한국통계학회논문집
2009, 16권, 3호, 437-447

통계 패키지에서의 데이터 접근 방식 비교

강근석[1,a]

[a]숭실대학교 정보통계보험수리학과

요 약

최근에 산업현장에서의 통계전문가들에게는 여러 가지 통계분석기법을 사용한 자료 분석 외에 다양한 형태의 자료 저장장치에서 추출 또는 생성의 과정을 거쳐 분석 목적에 적합한 자료를 구성해야하는 문제에 많이 부닥치고 있다. 본 논문에서는 현재 일반적으로 사용되고 있는 여러 통계 패키지들에서 제공하고 있는 데이터 접근방식을 살펴보고 각 기능들을 비교분석하고자 한다. 이들 방식에 대한 정확한 이해는 특히 데이터마이닝 등 대용량의 자료를 분석하고자 할 때 데이터 처리과정에서의 어려움으로 발생하는 비용과 시간을 감소시켜주어 통계전문가들이 통계분석에 더욱 많은 작업을 할애할 수 있도록 해줄 것이다.

주요용어: 통계 패키지, 데이터 접근, 데이터베이스 관리시스템.

1. 서론

통계학을 전공한 후 금융 또는 보험 등의 현장업무에 투입된 통계전문가들이 처음 부닥치는 어려움들 중 하나는 분석 자료의 구축 또는 가공 처리라 할 수 있다. 교육과정에서 통계분석을 배울 때는 대부분 단일 파일 형태로 제공되는 자료를 사용하여 분석하게 된다. 그렇지만 실제 현장에서는 대부분의 자료들이 데이터베이스에 존재하고, 또한 필요한 자료들이 여러 개의 파일에 나누어져 있는 경우가 많아 통계분석을 위해서는 필요한 변수들을 추출하여 합치거나, 기존의 변수들을 이용하여 새로운 변수들을 생성하는 과정이 필수적이다. 최근에 이와 같은 문제 해결을 위한 기법들을 소개하는 다양한 참고문헌들이 제공되고 있다 (손건태와 안상우, 2007; 최종후, 2008).

이와 관련하여 통계 패키지들에서도 접근 기능을 제공하여야 하는 데이터의 형태는 갈수록 다양해지고 있다. 다양한 표준 데이터 형식으로 존재하는 컴퓨터 파일들에 대한 접근 기능은 물론이고 많은 종류의 데이터베이스 관리시스템 내에 존재하는 분석 자료들에 대한 접근 기능을 제공하여야 한다.

오늘날 많은 양의 데이터를 다루는 기관이나 개인은 대부분 데이터를 특정 데이터베이스 관리시스템을 사용하여 데이터베이스 형태로 관리하고 있다. 데이터베이스를 구축하고 사용하는 이유는 데이터베이스 관리시스템, 특히 관계형 데이터베이스 관리시스템이 제공하는 데이터 관리의 편리성과 효율성 그리고 데이터 보안 기능 등 때문이다. 현재 많이 사용되는 관계형 데이터베이스 관리시스템들에는 Oracle, Sybase, DB2, SQL Server, Informix, PostgreSQL 등이 있다 (김형주, 2006; Silberschatz 등, 2005). 대부분의 프로그래밍 언어나 데이터 관련 소프트웨어에서는 이러한 데이터베이스들에 대한 접근을 위한 인터페이스가 제공되고 있으며 이를 효율적으로 지원하기 위한 다양한 데이터 접근 기술들이 개발되어 왔다. 각각의 데이터베이스 관리시스템이 제공하는 인터페이스를 사용한 접근기능들이 존재할 뿐만 아니라 표준화된 통합 데이터 접근 기술로 ODBC(Open DataBase Connectivity), OLE

본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음.
[1] (156-743) 서울 동작구 상도동 511 숭실대학교 정보통계보험수리학과, 교수. E-mail: gskang@ssu.ac.kr

DB(Object-Linking and Embedding Database), ADO(ActiveX Data Objects) 등이 주로 사용되고 있다 (Microsoft, 2009).

통계 패키지들도 다른 소프트웨어들과 마찬가지로 관계형 데이터베이스를 포함한 다양한 데이터 소스들에 대한 접근 기능들을 제공하고 있다. 대부분의 통계 패키지들은 접근을 지원하는 데이터 형태들을 더욱 다양화 할 뿐만 아니라, 접근 방식을 더욱 편리하게 그리고 효율적으로 지원하기 위하여 노력하고 있다. 특히 대용량의 자료를 분석하는 데이터마이닝의 경우에는 이러한 기능들이 필수적이며, 최근에는 여러 기업들에서 자료 유출을 막는 한 방안으로 모든 정보를 데이터베이스에 보관하면서 통계 패키지를 이용한 접근만을 허용하기도 한다.

