통계적 모형선택

Jong-June Jeon [1]

[1]Department of Statistics
University of Seoul

August 29, 2017

T-test

- T 검정은 스튜던트 t통계량의 분포를 귀무가설하에서 살펴봄으로써 가설의 기각 여부를 결정하는 의사결정 모형임
- 검정: $X_i \sim_{iid} N(\mu, \sigma^2)$ 이라고 가정하고, 귀무가설과 대립가설을 아래와 같이 놓자.

$$H : \mu = \mu_0 \qquad K : \mu > \mu_0.$$

귀무가설 즉, $\mu = \mu_0$하에서

$$(\bar{X} - \mu_0)/(S/\sqrt{n}) \sim T(n-1),$$

임이 알려져 있고 유의 수준 α에 따른 기각역이 결정된다 (Neyman-Pearson lemma).

T-test

- T-검정의 모형 가정:

$$X_i = \mu + \epsilon_i$$

  단, $\epsilon_i \sim_{iid} N(0, \sigma^2)$
- 가정에 위배되는 경우
  - X_i의 분포가 동일하지 않을 때.
  - X_i의 분산이 존재하지 않을 때 (꼬리가 두꺼운 분포) 혹은 X_i 분포가 대칭이 아닌경우;
  - 데이터가 독립이 아닌 경우.
- T-검정이 주는 기각역을 이용하여 의사결정을 했을 때, 이미 정해진 유의수준 α (1종 오류)과 다른 의사결정을 하게된다.

ANOVA

- 분산 분석은 (Analysis of variance: ANOVA) 여러 개의 모집단의 평균을 비교, 검정하는 방법이다.
- 1원배치 분산분석의 가정: $X_{ij} \sim_{iid} N(\mu_j, \sigma^2)$ for $j = 1, \cdots, p$ (p개의 처리)

$$\mu_1 = \mu_2 = \cdots = \mu_p$$

귀무가설 하에서

$$F = \frac{\text{급간 분산}}{\text{급내 분산}} \sim F(p-1, n-p-1)$$

임을 이용한다.

ANOVA

- 1원배치 분산분석은 j번째 처리에 대한 반응변수가

$$X_{ij} = \mu_j + \epsilon_{ij}$$

  ($\epsilon_{ij} \sim_{iid} N(0, \sigma^2)$)임을 가정한다.
- 가정에 위배되는 경우:
  - X_{ij}가 정규분포를 따르지 않는 경우;
  - X_{ij}의 분산이 이질적(heterogeneous)인 경우;
  - X_{ij}가 독립이 아닌 경우 (cluster effect);
- 검정이 주는 기각역을 이용하여 의사결정을 했을 때, 이미 정해진 유의수준 α (1종 오류)과 다른 의사결정을 하게된다.

- 2원 배치 분산분석의 가정:

$$X_{ijk} = \alpha + \mu_j + \gamma_k + \epsilon_{ijk}$$

  ($\epsilon_{ijk} \sim_{iid} N(0, \sigma^2)$)임을 가정한다.
- 가정에 위배되는 경우:
  - $EX_{ijk} \neq \alpha + \mu_j + \gamma_k$ (교호작용: interaction effects)
  - X_{ijk}의 분산이 이질적인 경우;
  - 데이터가 독립이 아닌 경우;
- 검정이 주는 기각역을 이용하여 의사결정을 했을 때, 이미 정해진 유의수준 α (1종 오류)과 다른 의사결정을 하게된다.

예제

잘못된 가정하의 t검정 결과

- $X_1 = 0$이라 하고 $X_{i+1} = 0.9X_i + \epsilon_i$ $(\epsilon_i \sim_{iid} N(0,1))$
- 위 모형에서는 $\mathrm{Cov}(X_{i+1}, X_i) \neq 0$ 로 관측치가 독립이 아니다. 단 $\mathrm{E}(X_i) = 0$ 는 성립한다.
- X_i $(i = 1, \cdots, 20)$ 를 생성하고 평균에 대한 t검정을 실시한다.
- 1000번의 반복실험을 통해 유의수준 α에서의 t검정의 결과와 실제 얻어진 1종오류를 비교한다.

