
2. $k = 0, 1, 2, \dots, p - 1$
 - 2.1 M_k 에 포함이 안된 $p - k$ 개의 예측변수를 하나씩 M_k 에 추가한 $p - k$ 개의 모형을 고려한다.
 - 2.2 이 중 RSS가 가장 작거나 R^2 가 가장 큰 모형을 M_{k+1} 라고 한다.
 - 2.3 포함된 예측변수 중 기준에 맞지 않는 변수를 제거한다.
3. 거쳐간 모든 모형 중 C_p , AIC, BIC, adjusted R^2 중 하나의 기준을 이용하여 가장 좋은 모형을 선택한다.

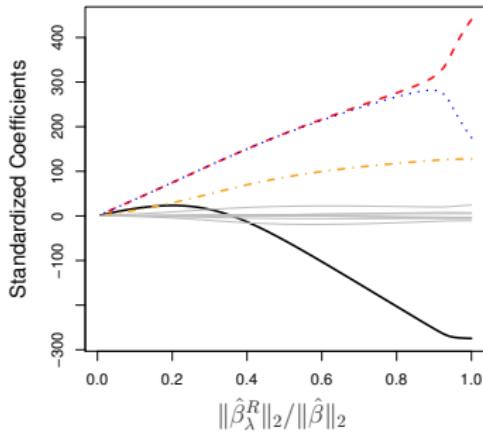
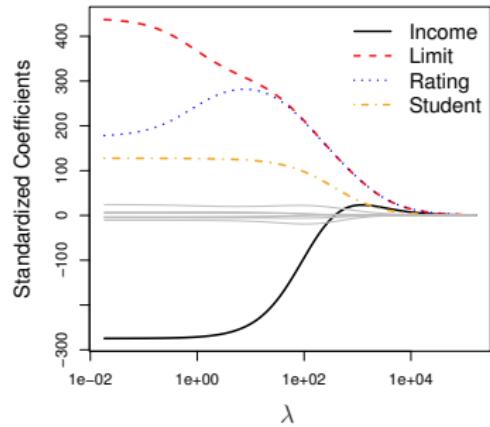
능선회귀(Ridge Regression)

능선회귀 추정량 $\hat{\beta}^R$ 은

$$\sum_{i=1}^n (y_i - (\beta_0 + \sum_{j=1}^p \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^p \beta_j^2 = RSS + \lambda \sum_{j=1}^p \beta_j^2$$

을 최소화하는 β 값으로 정의된다.

λ 의 변화에 따른 추정량의 변화



변수의 표준화

- 최소제곱추정량은 척도등변추정량(scale equivariant estimator)이다.

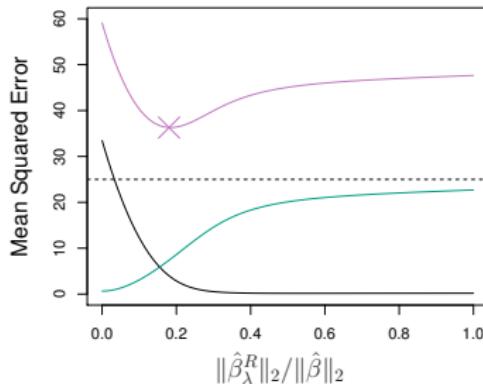
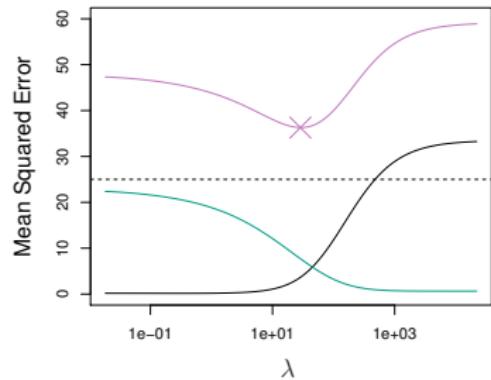
$$credit = \beta_0 + \beta_1 \times income + \epsilon$$

의 모형을 생각할 때 *income*을 천불단위로 하든 1불 단위로 하든 *credit*의 예측에는 변화가 없다.

- 능선회귀추정량은 척도등변추정량이 아니다. 단위에 따라 \hat{credit} 값이 달라질 수 있다.
- 능선회귀를 적용할 때는 모든 변수를 표준화시켜 같은 척도를 갖기를 추천한다.

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}.$$

능선회귀의 성능이 최소제곱법 보다 좋은 이유



λ 가 커지면서 편향의 제곱(검은색)은 커지면서 분산(녹색)은 작아진다.
최소제곱오차(붉은색)은 작아지다 커진다.

능선회귀 R 코드 |

```
Hitters <- na.omit(Hitters)
x <- model.matrix(Salary ~ ., Hitters)[, -1]
y <- Hitters[["Salary"]]

library(glmnet)

## Warning: package 'glmnet' was built under R version 4.3.2
## Loading required package: Matrix
## Warning: package 'Matrix' was built under R version 4.3.2
## Loaded glmnet 4.1-8

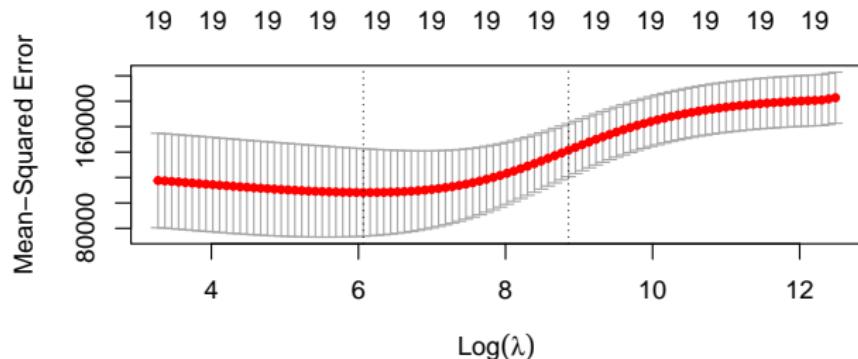
grid <- 10 ^ seq(10, -2, length = 100)
ridge.mod <- glmnet(x, y, alpha = 0, lambda = grid)
predict(ridge.mod, s = 50, type = "coefficients")[1:20, ]

##      (Intercept)          AtBat          Hits         HmRun        Runs
## 4.876610e+01 -3.580999e-01 1.969359e+00 -1.278248e+00 1.145892e+00
##          RBI          Walks          Years        CAtBat       CHits
## 8.038292e-01  2.716186e+00 -6.218319e+00  5.447837e-03 1.064895e-01
##         CHmRun         CRuns         CRBI        CWalks     LeagueN
## 6.244860e-01  2.214985e-01  2.186914e-01 -1.500245e-01 4.592589e+01
##    DivisionW        PutOuts        Assists        Errors NewLeagueN
## -1.182011e+02  2.502322e-01  1.215665e-01 -3.278600e+00 -9.496680e+00
```

능선회귀 R 코드 II

```
set.seed(1)
train <- sample(1:nrow(x), nrow(x) / 2)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 0)
plot(cv.out)
bestlam <- cv.out[["lambda.min"]]
bestlam

## [1] 431.0623
```



교차타당성(Cross-Validation, CV)

- 자료(training set) $D_n = \{(x_i, y_i), i = 1, \dots, n\}$ 으로부터 $\hat{y}_i = g_{D_n}(x_i)$ 을 얻었다고 하자. 여기서 $g_{D_n}(x)$ 는 prediction rule이다. 이때 prediction error는

$Err = E(L(y_0, \hat{y}_0))$, 이 때 y_0 는 D_n 과 독립적으로 관찰한 새로운 관측값.

