다중선형회귀분석2

Park Beomjin[1]
[1] *University of Seoul*

1 예제1 : 다항 회귀 모형

- "quad_dat.csv" 에는 반응변수에 해당하는 y 변수와 설명변수에 해당하는 x변수가 들어있다.
- 다음의 모형을 적합해보고 적합이 잘 되었는지 판단하여라.

$$Y = \beta_0 + \beta_1 X + \epsilon, \epsilon \sim N(0, \sigma^2) \tag{1}$$

- y와 x의 산점도를 확인하고 두 변수간의 어떠한 관계가 있는지 생각해보자.
- 설명변수에 X^2을 추가한 모형을 적합하여 보고 그 결과를 확인하여 보자.

```
PROC GLM data = reg.quad_dat;
model y = x x * x;
RUN; QUIT;
```

2 예제2: UFFI 데이터

- UFFI 데이터는 UFFI가 CH_2O 농도에 영향을 미치는지 알아보기 위해 수집된 데이터이다.
- 반응변수 Y : CH_2O, 설명변수 X_1 : UFFI, X_2 : Air Tightness
- 다음 세가지의 회귀모형을 고려해보자.
- Case1.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{2}$$

- Case2.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \tag{3}$$

- Case3.

$$Y = \beta_0 + \beta 2 x_2 + \epsilon \tag{4}$$

- 세가지 모형을 각각 적합하여 보고 추정된 모형을 UFFI = 1일 때와 0일때로 나누어 해석해보아라.

3 예제3: tree 데이터

- 산림지역에서 나무를 벌목할 때 해당 지역의 목재량을 조사할 필요가 있다.
- 그러나 나무의 부피를 직접 측정하는 것은 어렵기 때문에 상대적으로 측정이 쉬운 나무의 지름과 높이를 이용하여 부피를 추정하는 방법을 생각할 수 있다.
- tree 데이터에는 어느 지역에서 15그루의 나무를 표본으로 추출하여 부피(m^3), 지름(cm), 높이(m)를 측정한 값들이다.
- Q1 : 부피, 지름, 높이에 대한 산점도행렬을 작성하고, 변수들 간의 관계를 설명하라.
- Q2 : 부피를 반응변수로, 지름과 높이를 설명변수로 하는 선형회귀모형을 적합하여라.
- Q3 : 분산팽창인자(VIF)를 통해 설명변수간의 다중공선성을 확인하여라.

4 SAS CODE

- 예제1 SAS CODE

```
PROC REG data = reg.quad_dat;
model y = x;
RUN; QUIT;

PROC GLM data = reg.quad_dat;
model y = x x * x;
RUN; QUIT;
```

- 예제2 SAS CODE

```
PROC REG data = reg.uffi;
M1 : model y = x1 x2;
M2 : model y = x2;
RUN; QUIT;
```

```
PROC GLM data = reg.uffi;
model y = x1 x2 x1 * x2;
RUN; QUIT;
```

- 예제3 SAS CODE

```
PROC SGSCATTER data = reg.tree;
matrix volume -- height;
RUN; QUIT;

PROC REG data = reg.tree;
model volume = diameter height / VIF;
RUN; QUIT;

PROC GLM data = reg.tree;
model volume = diameter height diameter * height;
RUN; QUIT;

PROC GLM data = reg.tree plots = (diagnostics residuals);
model volume = diameter | height / tolerance;
RUN; QUIT;
```