단순선형회귀분석

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

1 예제1: 분포 유도 시뮬레이션

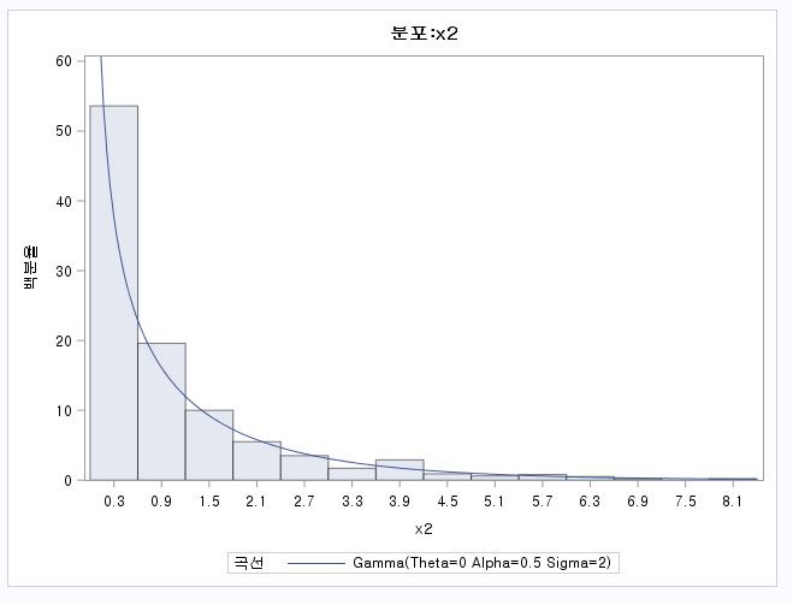
- 표준정규분포에서 카이제곱 분포 유도하기

  - 만약 $Z \sim N(0,1)$라 가정하면 $Z^2 \sim \chi^2(1)$ 이다.

  - 시뮬레이션을 통해 실제로 표준정규분포에서 생성된 random sample의 제곱이 카이제곱 분포를 따르는지 확인하여보자.

```
DATA normal(drop = seed n);
seed = 1;
call streaminit(seed);
do n = 1 to 1000;
        x1 = rand("Normal", 0, 1);
        x2 = x1 ** 2;
        output;
end;
PROC UNIVARIATE data = normal;
var x1 x2;
histogram x1 / normal(mu = 0, sigma = 1);
histogram x2 / gamma(alpha = 0.5 sigma = 2);
RUN;
```

2



2 예제2 : 경도 데이터

- 회귀분석을 통해 경도 데이터에서 온도가 경도에 영향을 미치는지 알아보자.
- 데이터 입력하기

```
libname reg "Library Path";
DATA reg.hardness;
input Temp Hardness @@;
cards;
30 55.8 30 59.1 30 54.8 30 54.6 40 43.1
40 42.2 40 45.2 50 31.6 50 30.9 50 30.8
60 17.5 60 20.5 60 17.2 60 16.9
;
RUN;
```

- 경도와 온도간의 관계를 눈으로 확인하기 위해 산점도를 그려본다.

```
PROC SGPLOT data = reg.hardness;
scatter X = Temp Y = Hardness;
RUN;
```

- 경도와 온도간의 상관관계가 있는지 알아보기위해 상관관계분석을 실시한다.

```
PROC CORR data = reg.hardness FISHER(rho0 = 0);
var Temp Hardness;
RUN;
```

- 경도와 온도간의 관계를 알아보고 특정 온도에 따른 경도의 예측값을 알아보기 위해 경도를 반응변수(Y) 온도를 독립변수 (X)로 설정한 회귀모형을 적합한다.

```
PROC REG data = reg.hardness;
model hardness = temp
RUN;
```

- 회귀분석시 "CLM" 옵션은 평균값의 신뢰대를 산출하여주고 "CLI" 옵션은 예측값의 신뢰대를 산출해준다.

```
PROC REG data = reg.hardness;
model hardness = temp / CLM CLI;
plot hardness * temp / pred;
RUN;
```

- 새로운 값 Temp = 55에 대해 추정된 회귀모형을 이용하여 예측하고 예측값에 대한 신뢰구간, 평균 값에 대한 신뢰구간 구하기

```
DATA reg.hardness_test;
set reg.hardness end = last;
output;
if last then do;
Hardness = .;
Temp = 55;
output;
end;
RUN;
PROC REG data = reg.hardness_test;
model hardness = temp / CLM CLI;
output out = test_out(where = (Hardness = .)) p = predicted ucl = UCL_Pred
lcl = LCL_Pred uclm = UCLM_mean lclm = LCLM_mean;
RUN; QUIT;
```

3 예제3 : 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).
- 예측값을 산출하기 위해 Training 데이터와 Test 데이터를 결합한다.

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

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

```
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;
```

- 실제 값과 예측값 비교하기

```
DATA reg.eval_pred;
merge reg.car_pred reg.test_y;
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;
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?