본 논문에서는 일반적으로 많이 쓰이고 있는 통계 패키지들인 SAS, STATISTICA, SPSS, R, Minitab을 중심으로 통계 패키지들이 제공하고 있는 최근의 데이터 접근 기능들을 조사하고 분석한다. 2장에서는 이들 패키지들의 데이터 접근 방법들을 각각 알아보고, 3장에서는 패키지들의 데이터 접근 방식들을 비교 분석하며, 4장에서는 결론을 보인다.

2. 통계 패키지에서의 데이터 접근 방법 분석

이 장에서는 통계 패키지 SAS, STATISTICA, SPSS, R, Minitab이 다양한 데이터 소스들에 대해 제공하는 데이터 접근 기술들을 차례로 조사하고 분석한다.

2.1. SAS

SAS(www.sas.com)에서의 데이터 접근은 기본적으로 SAS/ACCESS 소프트웨어를 기반으로 하여 제공된다. SAS/ACCESS는 다양한 플랫폼에 존재하는 여러 종류의 데이터 소스들을 위하여 제공되는 데이터 접근 엔진들이다. 이 엔진들은 SAS 소프트웨어로부터의 데이터 읽기, 쓰기 등의 데이터 접근 요청을 특정한 데이터 소스에 대한 적합한 명령으로 변환하여 수행하게 한다. 그리고 그 결과는 데이터 소스에 대한 논리적 view로 나타내어지거나 또는 SAS 데이터로 추출되어진다. SAS/ACCESS는 관계형 데이터베이스, 비관계형 데이터베이스 그리고 ERP 시스템들을 비롯하여 다양한 데이터 소스들에 대한 접근 인터페이스를 제공한다 (SAS, 2008).

SAS/ACCESS는 접근하는 데이터들을 SAS 소프트웨어의 기능들을 이용하여 직접 분석하고 그 결과를 볼 수 있게 지원한다. 주된 접근 방법으로는 LIBNAME문을 이용한 접근 방법이 있다. SAS LIBNAME문에서 데이터베이스 관리시스템 엔진 이름과 연결 옵션들을 명시하여 데이터 소스를 SAS LIBREF에 연결시킨 후 데이터베이스의 테이블, 뷰 그리고 스키마 등의 객체들을 SAS dataset을 사용하듯이 DATA STEP이나 SAS 프로시저에서 사용할 수 있게 된다. 이러한 접근 방식은 데이터베이스에 있는 데이터들을 직접 읽게 되어 항상 최신의 데이터를 사용하게 하며 디스크 공간 절약효과도 가진다. 즉, 외부 데이터들을 SAS 데이터인 것처럼 투명하게 접근(transparent access)할 수 있도록 한다. 그러므로 SAS 사용자들은 SQL과 같은 데이터베이스 언어를 사용하지 않고 SAS의 기능들을 사용하여 데이터베이스의 데이터를 접근하고 사용할 수 있다.

한편으로 특정 데이터베이스 관리시스템의 SQL 문법을 알 필요 없이 SAS SQL 프로시저에서 SAS에서 지원하는 ANSI 표준 SQL을 사용해서 데이터베이스의 데이터를 접근할 수 있으며 데이터의 변경, 새 테이블의 생성 등도 가능하다. 이 과정에서 더 효율적인 작업 수행을 위하여 많은 작업들을 DBMS에서 행할 수 있도록 데이터베이스 접근 요청을 분석해 명령들을 데이터베이스 관리시스템에 보내고 나머지 작업을 SAS에서 처리한다. 다음은 SQL 프로시저를 사용하여 Oracle 데이터베이스에 있는 'delay'라는 이름을 가진 테이블을 접근하는 예를 보여준다.

```
LIBNAME mydblib ORACLE USER=scott PASSWORD=tiger;
```

```
TITLE 'Flights to London and Frankfurt';
PROC SQL;
    SELECT dates FORMAT=datetime9., dest
    FROM mydblib.delay
    WHERE (dest eq "FRANKF") or (dest eq "LONDON")
    ORDER BY dest;
QUIT;
```

이 경우 성능 향상을 위해 WHERE 절과 ORDER BY 절 부분을 데이터베이스 관리시스템으로 보내 수행하게 한다. TITLE과 FORMAT은 SAS에서 표준 ANSI SQL에 확장 추가한 기능이다.

