다중선형회귀분석1

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

1 예제1 : p번의 단순회귀분석 VS 다중회귀분석

- expdata1에는 $y, x1, x2$ 변수가 순서대로 저장되어 있다.
- 변수 $y, x1, x2$간의 산점도 행렬을 그리고 산관관계 분석을 통해 변수간의 관계를 확인하여보자.
- 다음과 같은 세가지 회귀모형을 적합해보고 결과를 해석하여라.

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \epsilon \\ y &= \beta_0 + \beta_2 x_2 + \epsilon \\ y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \end{aligned} \tag{1}$$

- 세 모형을 적합한 후 x_1과 x_2의 계수를 확인하여 보자. 어떠한 차이점이 있는가?

2 예제2 : 중고차 가격 데이터

- 중고차 가격에는 어떠한 변수들이 영향을 줄지 생각해보자.
- 국내의 중고차 사이트에는 기본적으로 연식, 차종, 색, 사고 유무, 엔진 종류, 배기량, 연료 종류, 주행거리 등의 정보가 올라와 있다.
- 중고차 가격 데이터(usedcars.csv)는 국내 유명 중고차 사이트에서 2007년에 수집한 데이터로 가솔린 엔진의 중형차를 대상으로 각 중고차에 대한 가격(price), 연식(year), 주행거리(mileage), 배기량(cc), 변속기종류(automatic)의 변수가 존재한다.
- 연식의 단위는 개월, 주행거리는 km, 배기량은 cc 단위로 측정되었고 변속기는 수동(=0)과 자동(=1) 여부가 조사되어 있다.
- 데이터 읽기

```
libname reg "library path";
```

2

```
PROC IMPORT datafile = "data folder path\usedcars.csv" dbms = csv
replace out = reg.usedcars;
getnames = yes;
RUN;
```

- 중고차 가격 데이터의 모든 변수들간의 산점도를 확인하고 변수들간의 어떠한 관계가 있을지 예상해보자.

```
PROC SGSCATTER data = reg.usedcars;
matrix price -- automatic;
RUN;
/* 산점도 행렬의 대각에 해당 변수의 히스토그램과 추정한 density 그래프를 그려줌
PROC SGSCATTER data = reg.usedcars;
matrix price -- automatic / diagonal = (histogram kernel);
RUN;
*/
```

- 가격(price)와 다른 나머지 변수들간의 상관관계분석을 진행하고 산점도에서 예상했던 관계가 상관관계분석에 잘 반영되어 있는지 확인하여라.

```
PROC CORR data = reg.usedcars;
var price -- automatic;
RUN;
```

- 중고차 가격에 어떠한 변수들이 관계가 있는지 알아보고 예측하는 모형을 구성하여라.
- Model : $\text{Price} = \beta_0 + \beta_1 \text{year} + \beta_2 \text{mileage} + \beta_3 \text{cc} + \beta_4 \text{automatic} + \epsilon, \epsilon \sim N(0, \sigma^2)$

```
PROC REG data = reg.usedcars;
model price = year -- automatic;
RUN;
```

- 추정된 회귀모형의 결과를 확인하여보자.
  - 가격과 나머지 변수들의 관계와 유의성은 어떠한가?
  - 적합된 모형의 R^2는 어떠한가?
  - 적합된 모형의 분산분석 결과는 어떠한가?
  - Output으로 함께 제공된 Fit Diagnostics 그래프에서 residual의 산점도를 확인하여 보고 적합된 회귀모형이 가정에 만족하는지 확인하여라.

- 가격에 대해 다른 나머지 변수들을 각각 단순선형회귀(Simple linear regression)으로 적합하여 보고 다중선형회귀분석과 단순선형회귀분석을 p번 했을 때의 결과를 비교하여 보아라.

3 예제3 : 주택 판매가격 데이터

- 주택 판매가격 데이터(houseprice.csv)는 주택 판매가격(price; 천만원)와 이에 영향을 줄 것으로 판단되는 4가지 설명변수인 세금(tax; 만원), 대지평수(ground;평), 건물평수(floor; 평), 주택연령(year;년)을 27가구에 대해 조사한 데이터이다.
- Q1 : 5개의 변수들에 대한 산점도행렬을 작성하고 변수들 간의 관계를 예상하여라.
- Q2 : 주택 판매가격과 나머지 변수들간의 상관관계분석을 진행하고 결과를 확인하여라.
- Q3 : 주택 판매가격을 반응변수로, 나머지 4개의 변수를 설명변수로 하는 다중선형회귀모형을 적합하여라.
- Q4 : 추정된 회귀모형에서 주택 판매가격에 영향을 주는 변수들이 무엇인지 판단하여보고 분산분석 결과를 확인하여라.
- Q5 : 결정계수(R^2)와 MSE 그리고 잔차그래프를 확인하고 모형의 적합도와 모형의 가정을 만족하는지 확인하여라.

```
PROC IMPORT datafile = "data folder path\houseprice.csv"
dbms = csv replace out = reg.houseprice;
getnames = yes;
RUN;

/* SAS CODE for Q1 */
PROC SGSCATTER data = reg.houseprice;
matrix price -- year/ diagonal = (histogram kernel);
RUN;


/* SAS CODE for Q2 */
PROC CORR data = reg.houseprice;
var price -- year;
RUN;

/* SAS CODE for Q3*/
PROC REG data = reg.houseprice;
model price = tax -- year;
RUN;
```