통계학 실습

5. 상관분석과 가설검정

상관 관계 분석

- 상관 관계 분석은 변수들 간의 선형 관계를 나타내어 준다.

- 선형 관계의 값은 −1과 +1 사이에서 존재한다.
  - 절대값이 1에 가까울수록 더 강한 선형적인 관계를 띈다.
  - −1에 가까우면 음의 상관관계
  - +1에 가까우면 양의 상관관계

상관분석

```
PROC CORR DATA=SAS데이터;
  VAR 변수명1 변수명2;
RUN;
```

- 상관분석은 프로시저를 이용한다.
- CORR : 상관분석 프로시저
- VAR : 관계를 분석할 변수들로써 반드시 2개 이상이 입력되어야 한다.

모상관계수에 대한 가설 검정 (1/3)

```
DATA cars_data;
  SET SASHELP.CARS;
RUN;
```

- 가설 검정이란, 변수 간의 관계에 대한 가설을 세워두고 그 가설이 옳은지 분석을 통해 검정해보는 행위이다.
- 먼저, SASHELP에 존재하는 CARS 데이터를 불러온다.
- ‘자동차의 무게와 마력(힘)이 상관이 있을 것이다.’ 라고 가설을 만든다.

모상관계수에 대한 가설 검정 (2/3)

```
PROC CORR DATA=cars_data;
  VAR horsepower weight;
RUN;
```

- 자동차의 무게와 마력에 대한 상관분석을 시도한다.
- 상관분석의 출력 결과를 통해 관계를 예측한다.

CORR 프로시저

2 개의 변수:	Horsepower Weight

단순 통계량

변수	N	평균	표준편차	합	최솟값	최댓값	레이블
Horsepower	428	215.88551	71.83603	92399	73.00000	500.00000	
Weight	428	3578	758.98321	1531364	1850	7190	Weight (LBS)

피어슨 상관 계수, N = 428
H0: Rho=0 가정하에서 Prob > |r|

	Horsepower	Weight
Horsepower	1.00000	0.63080 <.0001
Weight Weight (LBS)	0.63080 <.0001	1.00000

모상관계수에 대한 가설 검정 (3/3)

- 결과에서 자동차의 무게와 마력은 상관계수는 0.63080이다. 이 상관계수가 통계적으로 유의한 값인지 모상관계수에 대한 가설검정을 수행한다.

- 1. 가설설정
  - 자동차의 무게와 마력은 상관 없다 / 상관 계수는 0이 아니다.
- 2. 유의확률의 계산
  - 유의확률이란, 귀무가설이 맞는데 잘못해서 기각할 확률이다.
  - 코드의 결과로 유의 확률이 0.0001보다 작다는 것을 확인할 수 있다.
- 3. 유의수준 α 결정
  - 유의수준이란, 귀무가설을 기각하는 검정 결과가 잘못될 확률이다.
  - 유의수준을 0.05로 결정한다. (결과의 5%는 잘못될 가능성이 있다는 뜻이다.)
- 4. 가설의 기각 여부 결정
  - 귀무가설이란, 의미 있는 차이가 없다는 경우의 가설이다.
  - 유의확률이 유의수준보다 작으므로 귀무가설을 기각할 수 있다.
  - 두 변수 간에는 상관관계가 존재하며, 계수가 0.63080이므로 강한 양의 상관 관계를 띄고 있다.

산점도 프로시저

- 상관 관계를 보여주기 위하여 자주 쓰이는 것이 산점도이다.

- 산점도는 두 변수의 대응되는 자료들을 좌표평면에 점으로 표시하는 것으로 이를 통해 변수 간의 선형 관계를 대략적으로 파악할 수 있다.

산점도 프로시저

```
PROC PLOT DATA=cars_data;
  PLOT horsepower*weight;
RUN;
```

- 이전에 사용한 CARS 데이터를 이용하여 산점도를 그려본다.
- 산점도를 통해 상관계수를 구하기 전에 대략적인 상관 관계를 파악할 수 있다.
- 자동차의 무게와 마력은 대략 양의 선형관계를 띄고 있음을 알 수 있다.

상관 관계 매트릭스

```
PROC CORR DATA=cars_data PLOTS=matrix;
  VAR horsepower weight;
RUN;
```

- 산점도를 조금 더 상관분석 관점에서 보기 위하여 CORR 기능에서 상관 관계 매트릭스를 그릴 수 있다.
- 주어진 두 변수 사이의 행렬을 구하여 해당 행렬을 산점도로 그려낼 수 있다.

