1장 Introduction

CONTENTS

1.1 서론

- 회귀분석(Regression Analysis)은 데이터로부터 독립변수(또는 예측변수)들의 함수를 다음과 같이

$$\hat{y} = f(x_1, x_2, \ldots, x_p)$$

으로 추정하여 종속변수(또는 반응변수)를 예측하는 방법이다. 여기서 $x = (x_1, x_2, \ldots, x_p)$ 은 주어지는 값이며 y는 주어진 x에서 측정되는 값으로 벡터로 주어질 수도 있다.

- 이 책에서는 강력한 통계분석 도구인 R을 이용한 다양한 회귀분석 문제를 다룬다. 이 장에서는 이 책의 전반에 걸쳐 사용될 R의 모형식 표현과 적합된 회귀모형에 대한 평가 방법을 소개한다.

1.2 R의 모형식 표현

- R의 대표적인 회귀모형 적합함수인 lm() 함수에 사용되는 모형식 표현을 소개한다. R에서 모형식의 기본 형식은

$$y\,(\text{반응변수}) \sim x_1 + x_2 + \cdots\,(\text{예측변수})$$

들 이며, 함수 내에서 다음과 같이 사용된다.

```
> lm(y ~ x + z)    # 중회귀 모형식
```

- R에서 모형식 표현에 사용되는 기호를 정리하면 아래의 [표 1.1]과 같다. 모형식의 간결한 표현을 위해 이들 기호는 매우 유용하다.

1.2 R의 모형식 표현

[표 1.1] R의 모형식 기호

기호	예	의미
1	1+(또는 +1)	절편을 포함
+	+x	변수를 포함
-	-x	변수를 제거
:	x:z	변수 간의 교호효과를 포함
*	x*y	각 변수와 변수 간의 교호효과를 포함
|	x|z	조건부: z 조건하에 x 포함
^	(x+z+w)^3	각 변수와 3차까지의 모든 교호효과를 포함
-	x-1	항을 제거

1.2 R의 모형식 표현

- 동일한 모형식도 다양하게 표현될 수 있다. 예를 들어, 다음의 모형식들은 동치이다.
  - y ~ x+z+w+x:z+x:w+z:w+x:z:w
  - y ~ x*z*w
  - y ~ (x+z+w)^3
- 마찬가지로, 아래의 모형식들도 서로 동치이다.
  - y ~ x+z+w+x:z+x:w+z:w
  - y ~ x*z*w-x:z:w
  - y ~ (x+z+w)^2

1.2 R의 모형식 표현

- 데이터프레임을 사용할 때 모형식에서 "."을 사용하면 매우 편리하다. 예를 들어, 데이터프레임 D가 4개의 변수(또는 컬럼) y, x, z, w를 가질 때 다음의 표현

```
> fit <- lm(y ~ ., data=D)
```

에서 "."은 y를 제외한 모든 변수를 예측변수로 포함한다는 의미이다. 즉, 위의 표현은 다음과 동치이다.

```
> fit <- lm(y ~ x+z+w, data=D)
```

1.2 R의 모형식 표현

- 유사하게, 아래의 두 표현은 동치이며

```
> fit <- lm(y ~ .-w, data=D)
> fit <- lm(y ~ x+z)
```

다음의 두 표현도 서로 동치이다.

```
> fit <- lm(y ~ .*w, data=D)
> fit <- lm(y ~ x+z+w+x:w+z:w)
```

1.3 모형평가

1.3.1 최적의 부분 회귀모형의 선택 기준

- 최적의 부분모형(best subset regression) 선택은 예측변수들의 모든 가능한 부분집합을 예측변수로 하는 회귀모형(all possible subset regression)을 적합하고, 이 가운데 아래의 기준에 가장 잘 부합하는 모형을 찾는 방법이다.

- **결정계수**: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

- **수정 결정계수**: $R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SST}$

1.3 모형평가

- **평균제곱오차**: $MSE = \frac{SSE}{n-p} = \frac{\sum(y_i - \hat{y}_i)^2}{n-p}$
- **Mallow's C_p**: $C_p = \frac{SSE_p}{S^2} - N + 2(P+1)$

- 위 식에서 p, SSE, $\widehat{\sigma^2}$은 다음과 같이 정의된다.
  - p = 모수의 수(절편항 포함).
  - SSE = p개의 예측변수로 적합한 모형의 오차제곱합.
  - $\hat{\sigma}^2$ = 모든 예측변수를 포함한 적합모형의 평균제곱오차.

