모형 점검

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

1 예제1 : Infludata

- Infludata에는 총 11개의 관측치가 포함되어 있으며 반응변수 Y와 설명변수 X가 존재한다.
- 반응변수 Y와 설명변수 X 간의 산점도를 그려보자.

```
PROC SGPLOT data = reg.infludata;
scatter X = x Y = y;
RUN; QUIT;
```

- 이 데이터에서 모형 $Y = \beta_0 + \beta_1 X + \epsilon$를 적합한다고 했을 때, 11개의 데이터를 모두 사용하여 추정한 $\hat{\beta}_0$와 $\hat{\beta}_1$, i번째 관측치를 제외하고 추정한 계수 $\hat{\beta}_{(-i)0}$와 $\hat{\beta}_{(-i)1}$, $i = 1, 2, \{10, 11\}$ 간에 어떠한 차이점이 있는지 생각해보자.
- 실제로 전체 데이터를 사용했을 때 회귀모형을 적합해보고 $1, 2, \{10, 11\}$번째 관측치를 제외하고 회귀모형을 적합하여 결과를 비교해보자.

```
DATA reg.infludata2;
set reg.infludata;
index = _n_;
RUN; QUIT;

PROC REG data = reg.infludata2 outest  = est1;
model y = x / OUTSEB noprint;
where index ^= 1;
PROC PRINT data = est1;
RUN; QUIT;

PROC REG data = reg.infludata2 outest  = est2;
model y = x / OUTSEB noprint;
where index ^= 2;
PROC PRINT data = est2;
```

```
RUN; QUIT;

PROC REG data = reg.infludata2 outest  = est3;
model y = x / OUTSEB noprint;
where index not in  (10, 11);
PROC PRINT data = est3;
RUN; QUIT;
```

2 예제2 : 중고차 가격 데이터

- 중고차 가격 데이터("usedcar.sas7bdat)"에서 모든 설명변수를 포함시켜 적합한 모형에 대하여 여러가지 영향력 측도들을 구하여 분석해보자.
- 스튜던트 잔차와 표준화 제외 잔차(Rstudentized residuals)을 기준으로 특이점을 식별해보자.
- 지렛값(leverage)을 통해 높은 지렛점을 식별해보자.
- 관측치 별 쿡의 거리(Cook's distance)와 DFBETAS를 보고 영향력 관측치를 식별해보자.
- COVRATIO를 통해 영향력 관측치를 식별해보자.
- DFFITS를 확인고 영향력 관측치를 식별해보자.

```
PROC REG data = reg.usedcar;
model price = year -- automatic / r influence;
RUN; QUIT;
```