실습

- 1. SASHELP의 BASEBALL 데이터를 불러온다.
- 2. 타석에 선 횟수와 안타를 친 횟수 간의 상관 분석을 해본다.
  - nAtBat : 타석에 선 횟수
  - nHits : 안타를 친 횟수

CORR 프로시저

2 개의 변수:	nAtBat nHits

단순 통계량

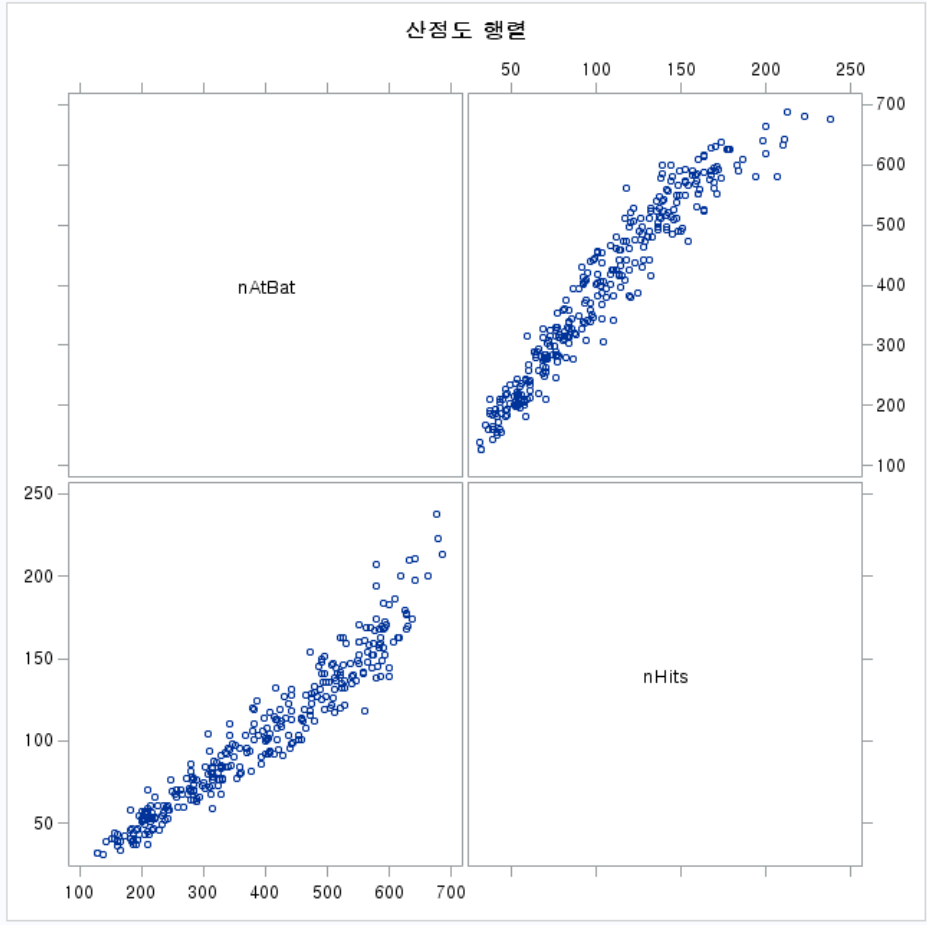
변수	N	평균	표준편차	합	최솟값	최댓값	레이블
nAtBat	322	390.07453	143.59584	125604	127.00000	687.00000	Times at Bat in 1986
nHits	322	103.39752	44.17951	33294	31.00000	238.00000	Hits in 1986

피어슨 상관 계수, N = 322
H0: Rho=0 가정하에서 Prob > |r|

	nAtBat	nHits
nAtBat Times at Bat in 1986	1.00000	0.96447 <.0001
nHits Hits in 1986	0.96447 <.0001	1.00000

실습 (2)

- 3. 상관관계 매트릭스를 통하여 두 변수 간의 관계를 파악한다.