- Err 의 추정량으로 \hat{Err} 을 고려한다.

$$\hat{Err} = \frac{1}{n} \sum_{j=1}^n L(y_{oj}, \hat{y}_{oj}), \quad \hat{y}_{oj} = g_{D_n}(x_{oj}).$$

- CV는 기존자료를 이용한 \hat{Err} 의 모방이다.

$$CV = \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{y}_{(i)}), \quad \hat{y}_{(i)} = (x_i, y_i) \text{ 없이 예측한 값(predicted value).}$$

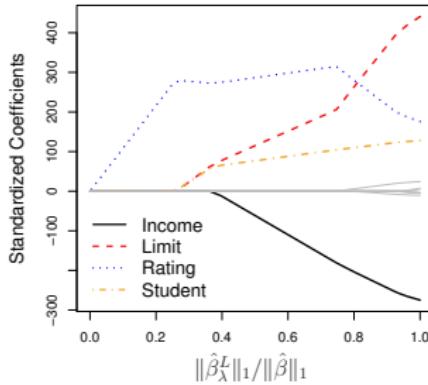
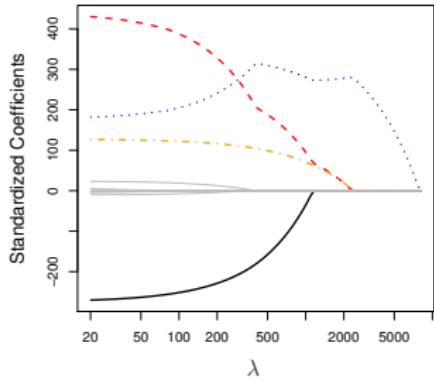
라쏘(Least Absolute Shrinkage and Selection Operator, Lasso)

- 라쏘 추정량은 변수의 회귀계수가 작을 경우 그 값을 정확히 0으로 놓는 성질이 있어, 변수선택의 효과가 있다.
- 라쏘는 변수선택과 축소를 동시에 한다.
- 라쏘추정량 $\hat{\beta}_\lambda^L$ 은

$$\sum_{i=1}^n (y_i - (\beta_0 + \sum_{j=1}^p \beta_j x_{ij}))^2 + \lambda \sum_{j=1}^p |\beta_j| = RSS + \lambda \sum_{j=1}^p |\beta_j|$$

을 최소화하는 β 값으로 정의된다.

λ 의 변화에 따른 추정량의 변화



능선회귀와 라쏘 구체화

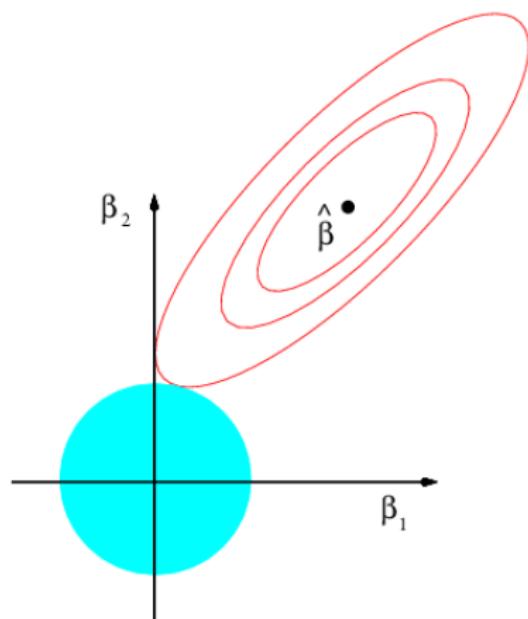
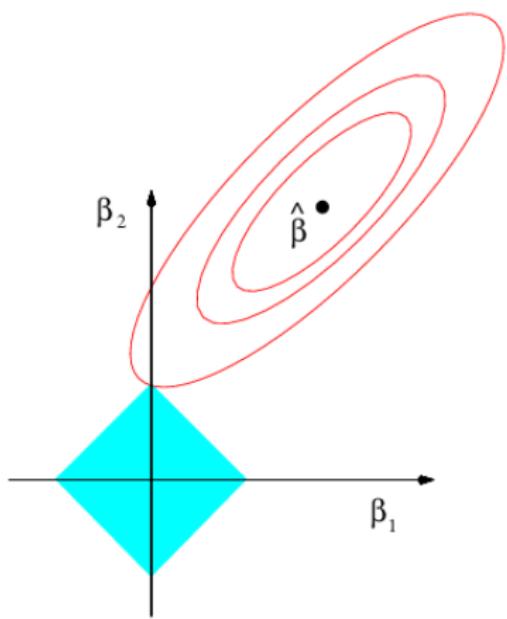
- 능선회귀의 구체화

능선회귀추정량 $\hat{\beta}_\lambda^R$ 는 적당한 s 에 대해 $\sum_{j=1}^p \beta_j^2 \leq s$ 조건하에서 $RSS = \sum_{i=1}^n (y_i - (\beta_0 + \sum_{j=1}^p \beta_j x_{ij}))^2$ 를 최소화하는 β 와 같다.

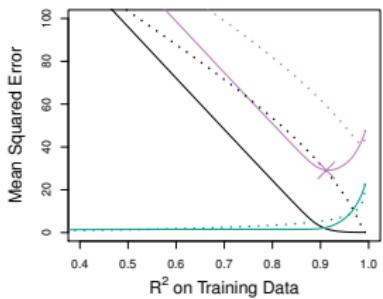
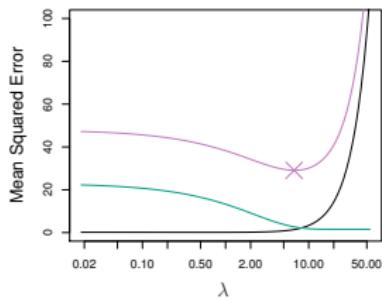
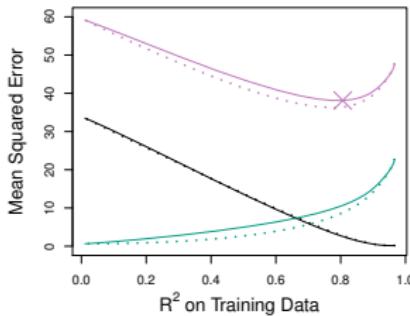
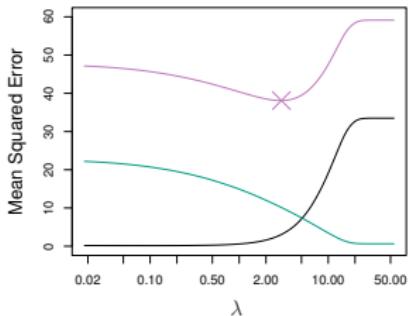
- 라쏘의 구체화

라쏘회귀추정량 $\hat{\beta}_\lambda^L$ 는 적당한 s 에 대해 $\sum_{j=1}^p |\beta_j| \leq s$ 조건하에서 $RSS = \sum_{i=1}^n (y_i - (\beta_0 + \sum_{j=1}^p \beta_j x_{ij}))^2$ 를 최소화하는 β 와 같다.

