다중선형회귀분석3

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

1 예제1: 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.

```
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;
```

2 예제2 : Wine 데이터

- Wine 데이터는 포르투갈의 북부에 있는 reg wine과 white wine 샘플 데이터이다.
- 데이터는 와인의 품질(quality) 변수와 11개의 화학 측정 결과를 나타내는 변수를 가지고 있다.
- 목표는 화학 측정 결과를 기반으로 와인 품질을 예측하는 모형을 모델링 하는 것이다.
- Q1 : red wine 데이터와 white wine 데이터를 결합하고 type이라는 새로운 변수를 생성하여 reg wine의 데이터인 경우 "red", white wine 데이터인 경우 "white" 값을 갖도록 하여라.
- Q2 : quality 변수를 반응변수로 하고 화학 측정 결과를 나타내는 11개 변수를 설명변수로 하는 회귀모형을 적합하여라.
- Q3 : red와 white 와인 별 각각 회귀모형을 적합하여라.
- Q4 : 각 회귀모형의 결과를 해석하고 vif를 계산하는 옵션을 추가한 후 vif를 확인하여라.
- Q5 : vif가 10 이상인 변수를 제외한 후 회귀모형을 다시 적합하고 결과를 해석하여라.

```
PROC IMPORT datafile = "C:\Users\User\Desktop\19 regression class\Data\Week9 Data
out = reg.whitewine;
getnames = yes;
delimiter = ";";
RUN;

DATA reg.wine;
set reg.redwine reg.whitewine(in = x);
if x = 0 then type = "red";
else type = "white";
RUN;

PROC REG data = reg.wine;
model quality = fixed_acidity -- alcohol / vif;
RUN;
```

```
PROC REG data = reg.wine (where = (type = "red"));
model quality = fixed_acidity -- alcohol / vif;
RUN; QUIT;

PROC REG data = reg.wine (where = (type = "whi"));
model quality = fixed_acidity -- alcohol / vif;
RUN; QUIT;

PROC REG data = reg.wine (where = (type = "whi"));
model quality = fixed_acidity -- total_sulfur_dioxide pH -- alcohol / vif;
RUN; QUIT;
```

3 예제3: 탑승객 데이터

- 탑승객 데이터(airline_passengers.csv)는 1949년 1월부터 1960년 12월까지의 어느 공항의 탑승객 수에 대한 데이터이다.
- Q1 : 연도를 X(설명변수) 탑승객 수를 Y(반응변수)로 하는 회귀모형을 적합 및 분산의 동질성(homogeneity of variance)에 대해 검정하고 적합된 모형에 대해 논의하여라.
- Q2 : 데이터를 이용하여 이분산성을 해결할 수 있는 가중치를 산출하는 방법을 생각해보고 가중치를 산출하여라.
- Q3 : 계산한 가중치를 이용하여 가중 최소제곱법(weighted least square)을 통해 회귀모형을 적합 및 분산의 동질성 검정을 하여라.

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

DATA reg.passengers2;
set reg.passengers;
ind = _n_;
RUN;

PROC SGPLOT data = reg.passengers2;
```

```
scatter X = month Y = passengers;
RUN; QUIT;

/* SAS CODE for Q1 */
PROC REG data = reg.passengers2;
model passengers = ind / spec r;
output out = reg.passenger_out p = pred r = resid;
RUN; QUIT;

/* SAS CODE for Q2 */
DATA reg.abs_resid;
set reg.passenger_out;
abs_resid = abs(resid);
RUN; QUIT;

PROC REG data = reg.abs_resid NOPRINT;
model abs_resid = ind;
output out = reg.passenger_abs_resid p = pred;
RUN; QUIT;

/* SAS CODE for Q3*/
DATA reg.weight_dat;
set reg.passenger_abs_resid;
w = 1 / (pred**2);
PROC REG data = reg.weight_dat;
model passengers = ind / spec r;
weight  w;
output out = reg.fpassenger p = wp student = wr r = wr0;
```