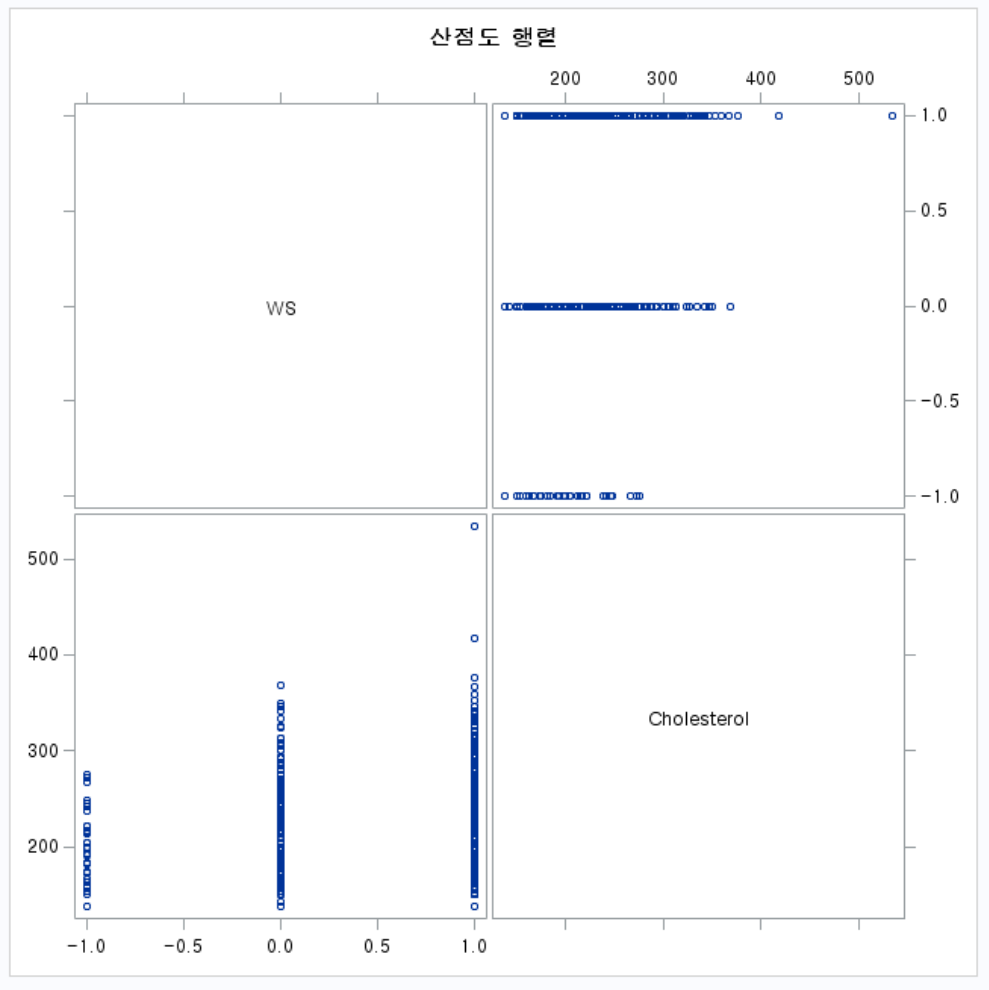


문제

- SASHELP 라이브러리에 있는 'HEART' 데이터를 이용한다. HEART 데이터에서 'Weight_Status'는 무게의 상태를, 'Cholesterol'은 콜레스테롤 수치를 나타낸다. 콜레스테롤 수치와 무게 상태의 연관성을 상관관계 매트릭스를 통해 확인하여 보자.
  - KEEP을 통해 필요 없는 변수들을 걸러낸다.
  - Weight_Status 변수는 문자열이기 때문에 CORR에서 쓰일 수 없다. 따라서 새로운 변수가 필요하다. 새로운 변수를 추가하여 Weight_Status의 상태에 따른 정수 값을 주도록 한다.

    Ex) Overweight –〉 1, Normal –〉 0, Underweight –〉 –1
  - 상관관계 매트릭스에는 한 번에 많은 데이터가 PLOT될 수 없다.
    1. SET 데이터에서 (OBS=1000)을 이용하여 데이터 ROW의 개수를 1000개로 줄이도록 한다.
    2. 새로운 변수와 Cholesterol을 PLOT 하도록 하자.

문제 결과 분석



- 상관관계 매트릭스를 통해 콜레스테롤 수치 상태가 무게 상태에 영향을 주지만, 크게 주지 않는다는 것을 알 수 있다.

- 유의수준이 0.05일 때, 유의확률이 유의수준보다 작으므로 귀무가설을 기각할 수 없다. 따라서 두 변수는 상관 관계가 존재하며, 상관계수가 0.147150이므로 매우 약한 양의 상관관계를 띄고 있음을 알 수 있다.

숙제 – 레포트

- 1. SASHELP의 BASEBALL 라이브러리를 이용한다.
- 2. 상관 관계를 확인하고 싶은 두 변수를 택한다.
- 3. 두 변수에 대하여 가설 검정을 세운다.
- 4. 상관 분석을 하고, 모상관계수에 대한 가설검정 절차를 통해 결론을 낸다.
- 5. 해당 절차를 모두 작성하여 레포트로 제출할 것.
  - 본 자료에 있는 내용을 이해하여 잘 따라오면 쉽게 할 수 있다.

숙제 – 레포트

- 레포트 파일이름
  - [5주차][학번][이름]통계학실습레포트
  - PDF로 제출하여야 하니 HWP나 Pages를 이용하세요.

- 레포트 내용
  - 제목 : 야구에서 변수 A와 변수 B의 상관 관계
  - 학번, 이름
  - 야구 라이브러리에서 선택한 변수 A와 변수 B에 대한 설명 및 가설 (짧게)
  - 상관 분석 코드 (글이나 이미지)
  - 상관 분석 결과 화면 (이미지)
  - 상관 분석 매트릭스 그림 (이미지)
  - 모상관계수에 대한 가설검정 절차 (4가지)

제출

- 1. 레포트를 PDF로 내보낸다.
- 2. PDF를 압축하여 아래 서식의 이름으로 만들어 이메일로 제출한다. 이메일 제목은 파일 이름과 동일하게 한다. (제출파일이름=압축파일이름=이메일제목)
- 서식
  - [5주차][학번][이름]통계학실습레포트
- 제출 이메일
  - gtsk623@gmail.com