기하학적 해석



라쏘와 능선회귀 비교



특별한 경우: $X = I_p$, $n = p$, $\beta_0 = 0$

최소제곱추정량

$$RSS = \sum_{j=1}^p (y_j - \beta_j)^2 \quad \text{라쏘}$$

$$\hat{\beta}_j = y_j, \quad j = 1, 2, \dots, p.$$

능선회귀

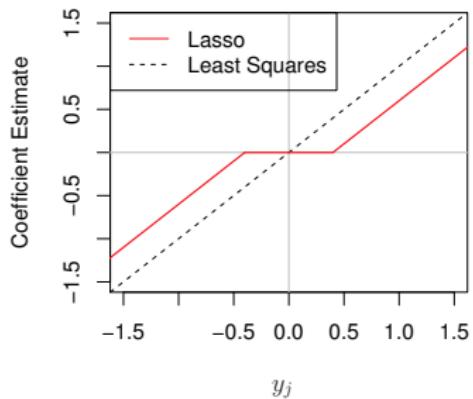
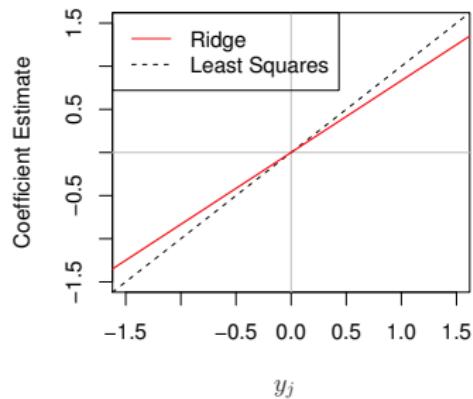
$$\sum_{j=1}^p (y_j - \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

$$\hat{\beta}_j^R = \frac{y_j}{1 + \lambda}, \quad j = 1, 2, \dots, p$$

$$\sum_{j=1}^p (y_j - \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2, & y_j > \lambda/2 \\ y_j + \lambda/2, & y_j < -\lambda/2 \\ 0, & |y_j| \leq \lambda/2, \end{cases}$$

축소의 형태



능선회귀와 라쏘: 베이지안 해석

- β 의 사전분포가

$$\pi(\beta) \propto e^{-\frac{\lambda}{\sigma^2} \sum_{j=1}^p \beta_j^2}$$

인 경우, 사후분포는

$$\pi(\beta|y) \propto e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 + \sum_{j=1}^p \beta_j x_{ij})^2} \times e^{-\frac{\lambda}{\sigma^2} \sum_{j=1}^p \beta_j^2}$$

가 된다. β 의 최대사후분포추정량(MAP)이 능선회귀추정량이 된다.

- β 의 사전분포가

$$\pi(\beta) \propto e^{-\frac{\lambda}{\sigma^2} \sum_{j=1}^p |\beta_j|}$$

인 경우, 사후분포는

$$\pi(\beta|y) \propto e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 + \sum_{j=1}^p \beta_j x_{ij})^2} \times e^{-\frac{\lambda}{\sigma^2} \sum_{j=1}^p |\beta_j|}$$

가 된다. β 의 최대사후분포추정량(MAP)이 라쏘추정량이 된다.

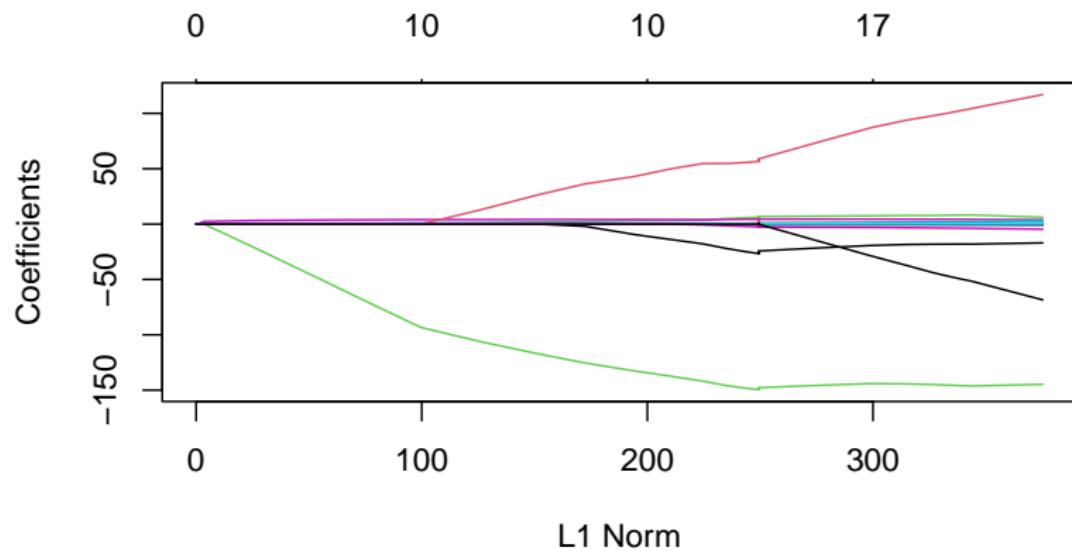
라쏘 R 코드 I

```
Hitters <- na.omit(Hitters)
x <- model.matrix(Salary ~ ., Hitters)[, -1]
y <- Hitters[["Salary"]]

library(glmnet)
grid <- 10 ^ seq(10, -2, length = 100)
lasso.mod <- glmnet(x[train, ], y[train], alpha = 1, lambda = grid)
plot(lasso.mod)

## Warning in regularize.values(x, y, ties, missing(ties), na.rm =
na.rm): collapsing to unique 'x' values
```

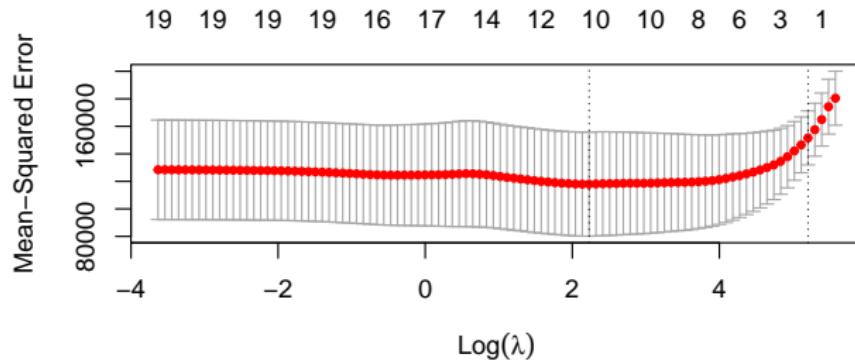
라쏘 R 코드 II



라쏘 R 코드 III

```
set.seed(1)
train <- sample(1:nrow(x), nrow(x) / 2)
cv.out <- cv.glmnet(x[train, ], y[train], alpha = 1)
plot(cv.out)
bestlam <- cv.out[["lambda.min"]]
bestlam

## [1] 9.286955
```



관련 방법

- Adaptive Lasso (Zou, 2006)

$$\hat{\beta}^{AL} = \arg \min(RSS + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j|), \quad \hat{w} = 1/|\hat{\beta}|, \quad \hat{\beta} = \text{OLSE or ridge est.}$$

- Elastic net (Zou and Hastie, 2005)

$$\hat{\beta}^{EN} = \arg \min(RSS + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p |\beta_j|^2).$$

- SCAD (Fan and Li, 2001)

$$\hat{\beta}^S = \arg \min(RSS/2 + \sum_{j=1}^p p_\lambda(|\beta_j|)),$$

$$p_\lambda(|\beta|) = \begin{cases} \lambda |\beta| & \text{if } |\beta| \leq \lambda \\ -\frac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases}$$

회귀분석: 모형, 목적

단순선형회귀, 다중선형회귀: 추론

가변수(지시변수)