데이터베이스에 대한 다른 접근 방법으로는 SAS 작업을 끝내지 않은 상태에서 데이터베이스 관리시스템에 직접 SQL 명령을 전달해 수행하게 하는 *Pass-Through* 기능을 이용하는 것이다. ANSI 표준 SQL을 사용해야 하는 일반 접근과는 다르게 이 경우 특정 데이터베이스 관리시스템이 지원하고 확장한 SQL 문법을 사용할 수 있게 한다. 또한 데이터베이스 관리시스템의 질의 최적화(query optimization) 기능을 사용하여 전체 질의를 최적화하게 하는 장점을 가진다. 다음은 pass-through 기능을 사용하여 Oracle DB에 접근하는 예이다.

```
PROC SQL;
    CONNECT TO ORACLE (USER=scott PASSWORD=tiger);
    CREATE VIEW budget2000 AS SELECT amount_b, amount_s
        FROM CONNECTION TO ORACLE
            (SELECT Budgeted, Spent FROM annual_budget);
QUIT;
```

PROC SQL을 호출하여 데이터베이스 관리시스템 이름과 적절한 연결 옵션과 함께 CONNECT문을 사용하면 원하는 데이터베이스와 연결된다. 연결된 데이터베이스 테이블로부터 데이터를 읽어 오기 위해서 SELECT문의 CONNECTION TO 절에 데이터베이스 관리시스템 고유의 문법을 사용하여 SELECT문을 작성한다. 그리고 이 SQL 명령을 직접 데이터베이스 관리시스템에 보내 실행하게 된다.

SAS/ACCESS LIBNAME문을 사용한 접근 방식이 일반적으로 더 빠르고 가장 직접적인 접근 방식으로 볼 수 있다. 하지만 ANSI 표준 SQL이 아닌 특정 시스템에서 지원하는 SQL을 사용해야 하는 경우에는 pass-through 기능을 활용할 수 있다. 또한 독립된 작업들의 동시 실행을 지원하는 'threaded reads' 기능도 지원하고 있으며 특히 이 기능은 병렬 데이터베이스 관리시스템을 효율적으로 접근하고자 할 때 유용하다.

비관계형 데이터베이스나 PC 파일들에 대해서도 *access descriptor*와 *view descriptor*를 작성함으로써 그 내부구조를 알 필요 없이 SAS 데이터인 것처럼 사용가능하다. 또한 이렇게 사용되는 데이터들은 SQL을 이용해서 다른 종류의 데이터들과 통합하여 다룰 수 있게 된다. SAS/ACCESS는 PC 파일들에 대해 view를 생성해 접근하는 방식 이외에 직접 데이터를 추출해 SAS 데이터 파일을 생성하는 방식도 허용한다.

ERP 시스템 데이터들에 대한 접근은 *SAS Data Surveyor*에 의해 제공되며 (SAS, 2009b), 각 ERP 시스템 데이터들의 복잡한 데이터 구조들을 탐색하고 이 데이터들을 SAS의 다른 데이터들과 통합 사용할 수 있게 하는 마법사 기능의 Graphical User Interface(GUI)가 지원된다. Data Surveyor가 제공되는 ERP 시스템으로는 Oracle, PeopleSoft, SAP, Siebel 등이 있다.

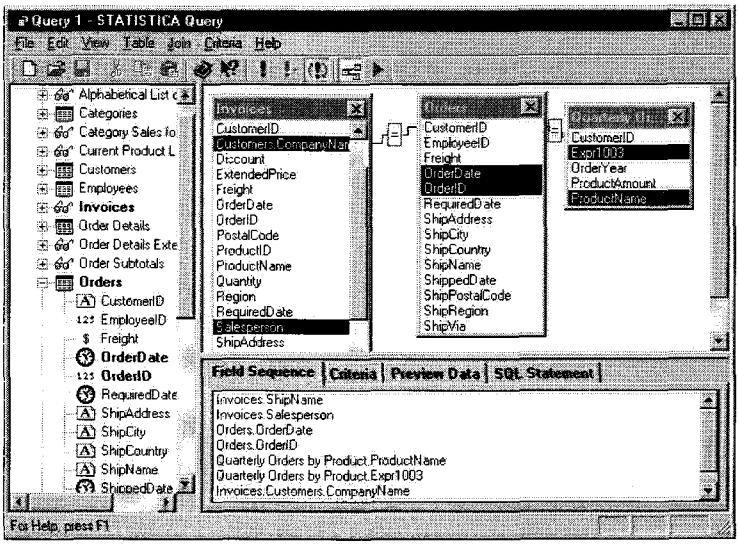

그림 1: *STATISTICA에서 복수 테이블 join 질의를 작성하는 예*

2.2. STATISTICA

STATISTICA(www.statsoft.com)에서의 데이터 접근은 STATISTICA Query를 통해 지원된다. 이 Query는 Microsoft의 OLE DB를 사용하여 관계형 데이터베이스를 포함한 다양한 데이터 소스로의 접근을 허용한다. 또한 SQL의 전문지식이 없이도 복수 테이블 질의를 작성할 수 있도록 QBE(Query By Example) 형태의 편리한 GUI를 제공하며, 질의 수행 결과는 STATISTICA 스프레드시트 형태로 얻어진다. 또한 다수의 외부 데이터베이스를 동시에 연결하는 기능도 있다 (Statsoft, 2009a). 그림 1은 Query 화면에서 GUI를 이용해 복수 테이블 join 질의를 작성하는 예를 보여 주고 있다.