예제

Rcode

Regression

- 선형회귀모형은 반응변수(Y) 와 설명변수 X의 관계를 모형화한 통계모형이다.
- 설명변수가 1개인 $X \in \mathbb{R}$인 선형회귀모형을 알아보자
- 모형의 가정:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

단, $\epsilon_i \sim_{iid} N(0, \sigma^2)$

Regression

- 선형회귀모형은 다음과 같은 관계를 만족한다.

$$\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$$

- 모형 가정의 위배;
  - $\mathrm{E}(Y_i|X_i) \neq \beta_0 + \beta_1 X_i$
  - $\mathrm{E}(Y_i|X_i)$이 존재하지 않는 경우.

Regression

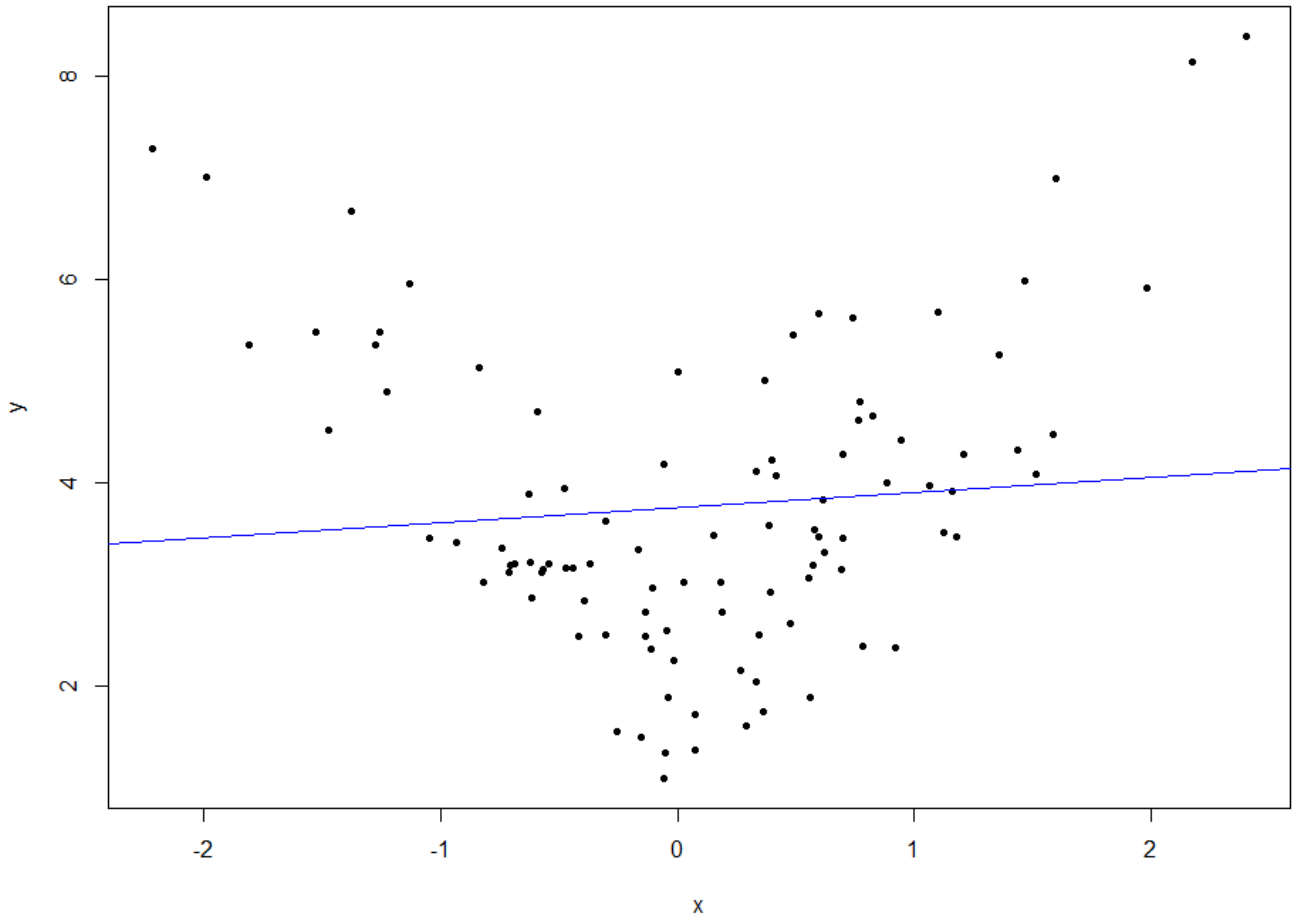
- $E(Y_i|X_i) \neq \beta_0 + \beta_1 X_i$인 예
  - $E(Y_i|X_i) = \beta_0 + \beta_1 X_i^2$
  - $E(Y_i|X_i) = f(X_i)$ 단, $f : \mathbb{R} \mapsto \mathbb{R}$.
  - ϵ_i 과 X_i 상관계수가 0이 아닌 경우.
- $E(Y_i|X_i)$가 존재하지 않는 경우?
  - $\epsilon_i \sim_{iid} t(1)$ (자유도가 1인 t분포).

