다항 회귀 및 변수 선택

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

1 예제1 : Cubic data

- "Cubic.csv" 데이터에는 반응변수 y와 설명변수 x를 포함하고 있다.
- x와 y간의 산점도를 확인하고 회귀모형 $Y = \beta_0 + \beta_1 x + \epsilon$이 적절하게 적합될지 논해보자.

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

- 실제로 회귀모형 $Y = \beta_0 + \beta_1 x + \epsilon$를 적합해보고 모형이 적절한지 논해보자.

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

- 해당 데이터에 적절한 모형을 적합하기 위한 변수를 생성해보자.

```
DATA reg.cubic2;
set reg.cubic;
x2 = x**2;
x3 = x**3;
RUN; QUIT;
```

- 생성한 변수를 이용하여 적절한 회귀모형을 적합하고 type I 제곱합과 type II 제곱합을 통해 각 변수의 유의성을 확인해보자.

```
PROC REG data = reg.cubic2;
M1 : model y = x;
M2 : model y = x x2;
M3 : model y = x x2 x3;
ods output ParameterEstimates=pval1;
PROC PRINT data = pval1;
```

```
RUN; QUIT;

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

2 예제2: 중고차 가격 데이터

- 중고차 가격 데이터에서 설명변수 year, mileage, cc, 그리고 automatic의 계수를 검정하고자 한다.
- type I과 type II 제곱합을 이용하여 계수에 대한 검정을 진행해보자.

```
PROC GLM data = reg.usedcar;
model price = year -- automatic;
RUN; QUIT;
```

- 중고차 가격 데이터에서는 type I과 type II 제곱합 중 어떤 검정 방법이 더 적절해보이는지 논해보자.
- 만약 변수의 중요도가 year, cc, mileage 그리고 automatic 순으로 중요하다고 했을 때, type I 제곱합을 통해 다시 검정해보고 이전 결과와 비교해보자.

```
PROC GLM data = reg.usedcar;
model price = year cc mileage automatic;
RUN; QUIT;
```

3 예제3: wine data

- "wine_tmp.csv"는 wine 데이터 중 일부만 추출한 데이터이다.
- 레드 와인에 대해서 화학적 측정 방법을 이용하여 와인 품질을 예측하는 모형을 모형을 만들고자 한다.
- 선형회귀모형을 적합해보고 잔차 그래프와 적절한 검정을 통해 모형이 가정에 만족하는지 확인해보자. 또한 영향력 관측치를 식별하기 위해 관측치별 영향력 측도를 산출해보자.

```
PROC REG data = reg.wine_tmp (where = (type = "red"));
model quality = fixed_acidity -- alcohol / spec dw;
output out = reg.wine_out COOKD = cookd DFFITS = diffit COVRATIO = covratio
RSTUDENT = rstudent H = hatval;
RUN; QUIT;
```

- Cook's Distance와 rstudent 잔차를 통해 영향력 관측치를 식별하고 식별된 영향력 관측치를 제거하고모형을 적합해보자. (Cook's Distance는 4/n보다 크면, rstudent 잔차는 절대값이 2.5보다 크면 영향력 관측치라고 식별한다.)

```
DATA reg.wine_inf;
set reg.wine_out;
obs_index = cookd < (4 / 483) & abs(rstudent) < 2.5;
RUN; QUIT;

/* Method1 */
PROC REG data = reg.wine_inf (where = (obs_index = 1 & type = "red"));
model quality = fixed_acidity -- alcohol / r;
RUN; QUIT;

/* Method2 */
DATA reg.tar_dat;
set reg.wine_inf (where = (obs_index = 1 & type = "red"));
RUN; QUIT;

PROC REG data = reg.tar_dat;
model quality = fixed_acidity -- alcohol  / r;
RUN; QUIT;
```

- 모든 가능한 회귀를 통해 변수 선택을 진행해보자.

```
PROC REG data = reg.tar_dat;
model quality = fixed_acidity -- alcohol / selection = rsquare mse cp aic bic sb
RUN; QUIT;
```

- 변수선택 방법 중 stepwise, forward 그리고 backward 선택법을 모두 적용해보고 최종 선택된 모형을 비교해보자.

```
PROC REG data = reg.tar_dat outest = reg.step_out;
model quality = fixed_acidity -- alcohol / selection = stepwise;
PROC REG data = reg.tar_dat outest = reg.forward;
model quality = fixed_acidity -- alcohol / selection = forward;
PROC REG data = reg.tar_dat outest = reg.backward;
model quality = fixed_acidity -- alcohol / selection = backward;
RUN; QUIT;

DATA reg.output;
set reg.step_out reg.forward reg.backward;
PROC PRINT data = reg.output;
RUN; QUIT;
```

- test 데이터를 이용하여 stepwise를 통해 변수를 선택한 모형과 모든 설명변수를 다 포함한 모형과의 평균제곱오차(MSE)를 비교하여보자.

```
PROC SCORE data = reg.test_wine score = reg.step_out out = reg.step_pred type = p
var fixed_acidity -- alcohol;
RUN; QUIT;

PROC REG data = reg.tar_dat outest = reg.full;
model quality = fixed_acidity -- alcohol;
RUN; QUIT;

PROC SCORE data = reg.test_wine score = reg.full out = reg.full_pred type = parms
var fixed_acidity -- alcohol;
RUN; QUIT;

DATA reg.full_pred;
set reg.full_pred;
class = "full";
RUN;

DATA reg.step_pred;
set reg.step_pred;
class = "step";
```

```
RUN;

DATA reg.total_pred;
set reg.full_pred reg.step_pred;
squre = (quality - model1)**2;
RUN; QUIT;

PROC means data = reg.total_pred mean;
class class;
var squre;
RUN;
```

4 예제4: KBO data

- "kbo.csv"는 2011년 1군 경기에서 1경기 이상 뛴 타자 162명의 선수들을 대상으로 연봉, 안타수, 홈런수, 타점수, 득점수, 볼넷수, 통ㅅ산안타수, 통산경력, 구단명, 선수명이 포함된 데이터이다.
- 연봉과 나머지 변수간의 산점도를 확인해보고 연봉 변수의 적절한 변환을 생각해보자.

```
PROC SGSCATTER data = reg.kbo;
matrix Y -- X7 / diagonal = (histogram kernel);
RUN; QUIT;
```

- 변환된 연봉 변수와 나머지 변수간의 산점도를 확인해보자.

```
DATA reg.kbo2;
set reg.kbo;
logY = log(Y);
RUN; QUIT;

PROC SGSCATTER data = reg.kbo2;
matrix logY X1 -- X7 / diagonal = (histogram kernel);
RUN; QUIT;
```

- 7개의 설명변수를 모두 포함한 모형을 적합하고 어느 변수들이 연봉과 유의한 관계를 가지는지 확인해보자.

- 멜로우즈 C_p를 변수선택 기준으로 사용하여 모형을 구축해보자.
- 단계별 회귀에 의한 모형을 구축해보자.

```
PROC REG data = reg.kbo2 outest = reg.kbo_out;
M1 : model logY = X1 -- X7;
M2 : model logY = X1 -- X7 / selection = rsquare adjrsq cp aic bic best = 1;
M3 : model logY = X1 -- X7 / selection = stepwise;
RUN; QUIT;
```