STATISTICA의 새로운 데이터 접근 기능 중 하나는 *In-Place Database Processing*(IDP)으로서 외부 데이터를 importing과 복사의 과정 없이 직접 접근할 수 있게 한다 (Statsoft, 2009b). 즉, IDP 기능을 사용하지 않는 경우에는 사용자가 제시한 질의를 수행한 후 그 결과를 STATISTICA의 데이터 파일로 복사해 오게 되는 반면, IDP 인터페이스를 선택하게 되면 데이터를 외부에 둔 상태에서 분석이 진행되며 필요시에만 데이터를 부분적으로 가져오게 된다. 이러한 이유로 IDP는 질의 실행 결과의 데이터 크기가 큰 경우에 특히 성능 향상을 보이게 된다. IDP 인터페이스는 Microsoft의 ADO와 COM 기술을 활용해 구현하였으며, STATISTICA 스프레드시트 인터페이스의 일부를 구현하기 때문에 사용자는 다른 STATISTICA 스프레드시트와 유사한 방법으로 편리하게 사용할 수 있다.

2.3. SPSS

SPSS(www.spss.com/statistics)의 SDAP(SPSS Data Access Pack)는 주로 ODBC를 이용하여 데이터베이스의 데이터를 접근하도록 지원한다. Oracle, SQL Server, DB2, Microsoft Access 등의 ODBC driver들을 기본으로 제공한다. OLE DB를 통한 데이터 접근도 제한적으로 지원하고 있다. Database Wizard를 이용한 데이터베이스의 접근도 지원하며 Database Wizard의 export 기능을 이용하여 외부 데이터베이스로의 쓰기도 가능하다. SAS 데이터의 import 기능도 제공하며 SAS 변수와 값들을 SPSS 형식으로 자동 변환해 준다. 이 외에 Excel 데이터 등과의 import/export 기능도 제공한다 (SPSS, 2008). 다음은 Access의 테이블로부터 ODBC를 사용해 데이터를 읽어 오는 예이다.

```
GET DATA /TYPE=ODBC /CONNECT=
 'DSN=MS Access Database;DBQ=C:\examples\data\dm_demo.mdb;' +
 'DriverId=25;FIL=MS Access;MaxBufferSize=2048;PageTimeout=5;'
 /SQL =
 'SELECT Age, Education, [Income Category]' +
    'FROM CombinedTable' +
 'WHERE ([Marital Status] <> 1 AND Internet = 1)'.
EXECUTE.
```

데이터베이스로부터 읽기 위해 GET DATA문이 사용되었으며 ODBC driver를 이용하여 접근하고 있다. 데이터소스와 연결하기 위해 CONNECT 옵션이 쓰인다. 주로 CONNECT에 사용할 문자열은 Database Wizard를 이용하여 작성한 것을 사용한다. 그리고, SQL 부분에 접근하는 데이터베이스에서 지원하는 SQL문을 사용하여 원하는 데이터를 표현한다.

2.4. R

R(www.r-project.org)은 데이터의 조작, 저장, 분석 그리고 그래픽 표현을 지원하는 통합 소프트웨어 환경이다. S 언어의 구현으로 시작한 R 시스템은 확장 가능성을 매우 중요한 요소로 간주하며 만들어진 소프트웨어 환경이다. 이러한 R의 특성에 따라 기본 R 환경은 짧은 기간에 우선적으로 개발되었으며 다양한 패키지들을 통해 지속적으로 그 기능이 확장되고 있다.

R에서는 다양한 함수와 데이터들이 패키지를 통해 제공된다. 기본 패키지에는 R을 작동시키기 위한 기본 함수들과 데이터, 그리고 표준 통계함수와 표준 그래픽함수들이 들어 있다. 표준 패키지 또는 추천 패키지로도 불리는 25개 정도의 기본 패키지와 CRAN(CRAN.R-project.org)을 통해 다양한 저자들에 의해 제공되는 수백 개의 패키지들이 존재한다.

R의 데이터 접근 또한 기본 패키지를 포함한 여러 패키지들에 의해 제공된다. 데이터가 테이블 형태로 되어 있으며 행과 열 이름을 포함하기도 하는 스프레드시트 형태의 데이터들과의 import/export 기능은 `read.table` 함수에 의해 제공된다. 고정넓이 형식의 파일을 위해서는 `read.fwf` 함수가 있다. `read.table`이나 `read.fwf`는 text 파일을 위한 `scan`함수를 이용하며 매우 큰 수치 테이블을 읽는 데는 `scan`을 직접 이용하는 것이 더 효율적이다. 다른 통계 시스템의 데이터 파일을 import 하는 데는 *foreign* 패키지가 유용하며 이 패키지는 SAS, SPSS, Minitab, Stata 등의 다양한 통계 시스템으로부터의 데이터 import 기능을 제공한다 (R, 2008).