모형진단

변수(모형) 선택: 최적부분집합선택, 능선회귀, 라쏘

다항회귀, 비모수회귀, 일반화 가법모형

다항회귀 |

d 차 다항회귀모형

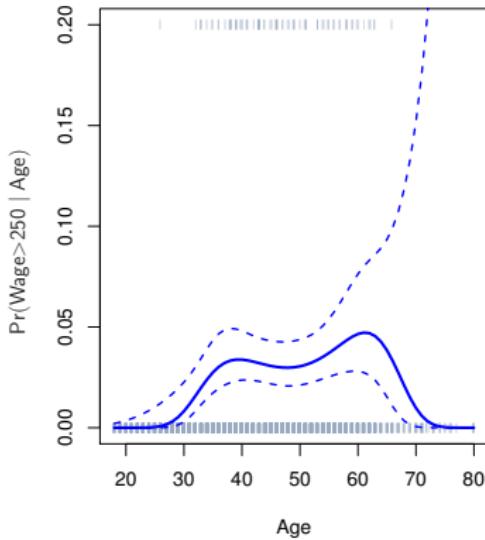
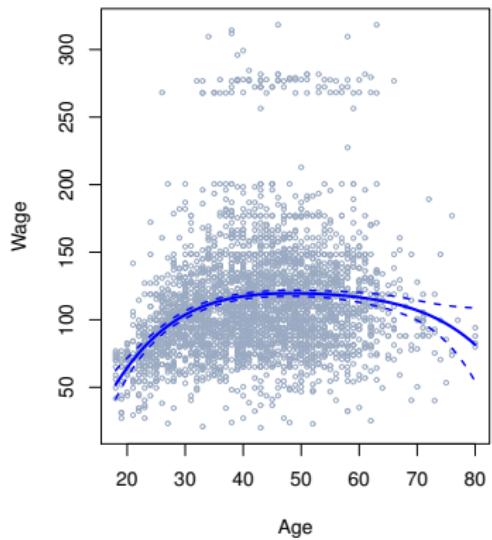
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_d x_i^d + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2), i = 1, 2, \dots, n$$

특성들

1. 가장 간단하게 x 의 비선형 회귀 함수를 표현할 수 있는 방법이다.
2. 4차가 넘어가면 함수의 모양이 너무 유연해져서, (특히 설명변수의 경계영역에서) 이상한 모양이 될 수 있다. 4차 이상의 모형은 잘 쓰지 않는다.

다항회귀 II

Degree-4 Polynomial



계단함수를 이용한 회귀모형 I

x 를 범주형변수들로 변환

$c_1 < c_2 < \dots < c_K$ 를 x 의 범위의 구분점들이라 하자. 다음과 같이 범주형 변수들을 정의한다.

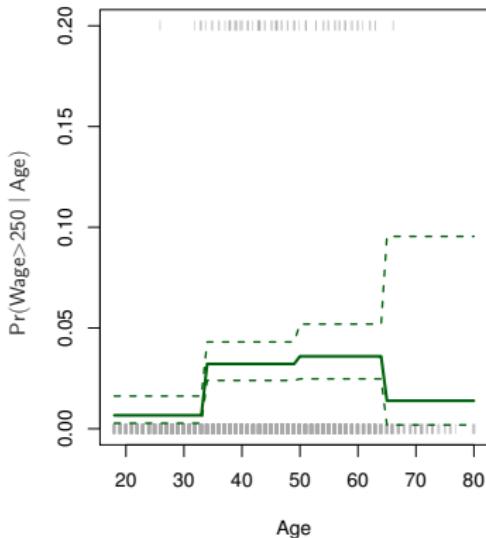
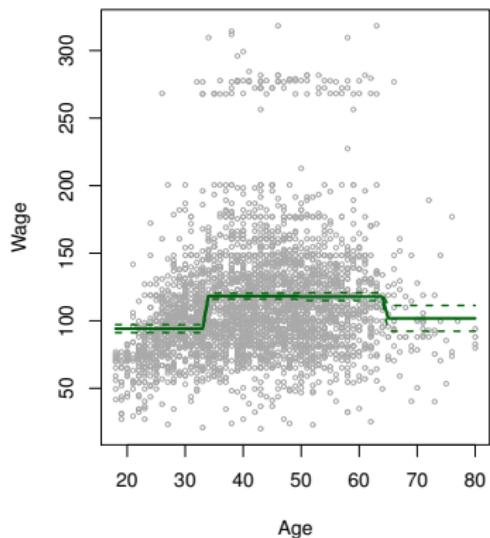
$$\begin{aligned} C_0(x) &= I(x < c_1) \\ C_1(x) &= I(c_1 \leq x < c_2) \\ &\vdots \\ C_{K-1}(x) &= I(c_{K-1} \leq x < c_K) \\ C_K(x) &= I(c_K \leq x). \end{aligned}$$

계단함수를 이용한 회귀모형

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \dots + \beta_K C_K(x_i) + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2), \quad i = 1, 2, \dots, n$$

계단함수를 이용한 회귀모형 II

Piecewise Constant



계단함수를 이용한 회귀모형 R 코드 I

```
table(cut(Wage[["age"]], 4))

##
## (17.9,33.5]  (33.5,49]   (49,64.5] (64.5,80.1]
##          750       1399       779        72

fit <- lm(wage ~ cut(age, 4), data = Wage)
coef(summary(fit))

##
##                               Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)           94.158392   1.476069 63.789970 0.000000e+00
## cut(age, 4)(33.5,49]  24.053491   1.829431 13.148074 1.982315e-38
## cut(age, 4)(49,64.5]  23.664559   2.067958 11.443444 1.040750e-29
## cut(age, 4)(64.5,80.1] 7.640592   4.987424  1.531972 1.256350e-01
```

노트.

`cut(age, 4)`는 변수 `age`를 4개의 영역으로 나눠준다. 격자의 분리점을 정하고 싶으면 `breaks` 옵션을 쓰면 된다.

기저함수(basis function)를 이용한 회귀모형 I

기저함수를 이용한 회귀모형

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \dots + \beta_K b_K(x_i) + \epsilon_i, \quad \epsilon_i \sim (0, \sigma^2), \quad i = 1, 2, \dots, n$$

여기서, b_1, \dots, b_K 는 기저함수이다.

참고사항

1. 다항회귀모형과 계단함수를 이용한 회귀모형도 기저함수를 이용한 회귀모형으로 볼 수 있다.
2. 이 외에도 푸리에 기저(Fourier basis), 웨이블렛 기저(wavelet basis) 등을 사용할 수 있다.

회귀스플라인(regression spline) I

1. 다항회귀모형의 문제점 : 다항회귀모형을 이용해서 모형의 유연성을 높이려면 다항식의 차수를 높여야 한다. 그런데 차수를 높이면 원치 않는 모양의 이상한 회귀함수 모양이 나타날 수 있다. 이는 어떤 관측치가 멀리 떨어진 곳의 함수 추정량에도 영향을 미치기 때문이다.
2. 계단함수의 문제점 : 이를 극복하기 위해서는 계단함수 같은 기저함수를 이용하면 된다. 그런데, 계단함수는 연속함수에 적용할 수 없다.
3. 스플라인 모형: 위의 문제점을 해결하기 위해 매듭(knot)으로 나누어진 구간에 낮은 차원의 다항식 모형을 적합하는 것을 고려한다. 조각별 다항식(piecewise polynomial)을 스플라인이라 한다.