1.3 모형평가

- 위의 기준에 따른 변수 선택은 다음의 절차를 따른다.

(i) 결정계수는 p에 대해 증가함수이다. ➡ 따라서 (모형의 단순성을 위해) 증가가 둔화되는 시점의 p를 선택한다.

(ii) 수정결정계수는 결정계수의 단점을 보완하여 설명력이 약한 예측변수가 추가될 때는 오히려 감소한다. ➡ 따라서 가장 큰 값을 가지는 p를 선택한다.

(iii) 평균제곱오차가 최소가 되는 p를 선택한다. 수정결정계수의 결과와 동일하다.

(iv) C_p의 값이 p와 가장 가까운 값을 가지는 p를 선택한다.

1.3 모형평가

1.3.2 정보기준과 PRESS

(a) 정보기준

- 회귀모형의 비교를 위해 다음의 여러 가지 정보기준(information criterion) 통계량이 사용된다.
- 몇몇 데이터 분석가들은 이들 정보기준이 C_p 통계량 보다 더 현실적인 방법으로 생각한다.

  그 이유는 C_p가 실제보다 모형 간에 더 큰 차이가 있는 것처럼 보이게 하는 경향 때문이다.

  - **Akaike's 정보기준**:

$$AIC = n\ln\left(\frac{SSE}{n}\right) + 2p$$

  - **Bayesian 정보기준**(or Schwartz's Bayesian Criterion):

$$BIC = n\ln\left(\frac{SSE}{n}\right) + p\ln(n)$$

1.3 모형평가

(b) 예측제곱합(PRESS: prediction sum of squres)

- PRESS는 모형의 예측력을 통해 평가하는 방법이다.
- 아래의 정의에서 $\hat{y}_{i(i)}$은 i-번째 자료를 제외하고 적합한 모형으로부터 i-번째 값을 추정한 것이다. PRESS의 값이 작을수록 예측력이 우수하다고 할 수 있다.
  - $PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i(i)})^2$

1.3 모형평가

(c) 예측 결정계수(predicted R^2)

- 예측 결정계수는 PRESS를 통해 다음과 같이 정의된다. 이 값은 PRESS 보다 더 직관적으로 해석될 수 있다.
- PRESS와 더불어 데이터셋을 훈련용과 검증용으로 별도로 나누지 않고 모형의 예측력을 비교할 수 있어 유용하다.
  - $R^2_{pred} = 1 - \frac{PRESS}{SST}$ (이 값이 영보다 작을 때는 영으로 간주함)
- PRESS와 예측 결정계수는 모형 추정에 포함되지 않은 자료를 이용하여 계산되므로 과적합(overfitting)을 방지하는데 도움이 된다.
- 과적합은 모형 적합에 사용된 데이터에 대해서는 우수한 적합을 제공하나, 새로운 관측치에 대해서는 유용한 적합을 보이지 못하는 것을 의미한다.

1.3 모형평가

1.3.3 교차타당법

- 교차 검증법(cross-validation method)은 데이터셋을 모형구축에 사용될 훈련용셋(training set)과 예측력 평가에 사용될 평가용셋(validation or prediction set)으로 나누어 모형을 평가하는 방법이다.
- 데이터 양이 충분히 많은 경우에는 두 데이터셋의 비율을 50%:50%로 랜덤하게 나누어 적용한다.

1.3 모형평가

- **K-중첩(fold) 교차타당법**:

  데이터의 양이 충분치 않은 경우에는 전체 데이터셋을 (동일한 크기의) K 조각으로 나누고, 이 가운데 한 조각을 제외한 나머지 $(K-1)$ 조각으로 모형을 구축한 뒤, 남겨둔 한 조각에 대해 예측을 수행한다. 남겨지는 조각을 바꾸어 가며 이 절차를 K번 반복한다. 각 조각에 대한 제곱예측오차를 더하여 교차타당법의 측도로 사용한다.

- **Leave-one-out 교차타당법**:

  K-중첩 교차타당법에서 $K = n$인 경우에 해당한다. 즉, 한 개의 데이터만 남기고 모형을 구축한 후 남겨진 한 개를 추정하는 과정을 반복한다. 계산량이 다소 많아질 수 있다. Leave-one-out 교차타당법 을 이용한 예측오차의 추정값은 PRESS와 동일하다.