예를 들어, 함수 `read.spss`와 `read.mtp`를 이용하면 SPSS와 Minitab의 자료를 읽어 올 수 있으며, SAS의 경우에는 자료가 SAS transport format인 XPORT (SAS, 1989)로 되어 있으면 함수 `read.xport`를 사용하여 아래와 같이 직접 읽어올 수 있다.

```
> ## 자료 sample.dat가 XPORT 형식으로 있는 경우
> read.xport("sample")
```

또한 자료가 일반적으로 많이 사용하는 SAS permanent dataset 형태로 있는 경우(보통 확장명이 `.ssd0x` 또는 `.sas7bdat`)에는 함수 `read.ssd`를 이용하면 먼저 XPORT 형태로의 자료변환을 하게 되고 순차적으로 `read.xport`를 이용한 읽기를 수행하게 할 수 있다. 이 경우 XPORT 형태로의 자료변환은 SAS 프로그램을 이용하므로 SAS 패키지가 있어야 한다. 예를 들어, SAS 실행파일이 "/Program Files/SAS/SAS 9.1"에 설치되어 있고, `statdata`라는 폴더 내에 있는 `survey.sas7bdat`를 읽고자 하는 경우에는

```
> mydata <- read.ssd("c:/statdata", "survey",
                              sascmd="/Program Files/SAS/SAS 9.1/sas.exe")
```

와 같이 실행하면 자료들이 SAS에서 선언된 변수명과 함께 R data frame으로 저장된다(Hmisc 패키지의 `sas.get` 함수를 사용할 수도 있다.).

데이터를 text 파일로 export하기 위한 기본 함수로는 `cat`이 제공된다. `write.table`과 `write` 함수를 이용하면 테이블 형태로 데이터를 export 할 수 있으며, 결측값이나 구분자 등 다양한 옵션을 명시할 수 있다. 행과 열 이름을 함께 출력하려면 `write.table`을 이용한다. 행렬을 효율적으로 출력하기 위한 `write.matrix` 함수가 MASS 패키지에서 제공된다. 다른 통계 패키지로의 export는 SPSS와 Stata로의 export가 foreign 패키지의 `write.foreign` 함수를 통해 지원된다. XML 패키지는 R에서의 XML 문서 읽기와 쓰기를 위한 일반적인 기능들을 제공한다.

R이 다루는 데이터는 모두 메모리에 상주(resident)하는 것으로 대용량의 데이터를 다루는 데는 적합하지 않다. 또한 R은 데이터에 대한 동시 접근을 잘 지원하지 못한다. 반면에 데이터베이스 관리 시스템들, 특히 관계형 데이터베이스 시스템들은 대용량 데이터에 대한 효율적이고 편리한 검색, 여러 client로부터의 동시 접근 지원, 보안 등 데이터관리를 위한 다양하고 강력한 기능들을 제공한다. R은 이러한 데이터베이스 관리시스템들과의 인터페이스를 CRAN의 다양한 패키지들을 통해 제공하고 있다. 이 패키지들은 데이터베이스의 데이터를 R의 데이터 프레임으로 가져올 수 있게 하고 반대로 데이터베이스에 R의 데이터를 저장할 수 있게도 한다. 또한 SQL 질의를 사용해 데이터베이스의 데이터 일부를 검색해 가져오는 기능도 제공한다. RODBC를 제외하고는 이 기능들은 특정 DBMS에 대한 인터페이스를 따로 제공한다. 하지만 이러한 인터페이스들도 R의 통합 front-end 패키지인 DBI를 통하여 사용할 수 있게 하기 위한 작업들이 진행 중에 있다. 이 통합 DBI 패키지를 지원하는 back-end 패키지로는 MySQL을 위한 RMySQL, Oracle을 위한 ROracle 그리고 SQLite를 위한 RSQLite 패키지들이 있다 (R, 2009).