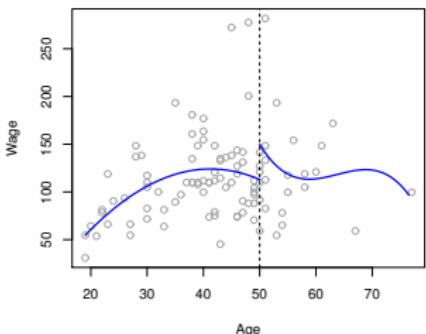
회귀스플라인(regression spline) II

회귀스플라인

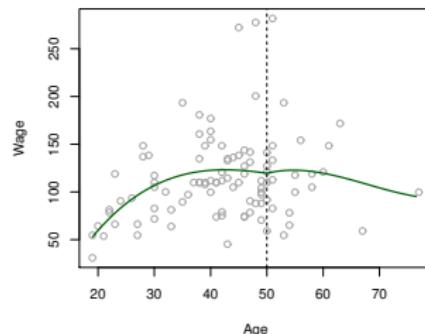
1. 차수가 d 이고 매듭이 ξ_1, \dots, ξ_K 인 스플라인이란 매듭으로 이루어진 각 구간에서 차수가 d 인 다항식이고 $d - 1$ 차 도함수가 연속인 함수를 말한다. 이것들을 회귀스플라인이라고 한다.
2. 3차 스플라인(cubic spline) : 차수가 $d = 3$ 인 스플라인. 각 구간이 3차 다항식이고, 2차 도함수가 연속. 매듭의 불연속성이 사람의 눈에 보이지 않은 가장 차수가 낮은 스플라인. 이 이상 높은 차수는 사용할 필요가 없다.
3. 1차 스플라인(linear spline) : $d = 1$ 인 스플라인.
4. 보통 $d = 0, 1, 3$ 이 가장 많이 사용된다.

회귀스플라인(regression spline) III

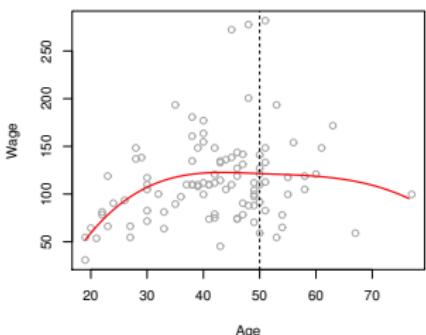
Piecewise Cubic



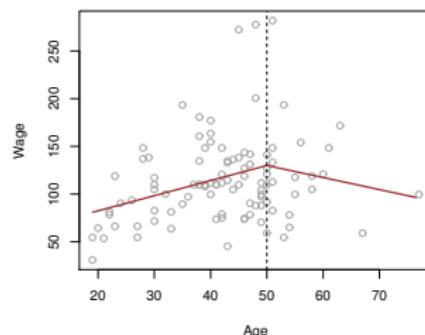
Continuous Piecewise Cubic



Cubic Spline



Linear Spline



회귀스플라인(regression spline) IV

회귀스플라인의 기저 표현

차수가 d 이고 매듭이 ξ_1, \dots, ξ_K 인 스플라인 모형은

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d \\&\quad + \beta_{d+1} h(x_i, \xi_1) + \beta_{d+2} h(x_i, \xi_2) + \cdots + \beta_{d+K} h(x_i, \xi_K) + \epsilon_i\end{aligned}$$

과 같이 나타낼 수 있다. 즉, 위 모형의 회귀함수는 매듭으로 구성되는 각 구간에서 d 차 다항식이고, 각 매듭에서 $d - 1$ 차 도함수가 연속이다.
여기서

$$h(x, \xi) := (x - \xi)_+^d = \begin{cases} (x - \xi)^d, & x > \xi \\ 0, & O.W. \end{cases}$$

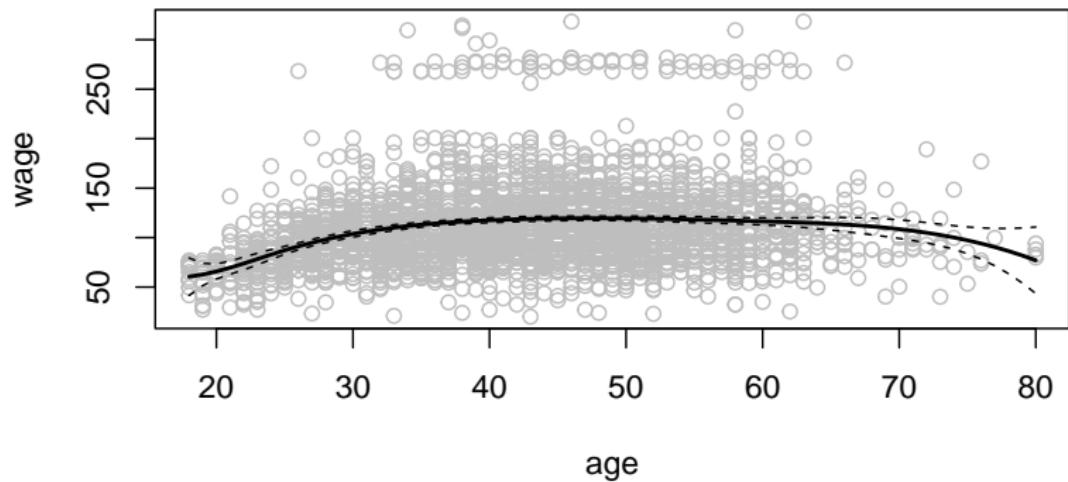
회귀스플라인 R 코드 |

```
agelims <- range(Wage[["age"]])
age.grid <- seq(from = agelims[1], to = agelims[2])

library(splines)
fit <- lm(wage ~ bs(age, knots = c(25, 40, 60)), data = Wage)
pred <- predict(fit, newdata = list(age = age.grid), se = TRUE)

plot(Wage[["age"]], Wage[["wage"]], col = "gray", xlab = "age",
     ylab = "wage")
lines(age.grid, pred[["fit"]], lwd = 2)
lines(age.grid, pred[["fit"]] + 2 * pred[["se.fit"]], lty = "dashed")
lines(age.grid, pred[["fit"]] - 2 * pred[["se.fit"]], lty = "dashed")
```

회귀스플라인 R 코드 II

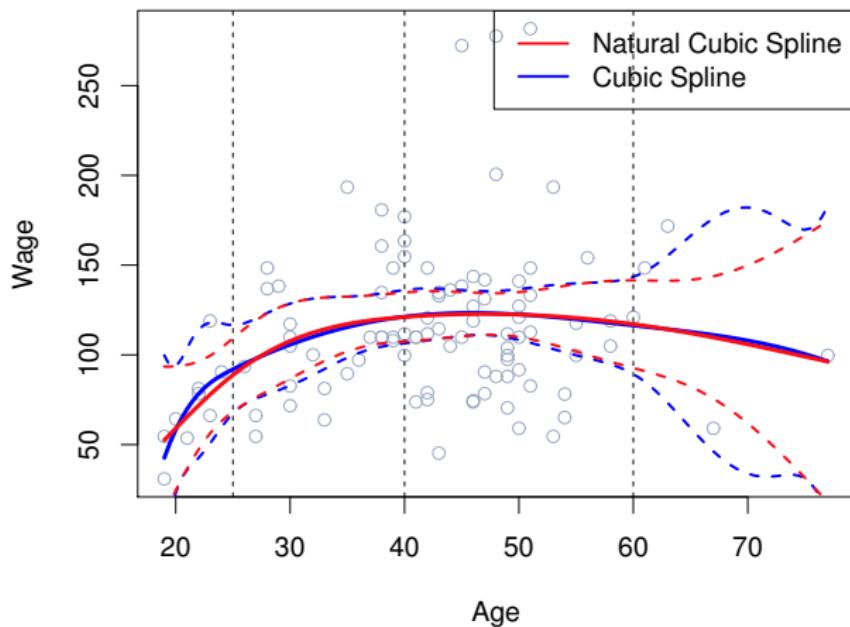


자연 스플라인(natural spline) I

자연 스플라인(natural spline)

1. 자연스플라인은 3차 스플라인에서 $x < \xi_1$ 혹은 $x > \xi_K$ 일 때, 1차 다항식으로 대치한 것을 말한다.
2. 회귀스플라인은 경계 부근에서 큰 분산을 갖는다. 자연스플라인은 경계부근에서 안정된 분산을 갖게 한다.

