Introduction to SAS 2

탐색적 자료분석 예제

Example **PennState STAT 480: Introduction to SAS에서 발췌한 것임**

- *subj* : subject id number
- *name*: patient's name
- *clinic*: where patient was treated
- *gender*: gender of subject (1: female, 2: male)
- *no_vis*: number of visits to a medical facility (0, 1, 2,...)
- *type_vis*: type of visit (101: gynecology, 190: physical therapy 187: cardiology)
- *expense*: medical charges in dollars.

```
DATA basic;
  input subj 1-4 name $ 6-23 clinic $ 25-28 gender 30 no_vis 32-33 type_vis
35-37 expense 39-45;
  DATALINES;
1024 Alice Smith        LEWN 1  7 101 1001.98
1167 Maryann White      LEWN 1  2 101 2999.34
1168 Thomas Jones       ALTO 2 10 190 3904.89
1201 Benedictine Arnold ALTO 2  1 190 1450.23
1302 Felicia Ho         MNMC 1  7 190 1209.94
1471 John Smith         MNMC 2  6 187 1763.09
1980 Jane Smiley        MNMC 1  5 190 3567.00
  ;
RUN;
/* Selecting observations */
PROC PRINT data = basic (FIRSTOBS = 2 OBS = 5);
   var subj name no_vis expense;
RUN;

PROC PRINT data = basic;
   var name no_vis type_vis expense;
   where no_vis > 5;
RUN;
PROC PRINT data = basic;
    var name gender no_vis type_vis expense;
       where name contains 'Smi';
RUN;

/* Sorting data*/
PROC SORT data = basic out = srtd_basic;
   by clinic no_vis;
RUN;

PROC PRINT data = srtd_basic NOOBS;
   var clinic no_vis subj name gender type_vis expense;
RUN;
```

```
PROC SORT data = basic out = srtd_basic;
   by descending clinic no_vis;
RUN;
PROC PRINT data = srtd_basic NOOBS;
   var clinic no_vis subj name gender type_vis expense;
RUN;

/* Column Totals*/
PROC PRINT data = basic;
   id name;
   var clinic no_vis;
   where type_vis = 190;
   sum no_vis;
RUN;

PROC SORT data = basic out = srtd_basic;
  by clinic;
RUN;

PROC PRINT data = srtd_basic;
   by clinic;
   var subj name no_vis type_vis expense;
   sum expense;
RUN;

/* Descriptive label  and Formatting Data Values*/
PROC PRINT data = basic LABEL;
   label name = 'Name'
         clinic = 'Clinic'
         expense = 'Expense';
   format expense dollar9.2;
   id name;
   var clinic expense;
RUN;
```

- **SASHELP안에 내장된 자료 class의 summary**

-수치형 변수: proc means, proc univariate

-범주형 변수: proc freq

```
proc contents data=sashelp.class position; run;
```

```
title 'Frequency of sex';
proc freq data=sashelp.class;
       tables sex;
run;
proc freq data=sashelp.class;
       tables sex/nocum;
run;
```

결측치가 있는 경우: 이를 반영하려면 missing, missprint 옵션 사용

```
proc means data=sashelp.class;run;

proc means data=sashelp.class; var  age height; run;

proc means data=sashelp.class maxdec=2 fw=10;
       var  age height;
run;

proc means data=sashelp.class maxdec=2 fw=10 sum range median ;
       var  age height;
run;
/* Grouping */
proc means data=sashelp.class maxdec=2 fw=10;
       var  age height;
       class sex;
run;
/*Summary Table*/
PROC SORT data =sashelp.class out = class2;
  by sex;
RUN;

PROC MEANS data=class2 NOPRINT;
  var age height;
       by sex;
  output out = clsummary
            mean = MeanAge MeanHeight
            median = MedianAge MedianHeight;
RUN;
proc contents data=clsummary;run;
/* univariate procedure*/
proc univariate data=sashelp.class normal plot;
       var height;
run;
```

- **SASHELP안에 내장된 자료 cars**

```
proc contents data=sashelp.cars;run;
```

상자그림

```
proc sgplot data=sashelp.cars;
title "Box Plot: Category = Origin";
vbox Horsepower / category=Origin;
run;

proc sgplot data=sashelp.cars ;
title "Box Plot: Group = Origin";
vbox Horsepower / group=Origin;
run;

proc sgplot data=sashelp.cars;
title "Box Plot: Category = Cylinders, Linear Scale";
vbox horsepower / category=cylinders; /* cylinders: numeric */
xaxis type=linear;
run;

proc sgplot data=sashelp.cars;
where Type in ('SUV' 'Truck' 'Sedan');
title "Box Plot: Category = Origin, Group = Type";
vbox horsepower / category=Origin Group=Type;
run;
```

산점도

```
proc sgplot data=sashelp.cars;
  title"Vehicles: All Origins";
  scatter x=wheelbase y=weight / markerattrs=(symbol=CircleFilled);
run;

proc sgplot data=sashelp.cars;
  title"Vehicles: All Origins";
  scatter x=wheelbase y=weight/markerattrs=(symbol=CircleFilled)
  group=origin;
run;

proc sgplot data=sashelp.cars;
 title"Vehicles: Origin=USA only";
 where origin="USA";
 scatter x=wheelbase y=weight / markerattrs=(symbol=CircleFilled);
run;
```

● 과제

9월 14일까지 제출

- bweight.csv 파일을 읽어 데이터셋을 만들고

- 2개의 수치형 변수와 2개의 범주형변수를 선택하여

- 표와 그래프를 사용하여 요약 정리하고

- 간단히 설명하여

- 보고서 형식으로 제출하기

bweight 자료 설명

National Center for Health Statistics (Koenker and Hallock 2001; Abreveya 2001)에서 제공한 1997년 출생 체중 데이터

흑인 또는 백인으로 분류된 18세에서 45세 사이의 어머니의 신생아 체중 기록

1	Weight	신생아 체중
2	Black	흑인 어머니 여부
3	Married	기혼 어머니 여부
4	Boy	남자 아이 여부
5	MomAge	어머니 나이
6	MomSmoke	흡연 어머니 여부
7	CigsPerDay	하루 흡연량
8	MomWtGain	어머니의 임신기간 체중 증가량
9	Visit	산전 병원 방문 여부
10	MomEdLevel	어머니의 교육수준