모형가정이 위배된 경우

- $X_i \sim_{iid} N(0,1)$ for $i = 1, \cdots, 100$.
- $Y_i = 3 + X_i^2 + \epsilon_i$ where $\epsilon_i \sim_{iid} N(0,1)$
- $\mathrm{E}(Y|\mathbf{X}_i = x) = 3 + x^2$ 는 x의 선형함수가 아니다.
- 선형회귀분석에서 가정한 모형은 $\mathrm{E}(Y|\mathbf{X}_i = x) = \beta_0 + x\beta_1$ 이므로 모형 공간은 다음과 같이 주어질 것이다.

$$\mathcal{F} = \{f : f(x) = \beta_0 + \beta_1 x, \beta_0, \beta_1 \in \mathbb{R}\}$$

모형가정이 위배된 경우



파란색 선은 잘못된 모형가정하에서의 LSE 혹은 MLE로 구해진 $E(Y|\mathbf{X}_i = x)$에 대한 추정량이다.

다변량선형회귀모형

- $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})' \in \mathbb{R}$
- $Y_i \in \mathbb{R}$
- $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$
- 다항선형회귀모형

$$\begin{aligned} Y_i &= X_{i1}\beta_1 + \cdots + X_{ip}\beta_p + \epsilon_i \\ &= \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i \end{aligned}$$

단, $\epsilon_i \sim_{iid} (0, \sigma^2)$

다변량선형회귀모형

- 실제 모형이 $Y_i = X_{i1} + X_{i2} + \epsilon_i$ 단, $\epsilon_i \sim_{iid} N(0,1)$
- $(X_{i1}, X_{i2})' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 단, $\boldsymbol{\mu} = (0,0)'$이고

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- X_{i1}만 관찰 가능하다고 하자. 회귀분석의 결과를 구해보자.
- 다음과 같은 상황일 것이다.

$$Y_i = X_{i1} + (X_{i2} + \epsilon_i) = X_{i1} + \tilde{\epsilon}_i$$

여기서 $\tilde{\epsilon}_i = X_{i2} + \epsilon_i$이며,

$$\mathrm{Cov}(X_{i1}, \tilde{\epsilon}_i) = \rho$$

예제

```
> set.seed(1)
> x <- sort(rnorm(100))
> y<- 3+x^2 + rnorm(100)
> plot(x, y, pch = 20)
> fit <- lm(y~x)
> abline(a = fit$coefficients[1],
+        b = fit$coefficients[2], col = 'blue'  )
> ytrue <- 3+ x^2
> lines(x, ytrue , lty = 2, col = 'black')
```

예제

```
> library(MASS)
> set.seed(1)
> rho = 0.5
> n = 100 ; mu.vec = c(0,0)
> Sigma.mat <- matrix(c(1,rho,rho,1),2,2)
> x <- mvrnorm(n, mu.vec, Sigma.mat)
> y<- x%*%c(1,1) + rnorm(100)
> fit <- lm(y~x[,1]-1)
```

예제

```
> set.seed(1)
> iter.num = 1000
> coef.vec <- rep(0,iter.num)
> for (i in 1:iter.num)
+ {
+   x <- mvrnorm(n, mu.vec, Sigma.mat)
+   y<- x%*%c(1,1) + rnorm(100)
+   fit <- lm(y~x[,1]-1)
+   coef.vec[i]<- fit$coefficients
+ }
> boxplot(coef.vec, col = 'orange', ylim = c(0,2))
> abline(h = 1, lty = 2, col = 'red')
```