자연 스플라인(natural spline) II



자연 스플라인(natural spline) III

자연 3차 스플라인의 기저표현 K 개의 기저는 다음과 같다.

$$N_1(x) = 1, \quad N_2(x) = x, \quad N_{k+2}(x) = d_k(x) - d_{K-1}(x), \quad k = 1, 2, \dots, K-2.$$

여기서,

$$d_k(x) := \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k}$$

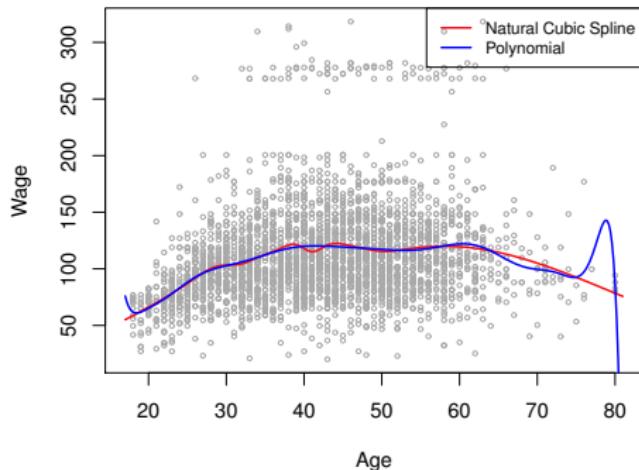
이다. N_k 의 2차와 3차 도함수는 $x > \xi_K$ 에서 0이다.

매듭의 위치와 개수를 정하는 법

1. 위치: 매듭이 촘촘하게 있는 구간에서 스플라인은 더 유연하다. 대개의 경우 설명변수의 분위수를 이용해 정한다. K 개의 매듭의 위치를 정하는 경우, $x_{(n \frac{i}{K+1})}$, $i = 1, 2, \dots, K$ 에 정한다.
2. 개수: 매듭의 개수 K 는 교차검증방법을 이용해서 정한다.

자연 스플라인(natural spline) IV

다항회귀모형과 스플라인모형의 비교



차수가 15인 다항회귀모형과 자연스플라인 모형이 비교: 다항회귀인 경우 경계부근에서 원치않는 효과가 있음.

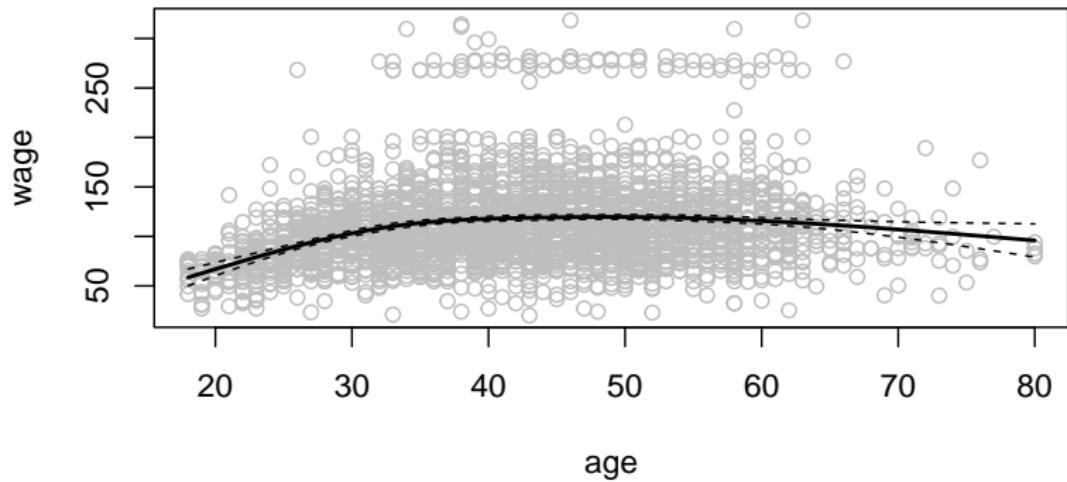
자연스플라인 R 코드 |

```
agelims <- range(Wage[["age"]])
age.grid <- seq(from = agelims[1], to = agelims[2])

library(splines)
fit2 <- lm(wage ~ ns(age, df = 4), data = Wage)
pred2 <- predict(fit2, newdata = list(age = age.grid), se = TRUE)

plot(Wage[["age"]], Wage[["wage"]], col = "gray", xlab = "age",
     ylab = "wage")
lines(age.grid, pred2[["fit"]], lwd = 2)
lines(age.grid, pred2[["fit"]] + 2 * pred2[["se.fit"]], lty = "dashed")
lines(age.grid, pred2[["fit"]] - 2 * pred2[["se.fit"]], lty = "dashed")
```

자연스플라인 R 코드 II



함수 `ns`는 자연스플라인기저로 생성되는 계획(모형)행렬을 생성한다.
자연스플라인모형은 이 기저에 `lm` 함수를 적용한다.

평활스플라인(smoothing spline) I

정의

$$\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int g''(x)^2 dx$$

를 최소화 하는 함수 g 를 평활스플라인이라고 한다.

1. λ : 조율파라미터(tuning parameter)
2. $\sum_{i=1}^n (y_i - g(x_i))^2$: 손실함수(loss function)
3. $\int g''(x)^2 dx$: 벌점(penalty). 이차도함수는 함수의 구불구불한 정도를 나타낸다.

평활스플라인(smoothing spline) II

평활스플라인은 자연3차스플라인

평활스플라인은 매듭이 중복되지 않는(distinct) x_1, \dots, x_n 인 자연3차스플라인이다.

유효자유도(effective degree of freedom)

g 가 회귀함수일 때, $(g(x_1), \dots, g(x_n))'$ 의 추정값을

$$\hat{g}_\lambda = S_\lambda y$$

와 같이 나타낼 수 있다. 이 때, 유효자유도는

$$df_\lambda := \text{tr}(S_\lambda)$$

로 정의된다.

평활스플라인(smoothing spline) III

조율파라미터의 결정

λ 의 값은 교차검증을 이용해 결정할 수 있다. 그런데, 평활스플라인에서 하나빼기교차검정오차는 다음과 같이 수식이 알려져 있어, 빠르게 계산할 수 있다.

$$RSS_{cv}(\lambda) := \sum_{i=1}^n (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^n \left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - (S_\lambda)_{ii}} \right]^2$$

여기서, $\hat{g}_\lambda^{(-i)}(x)$ 는 i 번째 관측치를 제외하고 구축한 함수의 추정치이고, $(S_\lambda)_{ii}$ 는 S_λ 의 i 번째 대각행렬이다.