다음은 DBI를 이용한 MySQL 데이터베이스 접근 예를 보여 주고 있다.

```
> library(RMySQL) # RMySQL과 함께 DBI load
##  MySQL 데이터베이스 "test" 에 연결
> con <- dbConnect(dbDriver("MySQL"), dbname = "test")
## 데이터베이스의 테이블을 나열
> dbListTables(con)
## "USArrests" data frame을 데이터베이스에 load
> data(USArrests)
> dbWriteTable(con, "arrests", USArrests, overwrite = TRUE)
> dbListTables(con)
[1] "arrests"
## 테이블 arrests를 읽어옴
> dbReadTable(con, "arrests")
Murder Assault UrbanPop Rape
Alabama 13.2 236 58 21.2
Alaska 10.0 263 48 44.5
...
## select query 실행
> dbGetQuery(con, paste("select row_names, Murder from arrests",
```

```
"where Rape > 30 order by Murder"))
row_names Murder
1 Colorado 7.9
2 Arizona 8.1
...
> dbRemoveTable(con, "arrests")
> dbDisconnect(con)
```

DBI와 RMySQL을 사용하여 MySQL 시스템의 데이터베이스를 접근하는 위의 예를 보면, 먼저 `dbDriver("MySQL")`을 호출하여 connection manager 객체를 돌려받고 이와 연관된 `dbConnect`를 호출하여 MySQL 데이터베이스에 연결하게 된다. 연결된 후 `dbGetQuery`는 SQL 질의를 보내 그 결과를 data frame에 받는다.

이와는 다르게 `dbSendQuery`는 질의를 보내 그 결과를 DBIResult의 하위 객체에 돌려받는다. 질의 결과의 일부 또는 전부를 가져오려면 `fetch` 함수를 적용하면 된다. 또한 `dbReadTable`과 `dbWriteTable`들을 이용하면 데이터베이스 테이블과 R의 자료구조 상호간의 변환을 쉽게 할 수 있다.

RODBC 패키지가 제공하는 기능들을 이용하면 ODBC 인터페이스를 지원하는 데이터 소스들을 접근할 수 있다. 대부분의 데이터베이스 시스템들이 ODBC를 통한 접근을 지원하기 때문에 R에서도 같은 프로그램을 사용하여 다양한 데이터베이스 시스템에 접근할 수 있다는 장점을 가진다.

2.5. Minitab

Minitab(www.minitab.com)의 주 window는 네 종류의 sub-window들을 포함할 수 있다. 하나 이상의 Data window들이 데이터 칼럼들을 보여주며, 프로젝트의 각 워크시트에 대해 한 개의 Data window가 존재한다. 결과를 보여주는 Session window, Minitab의 그래프 명령들에 의해 만들어지는 고해상도의 그래프들을 위한 Graph window 그리고 Project Manager들이 있다.

Minitab은 ODBC driver가 설치되면 원하는 데이터베이스 파일에 연결하여 Query Database 메뉴를 사용하여 데이터를 가져올 있도록 한다 (Minitab, 2009). Query Database의 대화상자를 이용하여 한 테이블의 행과 열 일부를 선택하는 단순한 질의만 수행할 수 있으며, ODBC session 명령을 사용하면 여러 테이블들을 연결(join)하는 질의도 수행할 수 있다. Excel, Quatro, Lotus, dBase 등의 다양한 PC 파일들에 대한 import 기능을 제공하고 export 기능도 일부 제공하고 있다.

3. 통계 패키지에서의 데이터 접근 비교 분석

ACCESS 소프트웨어를 사용하여 데이터 접근 기능들을 제공하는 SAS는 ODBC, OLE DB를 통한 데이터 접근과 함께 다양한 접근 기능들을 제공하고 있다. 기본적인 import/export 기능과 함께 데이터베이스 관리시스템에 접속을 유지하면서 데이터베이스의 테이블, 뷰 등에 접근하게 하는 'direct access', 데이터베이스의 데이터를 접근하면서 사용자가 SAS 데이터인 것처럼 다룰 수 있게 하는 'transparent access' 기능을 제공한다. 또한, 필요한 경우 특정 DBMS의 SQL 문법을 사용하여 데이터베이스에 접근할 수 있도록 SQL pass-through 기능을 제공한다. 자체 데이터에 대한 SQL 사용이 가능하며 query 화면에서 GUI를 이용한 SQL 작성을 지원한다.

STATISTICA는 ODBC, OLE DB 접근을 지원하며 데이터베이스 연결 후 STATISTICA 스프레드시트를 통한 transparent access 기능을 제공한다. 질의 수행결과를 바로 STATISTICA로 가져오지 않고 분석에 필요한 경우에만 가져오게 하는 in-place database processing 기능을 새롭게 지원하고 있다.

표 1: 통계 패키지들에서의 데이터 접근 방식 비교

기능	SAS	STATISTICA	SPSS	R	Minitab
데이터 접근 표준 지원	ODBC, OLE DB	ODBC, OLE DB	ODBC, OLE DB	ODBC	ODBC
SQL 지원 (접근 데이터)	직접 접근, 투명 접근, pass-through	직접 접근, 투명 접근	직접 접근	직접 접근	
SQL 지원 (기타)	*threaded reads*, 다중 DB 접근, 자체 데이터 SQL 지원	in-place 접근			
기타	ERP 데이터 접근, *SQL GUI*	SQL GUI	SQL GUI		SQL GUI

SPSS는 ODBC, OLE DB를 통한 데이터 접근을 지원하며 SAS로부터의 데이터 import와 Excel 데이터로의 *export* 기능을 제공한다. 데이터베이스 마법사에서 SQL GUI 기능을 제공한다.