예제

```
> set.seed(1)
> iter.num = 1000
> rho.vec = seq(-0.7, 0.7, by = 0.1 )
> coef.mat <- matrix(0,iter.num, length(rho.vec))
> for (j in 1:length(rho.vec))
+ {
+   rho = rho.vec[j]
+   Sigma.mat <- matrix(c(1,rho,rho,1),2,2)
+   for (i in 1:iter.num)
+   {
+     x <- mvrnorm(n, mu.vec, Sigma.mat)
+     y<- x%*%c(1,1) + rnorm(100)
+     fit <- lm(y~x[,1]-1)
+     coef.mat[i,j]<- fit$coefficients
+   }
+ }
> colnames(coef.mat)<- paste0('rho=',round(rho.vec,2))
> boxplot(coef.mat, col = 'orange', ylim = c(0,2))
> abline(h = 1, lty = 2, col = 'red')
```

로지스틱 회귀분석

- 모형가정:

$$Y_i|X_i \sim \text{Bernoulli}(\theta(X_i))$$

단, $\theta(X_i) = \frac{\exp(\beta_0+\beta_1 X_i)}{(1+\exp(\beta_0+\beta_1 X_i))}$.

Classification

- 로지스틱 회귀모형은 $\Pr(Y_i|X_i) = \theta(X_i)$ 에 대하여 다음과 같은 가정을 한다.

$$\theta(X_i) = (Y_i|X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)},$$

  즉 $\theta(X_i)$의 로짓(logit), $\log \theta(X_i)/(1 - \theta(X_i))$, 이 $\beta_0 + \beta_1 X_i$ 임을 가정한다.
- 가정의 위배;
  - link misspecification:

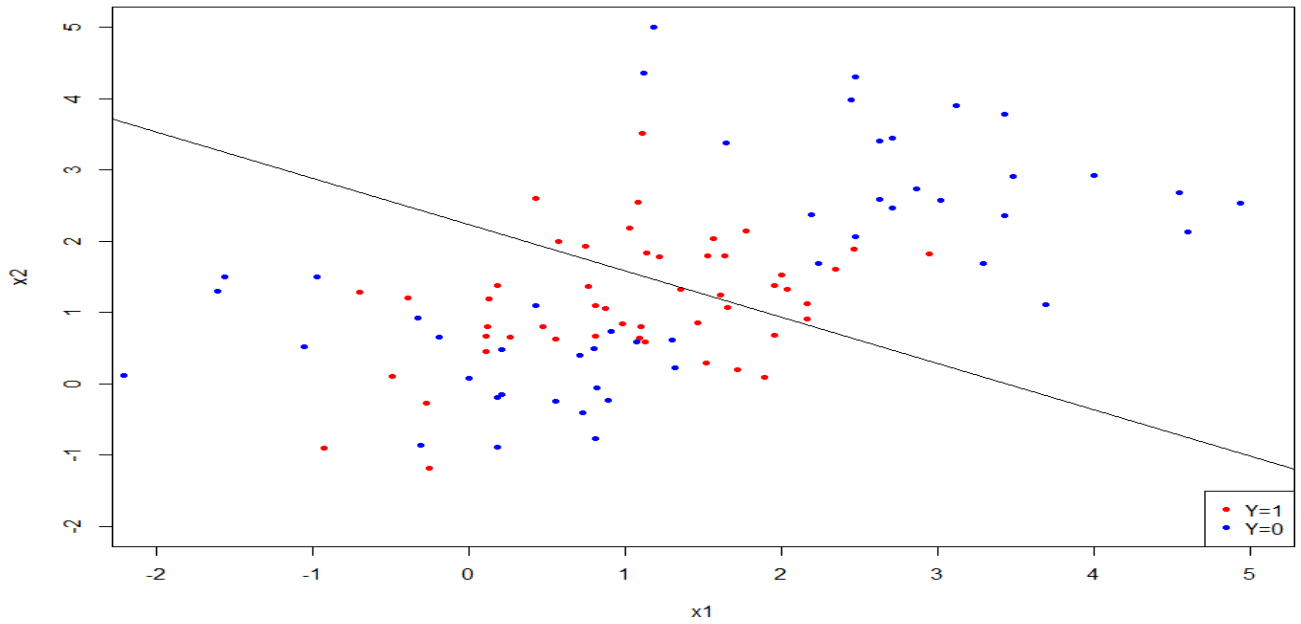
$$\theta(X_i) = \Phi(\beta_0 + \beta_1 X_i)$$

    단, $\Phi(\cdot)$ 표준 정규분포의 cdf.
  - nonlinear model:

$$\theta(X_i) = \frac{\exp(f(X_i))}{1 + \exp(f(X_i))},$$

로지스틱 모형

- 결정경계: $\{\mathbf{x} : f(\mathbf{x}) = 0\}$
- 즉, 결정경계는 $\{\mathbf{x} : \theta(\mathbf{x}) = 0.5\}$과 같다.
- f가 $\mathbf{x}$의 선형함수인 경우에 결정경계는 항상 선형으로 나온다.