평활스플라인 R 코드 |

```
agelims <- range(Wage[["age"]])
age.grid <- seq(from = agelims[1], to = agelims[2])

fit <- smooth.spline(Wage[["age"]], Wage[["wage"]], df = 16)
fit2 <- smooth.spline(Wage[["age"]], Wage[["wage"]], cv = TRUE)

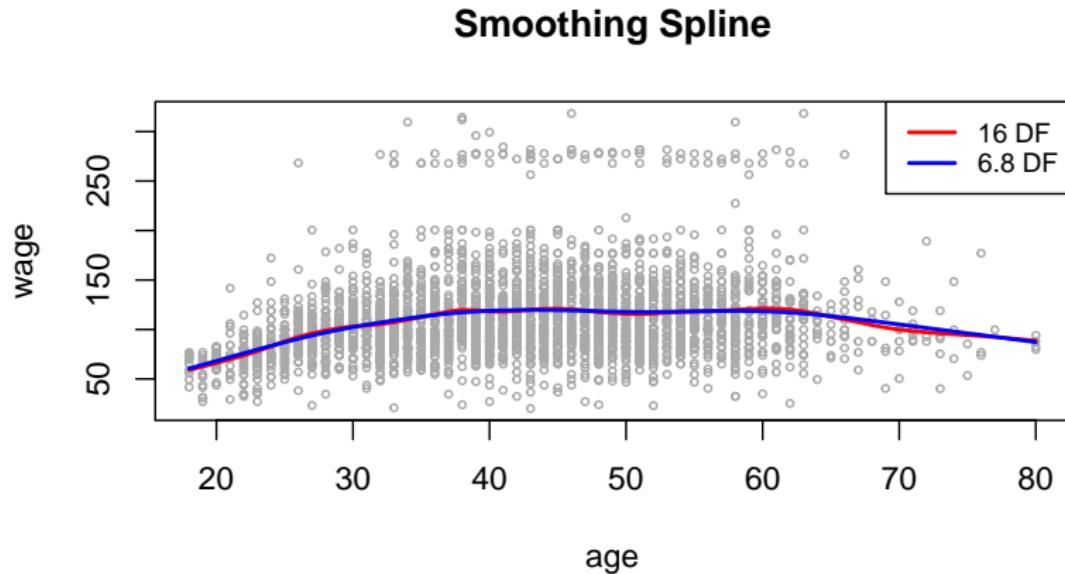
## Warning in smooth.spline(Wage[["age"]], Wage[["wage"]], cv = TRUE):
## cross-validation with non-unique 'x' values seems doubtful

fit2[["df"]]

## [1] 6.794596

plot(Wage[["age"]], Wage[["wage"]], xlim = agelims, cex = 0.5,
      col = "darkgrey", xlab = "age", ylab = "wage")
title("Smoothing Spline")
lines(fit, col = "red", lwd = 2)
lines(fit2, col = "blue", lwd = 2)
legend("topright", legend = c("16 DF", "6.8 DF"), col = c("red", "blue"),
       lty = 1, lwd = 2, cex = 0.8)
```

평활스플라인 R 코드 II



평활스플라인은 `smooth.spline` 함수를 이용해서 적합한다. `fit`은 $df=16$ 이 주어져서 조율파라미터가 결정이 되었다. `fit2`는 교차검정을 이용해서 조율파라미터가 결정되었다.

국소회귀(local regression) I

$x = x_0$ 에서 \hat{g} 의 계산

x_0 에 가까운 $s = k/n$ 개의 관측치에 가중치 $K(x_i, x_0)$ 를 주고 1차 회귀모형을 적합한다. 즉,

$$\sum_{i=1}^n K(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i)^2$$

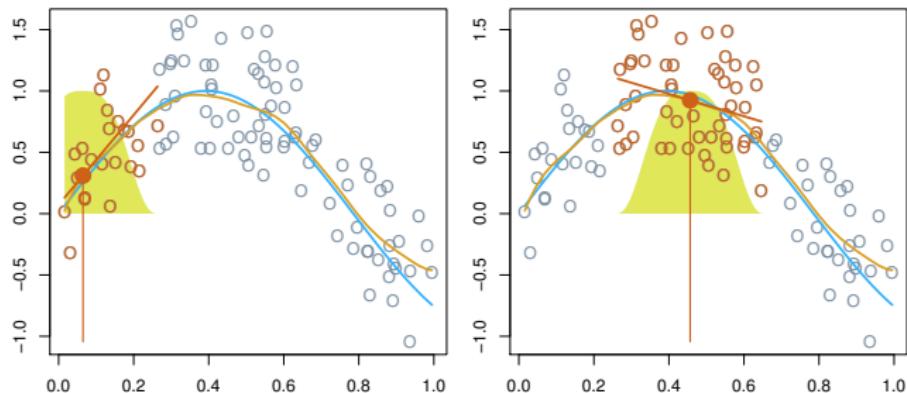
을 최소화하여 β_0 와 β_1 을 얻는다. 예측치는

$$\hat{f}(x_0) := \hat{\beta}_0 + \hat{\beta}_1 x_0$$

가 된다. 위에서 선형회귀 대신 p 차 회귀모형을 쓸 수 있다. 위에서 s 를 펼침(span)이라 부른다.

국소회귀(local regression) II

Local Regression

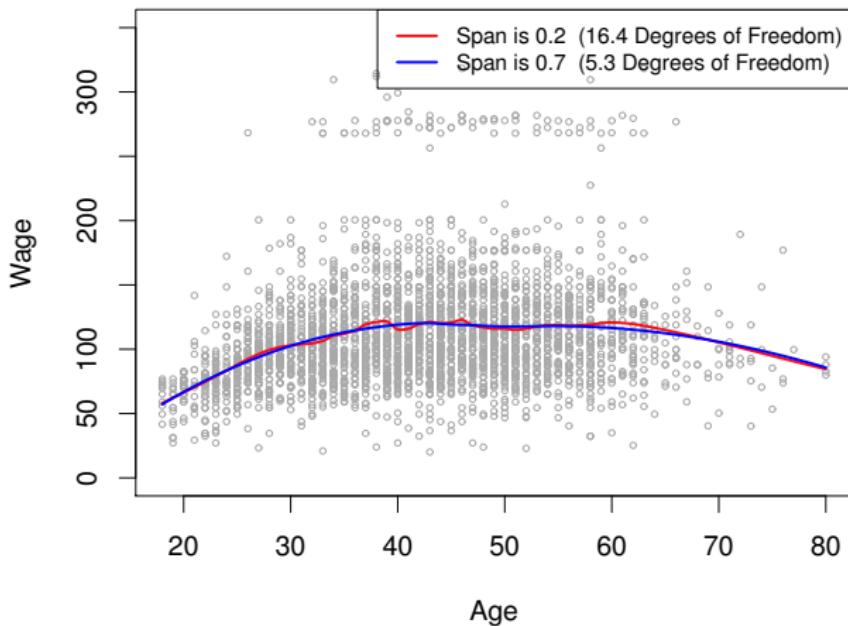


펼침 s 의 결정

국소회귀를 적합할 때, 가중치 K , 회귀모형의 차수, 펼침 s 를 결정해야 한다. 이 중 s 의 결정이 제일 중요하다. s 는 모형의 유연성을 결정하고, 교차검증으로 결정할 수 있다.

국소회귀(local regression) III

Local Linear Regression



국소회귀(local regression) IV

국소회귀모형의 확장

1. 일부 설명변수에는 국소회귀모형을 쓰고, 일부 설명변수에는 전역회귀모형을 쓸 수 있다. 시간이 설명변수일 때, 시간을 국소회귀모형으로 쓰면 시간변동계수모형(time varying coefficient model)이 된다.
2. 국소회귀모형은 고차원 자료에 확장이 잘 된다.
3. 회귀모형의 차수 p 가 3, 4보다 크면 성능이 나빠질 수 있다.