R은 ODBC를 사용한 데이터베이스 접근을 기본적으로 사용하며 다른 데이터베이스 통합 인터페이스로서 DBI를 통한 데이터베이스 접근기능도 다양한 DBMS에 대하여 추가해 가고 있다. 그 밖에, 스프레드형식의 데이터와 사용자 정의 형식의 데이터 *import/export*, SAS, SPSS, Minitab, Stata의 데이터 import 기능들이 존재하며 XML 문서 읽기와 쓰기도 지원된다.

Minitab은 ODBC를 사용한 데이터베이스 접근을 허용하며, Query Database 명령을 이용하여 한 테이블에 대한 단순한 명령만 GUI를 사용하여 표현가능하며 복수 테이블 연결은 ODBC session 명령을 사용하여 작성해야 한다. 표 1은 지금까지 비교 분석한 내용을 정리한 것이다.

SAS는 자체 데이터를 조작하는 기능도 데이터베이스 관리시스템 수준으로 높이 제공하지만 가장 강력하고 다양한 데이터 접근 기능들을 제공하고 있으며 지속적으로 기능을 개선해 나가고 있다. STATISTICA도 비교적 광범위한 데이터 접근 기능을 제공하고 있으며 최근에 기능들을 더 보완하고 추가하여 SAS 다음으로 강력한 기능들을 제공한다고 볼 수 있다. SPSS와 R은 아직은 상대적으로 평범한 데이터 접근 기능을 제공하고 있으며 Minitab은 ODBC를 사용한 데이터 import 기능 정도의 가장 미약한 데이터 접근 기능을 제공하고 있다.

4. 결론

통계 패키지들은 산업 현장 또는 통계분석 기법에서의 요구에 맞추어 데이터베이스를 포함하여 다양한 형태의 데이터 소스에 대한 편리한 데이터 접근 기능들을 제공하고 있고, 지속적으로 그 기능들을 개선해 나가면서 새로운 기능들을 추가해 나가고 있다.

본 논문에서는 일반적으로 많이 쓰이고 있는 통계 패키지들인 SAS, STATISTICA, SPSS, R, Minitab이 제공하고 있는 최근의 데이터 접근 기능들을 조사 분석하고 비교하였다. SAS와 STATISTICA가 가장 강력한 데이터 접근 기능들을 제공하고 있다. ODBC, OLE DB를 사용한 데이터 접근뿐만 아니라 데이터 소스와의 연결 상태에서 필요시 직접 데이터들을 접근 할 수 있고 사용자들은 SQL과 같은 데이터베이스 접근 언어를 배울 필요 없이 자체 시스템의 데이터와 같은 방법으로 데이터 소스를 접근하고 사용할 수 있도록 지원하고 있다. 더 나아가 자체 데이터들에 대해서도 SQL을 사용해 원하는 데이터를 표현할 수 있으며 이 경우 편리한 GUI를 지원하고 있다.

통계 패키지들이 지원하는 데이터 처리 기능과 데이터 접근 기능은 현재도 계속 보강되고 더욱 강력해지고 있다. 특히 SAS의 경우 SAS 엔진과 데이터베이스 관리시스템을 더욱 밀접하게 연결시켜 SAS의 일부 기능을 데이터베이스 관리시스템으로 옮기는 SAS In-Database Processing이라고 부

르는 기술을 지원하기 위한 계획을 발표하였다 (SAS, 2007). 그리고 데이터 처리와 접근뿐 아니라 ETL(Extract, Transform and Load), data synchronization, data migration, data federation 등의 기능을 모두 지원하는 데이터 통합환경을 구축하여 제공하고 있다 (SAS, 2009a).

통계 패키지들에서 데이터베이스 관리시스템의 기능들을 최대한 활용하려는 노력들과 함께 데이터베이스 관리시스템들에서도 자체적으로 통계 처리 기능들을 제공하려는 시도들도 존재한다. 대표적으로 Oracle에서는 Oracle OLAP을 통해 다차원자료에 대한 다양한 분석기능을 직접 제공하고 있으며, 더 나아가 Oracle Business Intelligence 제품들에서는 강력한 통계분석기능을 포함한 통합환경을 제공하고 있다 (Oracle, 2005).

