단순선형회귀분석3

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

1 예제1 : 미국 자동차 판매 데이터

- 미국 자동차 판매 데이터는 1993년도 미국 자동차의 브랜드, RPM, 가격 등이 기록되어 있는 데이터이다.

- 자동차 가격에 자동차 무게가 영향을 미치는지 알아보고 자동차 무게를 통해 자동차 가격을 예측하는 회귀모형을 적합하여라.

- Training과 Test데이터 읽기

```
PROC IMPORT datafile = "Data File Path\train_car.csv" dbms = csv replace
out = reg.train_car;
getnames = yes;
RUN;
PROC IMPORT datafile = "Data File Path\test_car.csv" dbms = csv replace
out = reg.test_car;
getnames = yes;
RUN;
PROC IMPORT datafile = "Data File Path\test_y.csv" dbms = csv replace
out = reg.test_y;
getnames = yes;
RUN;
```

- Training 데이터에서 자동차 가격(Price)와 자동차 무게(Weight)의 산점도와 표본상관계수를 산출하고 모상관계수가 0인지 검정하여라(실습).

- 자동차 가격과 무게와의 관련성을 알아보고 자동차 무게를 통해 가격을 예측하는 모형 만들기.

- 자동차의 가격(Y)와 무게(X)로 이루어진 회귀모형 적합 (평균함수 신뢰대와 예측값 신뢰대 포함)

```
PROC reg data = train_car outest = reg_out;
car_hat : model price = weight / R CLM CLI;
RUN; QUIT;
```

- R^2와 분산분석 결과는 어떠한가? 우리가 구성한 회귀 모형이 잘 적합했다고 볼 수 있는가?
- 자동차 무게에 따른 평균 가격 예측하기

```
PROC score data = reg.test_car score = reg_out out = new_pred type = parms;
var weight;
RUN;

DATA residuals (keep = price car_hat);
merge reg.test_y new_pred;
RUN;
```

- 자동차 무게에 따른 평균 가격 예측하기 (평균 모수의 신뢰대와 예측값의 신뢰대 포함하여 예측하기)

```
DATA reg.car;
set reg.train_car reg.test_car (in = new);
w = not(new);
RUN;

PROC REG data = reg.car;
model Price = Weight / CLM CLI;
output out = reg.car_pred(where = (w = 0)) p = pred lcl = LCL_pred ucl = UCL_pred
RUN; QUIT;

DATA residuals2 (keep = price pred LCL_pred UCL_pred LCLM UCLM);
merge reg.test_y reg.car_pred(drop = price);
RUN;
```

- 설명변수(X)를 EngineSize로 바꾼 후 자동차 가격(Price)와 산점도 및 표본 상관계수를 산출하여라 (실습).
- EngineSize(X)와 자동차 가격(Price)간의 단순선형회귀분석을 진행하여라.

```
PROC REG data = reg.car outest = reg_out2;
car_hat2 : model Price = EngineSize / CLM CLI;
output out = reg.car_pred2(where = (w = 0)) p = pred lcl = LCL_pred ucl = UCL_pre
RUN; QUIT;

DATA reg.eval_pred2;
merge reg.car_pred2 reg.test_y;
RUN;
```

- 어느 설명변수가 자동차 가격을 더 잘 설명하는가?

2 예제2 : 연습문제 2.6 hooker data

- hooker 데이터는 히말라야 산맥의 다양한 지점에서 물의 끓는 온도(TEMP)와 기압(AP)을 측정한 데이터이다.
- TEMP와 AP간의 산점도를 그려보고 선형회귀모형이 적절한지 알아보자.
- 데이터 읽고 산점도 그리기

```
PROC IMPORT datafile = "C:\Users\User\Desktop\19학년도 회귀분석 및 실습\Data\Week4
out = reg.hooker;
getnames = yes;
delimiter = ",";
RUN;

PROC SGPLOT data = reg.hooker;
scatter X = AP Y = BT;
RUN;
```

- TEMP와 AP간의 산점도를 봤을 때, 선형회귀모형이 적절하게 적합될 것 같은가? X = 100log(AP)로 변환하고 TEMP와 X간의 산점도를 그려보자.

```
DATA reg.trans_hooker;
set reg.hooker;
log_AP = 100 * log(AP);
RUN;
```

```
PROC SGPLOT data = reg.trans_hooker;
scatter X = log_AP Y = BT;
RUN;
```

- TEMP를 각각 AP와 X가 설명변수인 회귀모형에 적합하고 두 모형을 비교하여라. 어느 모형이 더 적절한가?