국소회귀 R 코드 |

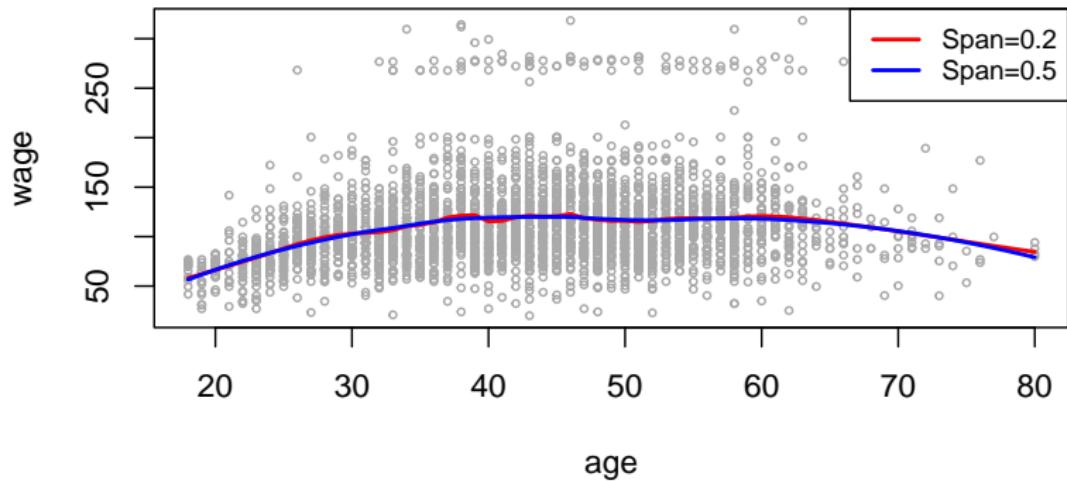
```
agelims <- range(Wage[["age"]])
age.grid <- seq(from = agelims[1], to = agelims[2])

fit <- loess(wage ~ age, span=0.2, data = Wage)
fit2 <- loess(wage ~ age, span=0.5, data = Wage)

plot(Wage[["age"]], Wage[["wage"]], xlim = agelims, cex = 0.5,
     col = "darkgrey", xlab = "age", ylab = "wage")
title("Local Regression")
lines(age.grid, predict(fit, data.frame(age = age.grid)), col = "red",
      lwd = 2)
lines(age.grid, predict(fit2, data.frame(age = age.grid)), col = "blue",
      lwd = 2)
legend("topright", legend = c("Span=0.2", "Span=0.5"),
       col = c("red", "blue"), lty = 1, lwd = 2, cex = 0.8)
```

국소회귀 R 코드 II

Local Regression



일반화가법모형(generalized additive models) I

모형

여러 개의 설명변수가 있을 때, 각 변수에 1변수 회귀함수를 적용하는 것을 말한다.

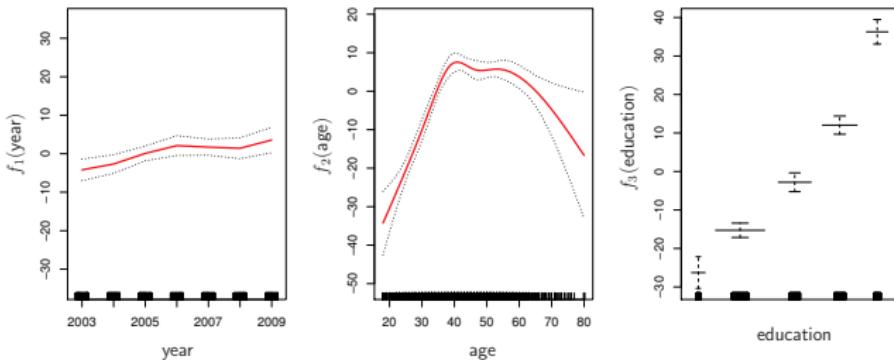
$$y_i = \beta_0 + f_1(x_{i1}) + \dots + f_p(x_{ip}) + \epsilon_i, \quad i = 1, 2, \dots, n.$$

연봉자료의 예

$$wage = \beta_0 + f_1(year) + f_2(age) + f_3(education) + \epsilon$$

일반화가법모형(generalized additive models) II

<HS HS <Coll Coll >Coll



역적합(backfitting)

설명변수 하나씩 돌아가면서 나머지 설명변수와 그의 적합된 함수를 고정한 후, 한 개의 변수에 관해서만 회귀함수를 구하는 방법

일반화가법모형(generalized additive models) III

일반화가법모형의 특징

1. 각 설명변수에 비선형 함수 f_j 를 적용하기 때문에, 설명변수를 변환할 필요가 없다.
2. 설명변수의 비선형 함수는 예측 성능을 향상시킬 수 있다.
3. 모형이 가법이기 때문에, 다른 설명변수들을 고정시켰을 때, 한 설명변수의 효과가 f_j 이다.
4. f_j 의 유연성은 자유도로 요약된다.
5. 가법모형이기 때문에 교호작용을 표현하지 못한다. 하지만, 교호작용을 표현하고 싶으면 $f(x_i, x_j)$ 를 적합하면 된다.

이산형 반응변수와 가법모형

반응변수가 이산형일 때도 가법모형을 적용할 수 있다.

$$\log \left(\frac{p(x)}{1 - p(x)} \right) = \beta_0 + f_1(x_1) + \dots + f_p(x_p)$$

일반화 가법모형 R 코드 I

```
gam1 <- lm(wage ~ ns(year, 4) + ns(age, 5) + education, data = Wage)
```

예측변수 모두가 자연스플라인 기저로 구성된 가법모형을 적용하려면 lm을 이용하면 된다. 가법모형이 큰 선형모형일 뿐이기 때문이다.

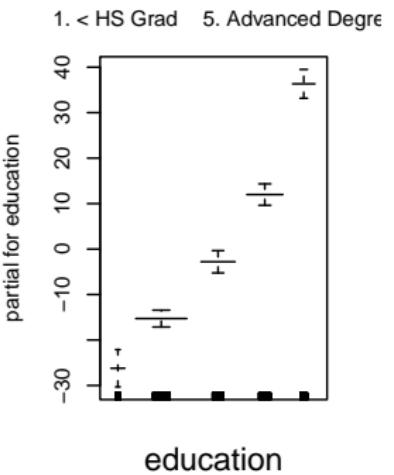
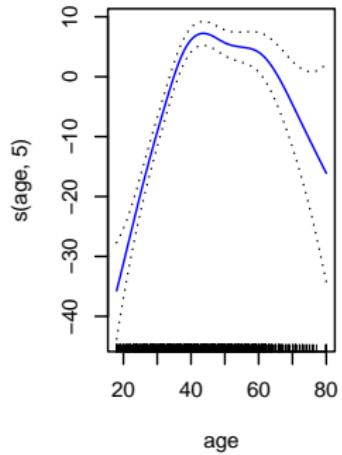
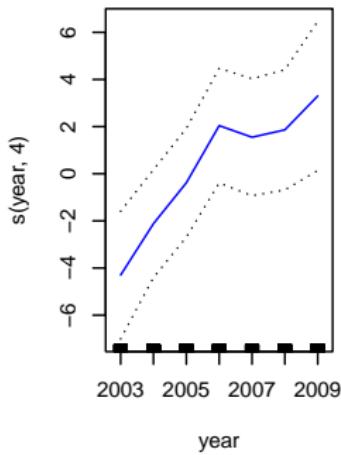
일반화 가법모형 R 코드 II

```
library(gam)

## Warning: package 'gam' was built under R version 4.3.2
## Loading required package: foreach
## Warning: package 'foreach' was built under R version 4.3.2
## Loaded gam 1.22-3

gam.m3 <- gam(wage ~ s(year, 4) + s(age, 5) + education, data = Wage)
par(mfrow = c(1, 3))
plot(gam.m3, se = TRUE, col = "blue")
```

일반화 가법모형 R 코드 III



하나 이상의 예측변수에 평활스플라인을 적용하려면 gam 패키지의 함수 `gam()`이 필요하다. 함수 `gam` 안의 함수 `s`는 평활스플라인의 기저를 생성하는 함수로 `gam` 패키지 안에 있는 함수이다.

참고문헌

이 강의자료는 서울대학교 통계학과 오희석 교수의 2022년 강의자료를 바탕으로 하였다. 그림의 출처는 아래의 책이다.

- [1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. 2nd Edition, Springer, 2021.