결론적으로 통계 패키지들은 데이터베이스 관리시스템을 포함한 다양한 데이터 소스에 대하여 데이터베이스 기술을 활용한 강력한 데이터 접근 기능들을 추가해 나가고 있을 뿐만 아니라 통계 패키지와 데이터베이스 관리시스템의 경계를 무너뜨리는 두 시스템의 융합 현상까지 나타나고 있다. 이러한 시도는 특히 대용량 자료를 취급하면서 다양한 형태의 분석기법을 사용하고 있는 산업현장의 요구를 수용해나가는 것으로 볼 수 있다. 실제로 최근에 일반 기업으로부터 이러한 작업들을 수행할 수 있는 인력에 대한 지원요청이 급증함을 경험하고 있다. 따라서 통계학 교육과정에서도 단순히 주어진 자료를 분석하는 기법을 가르치는 것 외에도 통계 패키지에 포함된 다양한 데이터 접근 기능들을 활용하여 필요한 자료를 구축하고 가공하는 기법을 교육내용에 포함시킬 필요가 있다고 할 수 있다.

참고 문헌

김형주 (2006). <데이터베이스 시스템>, 한국맥그로힐, 서울.

손건태, 안상욱 (2007). <SAS DATA STEP : 기초편>, 자유아카데미, 서울.

최종후 (2008). <SAS DATA STEP>, 자유아카데미, 서울.

Microsoft (2009). Win32 and COM Development, *MSDN Library,* Available from `http://msdn.microsoft.com/en-us/library/aa968814.aspx`.

Minitab (2009). Data and File Management, *Online Documentation,* Available from `http://www.minitab.com/products/minitab/features`.

Oracle (2005). Oracle Business Intelligence: Concepts Guide, *Online Documentation,* `http://download.oracle.com/docs/cd/B14099_19/bi.1012/b16378.pdf`.

R Development Core Team (2008). R Data Import/Export, Version 2.8.0, Available from `http://cran.r-project.org/doc/manuals/R-data.html`.

R Development Core Team (2009). The R interface packages, Available from `http://cran.r-project.org/doc/manuals/R-data.html#R-interface-packages`.

SAS Institute Inc. (1989). The Record Layout of a Data Set in SAS Transport (XPORT) Format, *SAS Technical Support document TS-140,* `http://support.sas.com/techsup/technote/ts140.pdf`.

SAS Institute Inc. (2007). SAS In-Database Processing: A Roadmap for Deeper Technical Integration with Database Management Systems, *Technical Paper,* `http://support.sas.com/resources/papers/InDatabase07.pdf`.

SAS Institute Inc. (2008). SAS/ACCESS 9.2 for Relational Databases: Reference. Cary, NC: SAS Institute Inc., Available from `http://support.sas.com/documentation/cdl/en/acreldb/59618/PDF/default/acreldb.pdf`.

SAS Institute Inc. (2009a). The New Data Integration Landscape: Moving beyond ad-hoc ETL to an enterprise data integration strategy, *White Paper,* Available from `http://support.sas.com/apps/whitepaper/index.jsp?cid=3498`.

SAS Institute Inc. (2009b). SAS Data Surveyors, *Online Documentation,* Available from `http://www.sas.com/technologies/dw/etl/surveyors`.

Silberschatz, A., Korth, H. F. and Sudarshan, S. (2005). *Database System Concepts,* McGraw-Hill, New York.

SPSS Inc. (2008). Data Access Pack Installation Instructions for Windows, *Online Documentation,* Available from `ftp://ftp.spss.com/pub/web/drivers/sdap/Documentation/SDAP/en-us/sdapwin.pdf`.

Statsoft (2009a). STATISTICA Query, *Online Documentation,* Available from `http://www.statsoft.com/uniquefeatures/query.html`.

Statsoft (2009b). The In-Place Database Processing(IDP) Technology, *Online Documentation,* Available from `http://www.statsoft.com/products/idp.html`.

2009년 1월 접수; 2009년 2월 채택

Comparing Data Access Methods in Statistical Packages

Gunseog Kang[1,a]

[a]Department of Statistics & Actuarial Science, Soongsil University

Abstract

Recently, in addition to analyzing data with appropriate statistical methods, statistical analysts in the industrial fields face difficulties that they have to compose proper datasets for analysis objectives via extracting or generating processes from diverse data storage devices. In this paper we survey and compare many state-of-the-art data access technologies adopted by several commonly used statistical packages. More understanding of these technologies will help to reduce the costs occurring when analyzing large size of datasets in especially data mining works, and so to allow more time in applying statistical analysis methods.

Keywords: Statistical packages, data access, database management system.

This work was supported by the Soongsil University Research Fund.

[1] Professor, Department of Statistics & Actuarial Science, Soongsil University, Sangdo-Dong, Dongjak-Gu, Seoul 156-743, Korea. E-mail: gskang@ssu.ac.kr