Essentially, all models are wrong, but some are useful.

(George E. P. Box)

model and sub-models

- 여기서는 다변량선형회귀모형을 중심으로 모형, 부모형, 모형선택의 개념을 알아보겠다.
  - $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})^T \in \mathbb{R}^p$
  - $Y_i \in \mathbb{R}$
  - 모형:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$$

  - 실제 참모형이 $Y_i = f(\mathbf{X}_i) + \epsilon_i$ (여기서 f는 부드러운 함수)라고 하자
  - 충분히 큰 p에 대하여 $f(\mathbf{X}) \simeq \beta_0 + \sum_{j=1}^{p} \beta_j x^j$ 이므로

$$(Y_i|\mathbf{X}_i) \simeq \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$$

로써 기대값을 근사(approximation)시킬수 있다.

부모형

- 가장 간단한 모형: $y_i = \beta_0 + \epsilon_i$. 즉, 완전모형(full model)의 입장에서는

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

  (단, $\beta_j = 0$ for $j = 1, \cdots, p$)로 주어지는 특별한 경우에 해당한다.
- 1개의 변량에 대응되는 회귀계수만 0이 아닌 부모형도 생각할 수 있다.
  - $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$
  - $\cdots$
  - $y_i = \beta_0 + \beta_p x_{ip} + \epsilon_i$
- 2개의 변량에 대응되는 회귀계수만 0이 아닌 부모형도 생각할 수 있다.
  - $y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \epsilon_i$
  - $\cdots$

부모형의 개수

- 부모형의 개수는 $2^{p+1} - 1$이다.
- $p = 30$인 경우, PC가 1초에 1000번의 회귀모형 적합을 할 수 있다고 가정하자. 이 때, 모든 부모형에 대한 계산시간은 대략 24일이다.
- $p = 50$ 인 경우 대략 7만년이 걸린다.
- p가 큰 경우 부모형에 대한 모형적합을 가능한 시간 내에 다 할 수가 없다.

부모형 구성 전략 (전진법)

- 가장 간단한 모형으로 부터 변량을 하나씩만 추가해 나간다.
- 어떤 변량을 넣을 것인가??
  - 1개의 변량을 가지는 부모형을 적합하는 것을 고려하자.
  - $\min \sum_{i=1}^{n}(y_i - \beta_0 - \beta_j x_{ij})^2$ $(j = 1, \cdots, p)$ 값을 계산한 후에, 가장 작은 값을 가지는 j를 선택하여 모형에 반영한다.
  - $j = 3$이라고 가정하자. 다시 $\min \sum_{i=1}^{n}(y_i - \beta_0 - \beta_3 x_{i3} - \beta_j x_{ij})^2$ for $j \in \{1, \cdots, p\} - \{3\}$를 계산한 후에 가장 작은 값을 가지는 j를 선택한다.
  - 이 과정을 반복한다.
- 이 때 계산 횟수는 p^2이하이며 $p = 1000$에 대해서도 8분 안에 계산을 끝낼 수 있다.
- 여기서 얻어지는 부모형의 개수는 $(p + 1)$개다.

부모형 구성전략(후진법)

- 완전모형에서 하나씩 변수를 빼나간다.
- 어떤 변수를 빼 나갈 것인가?
  - $(p-1)$ 변수를 가지는 모형을 찾는 경우를 고려하자.
  - $\min \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j \notin B} \beta_j x_{ij})^2$ for $B = \{k\}$ 을 계산하고 가장 작은 값을 가지는 k를 선택한다.
  - $k = 3$이라고 하자. 다시 $\min \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j \notin B} \beta_j x_{ij})^2$ for $j \in \{1, \cdots, p\} - \{3, k\}$ 를 계산하고 가장 작은값을 갖는 $k \neq 3$ 를 선택한다.
  - 이를 반복한다.

LASSO를 이용한 부모형의 생성

- LASSO:

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

- λ를 조정하면서 부모형을 구성할 수 있다.
  - $\lambda = 9 \Rightarrow \hat{\beta}(\lambda) = (\hat{\beta}_0, 0, 0, 0, \cdots, 0)$
  - $\lambda = 7 \Rightarrow \hat{\beta}(\lambda) = (\hat{\beta}_0, 0, 0, \hat{\beta}_3(\lambda), \cdots, 0)$
  - $\lambda = 4 \Rightarrow \hat{\beta}(\lambda) = (\hat{\beta}_0, 0, \hat{\beta}_2(\lambda), \hat{\beta}_3(\lambda), \cdots, 0)$
  - $\lambda = 1 \Rightarrow \hat{\beta}(\lambda) = (\hat{\beta}_0, \hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda), \hat{\beta}_3(\lambda), \cdots, 0)$
  - $\cdots$
- λ 조정함으로써 부모형을 쉽게 생성할 수 있다.

정규화 방법론 연구

정규화 방법론은 다음 함수를 최소화 하는 추정량에 대한 연구다.

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

위 식은 아래와 같이 이해할 수 있다.

risk function based on data + penalty function on the model complexity

모형 선택

F 검정을 통한 모형선택

- 전진법
  - 변수를 하나씩 추가해가면서 부모형들을 만든다.
  - 변수를 추가할 때마다, 위험함수값 (eg. SSE)이 얼마나 줄어드는지 계산한다.
  - 변수를 추가하면 항상 위험함수값은 늘어날 수 없다.
  - F-검정을 통해 위험함수값이 유의하게 줄어드는지 확인한다.
  - 즉, F 통계량은 유의미한 위험함수의 감소 다시말해, 유의미한 변수가 모형으로 들어왔는지 그렇지 않은지를 위험함수를 통해 판별한다.
  - 유의미한 위험함수의 감소가 보이지 않을 때 까지 변수를 추가한다.

F 검정을 통한 모형선택

- 후진법
  - 변수를 완전모형으로 부터 하나씩 빼가면서 부모형을 만든다.
  - 변수를 제거할 때마다, 위험함수값 (eg. SSE)이 얼마나 늘어나는지 계산한다.
  - 변수를 제거하면 항상 위험함수값은 줄어들 수 없다.
  - F-검정을 통해 위험함수값의 감소량에 유의한 증거가 없는지 확인한다.
  - 유의미한 위험함수의 감소가 보일 때 까지 변수를 추가한다.

모형선택 기준에 의한 모형선택

$\sigma^2 = 1$이라고 가정하자

- AIC (Akaike information criteria)

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + 2k$$

단 $k = \#$ of nonzero coefficients in $\hat{\beta}_j$ for $j = 1, \cdots, p$.

모형선택 기준에 의한 모형선택

- BIC (Bayesian information criteria)

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2 + k\log(n)$$

단 $k = \#$ of nonzero coefficients in $\hat{\beta}_j$ for $j = 1, \cdots, p$.

데이터 기반 통한 모형선택

- 검증 데이셋(Validation set)의 이용
  - 훈련 집합과 독립인 validation set의 확보
  - 훈련 집합의 데이터를 이용하여 부모형들을 만들고 검증데이터를 이용해 모형을 평가(우도, 정확도 등)
  - 검증 데이터 셋에서 좋은 성능을 보이는 부모형을 선택/

데이터 기반 통한 모형선택

- 교차 검증방법 (Cross validation:CV)
  - 데이터 집합을 k개의 분할로 만듬. 그 중 하나를 검증 데이터 셋으로 나머지를 훈련 데이터 셋으로 선택.
  - 훈련 집합의 데이터를 이용하여 부모형들을 만들고 검증데이터를 이용해 모형을 평가(우도, 정확도 등)
  - 훈련 데이터 중 하나를 검증 데이터 셋으로 선택하고, 나머지를 다시 훈련 데이터셋으로 놓음.
  - 훈련 집합의 데이터를 이용하여 부모형들을 만들고 검증데이터를 이용해 모형을 평가(우도, 정확도 등)가 하고 같은 작업을 반복함.
  - 총 k개의 평가 결과를 이용하여 부모형을 선택함
- 일반화 교차검증(Generalized Cross Validation: GCV): 계산의 복잡성을 피하기 위해 개발됨. GCV는 식의 형태로 주어져 빠른 시간내에 계산이 가능함.