

> SPSS를 이용한 “R”연동 기능소개와 분석 기능의 시너지 효과

2011.10.28(금)
SPSS Korea Data Solution Inc.,
허준

> R User Conference & SPSS

주위의 반응



> R이란 무엇인가? 그리고 R의 장점은?

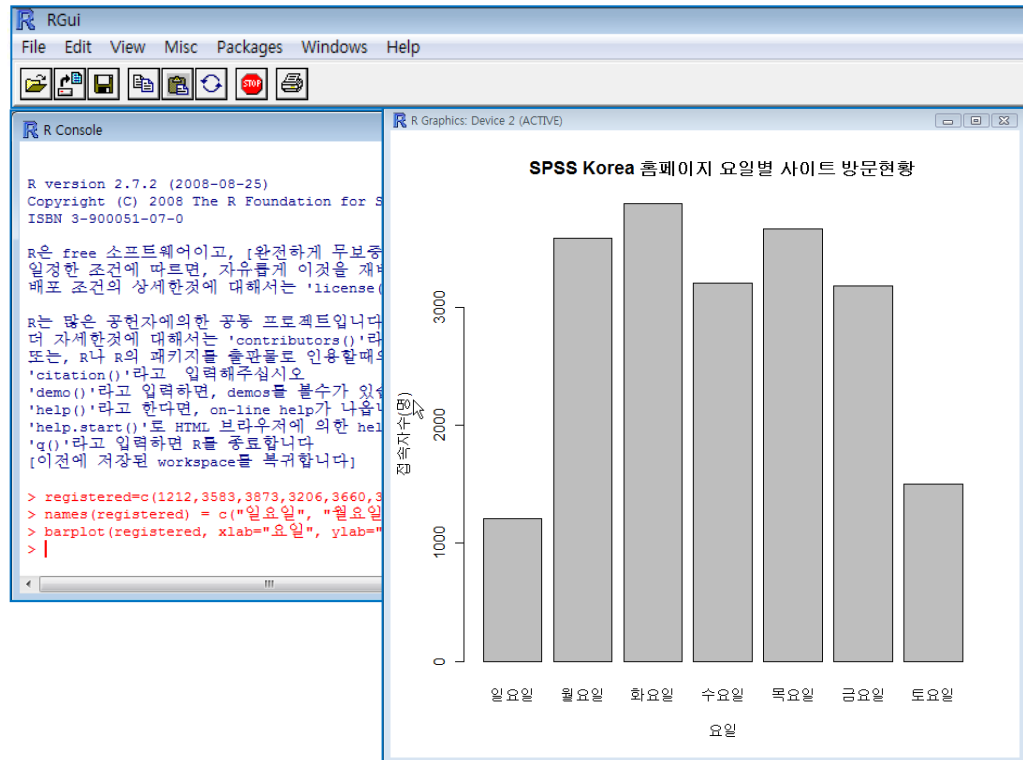
What's R?

- R은 R코어팀(Core Team)에서 R 프로젝트라는 이름으로 지금도 계속 개발되고 있는 통계적 계산과 그래픽을 위한 컴퓨터 언어이며 환경임. 처음 뉴질랜드의 오클랜드 대학교에서 교육용으로 개발되었으나 소스 코드를 인터넷에 공개하면서 리눅스의 보급과 더불어 급속히 확산 되어 현재 전세계적으로 사용되고 있음.(현재 2.13 버전)

R 소개

R 장점 및 특징

- 다양한 통계적 분석 기법 보유
- 우수한 그래픽 방법 제공
- 뛰어난 프로그램 기능을 통해 사용자가 새로운 함수를 작성하여 추가할 수 있음
- AT&T사의 Bell 연구소에서 개발한 S언어를 윈도 체계에서 구현한 상용 통계패키지 S-PLUS와 유사함. 따라서 현재 사용중인 S-PLUS에 관한 매뉴얼이나 책을 R에서 그대로 사용해도 됨 (약간의 차이점 존재)
- 무료 소프트웨어
- 개발자(분석자)들이 자신의 패키지(분석기법)을 만들어서 상호 간 공유하는 형태로 일단 표준 공통이 없고, 자유로운 형태의 관리 및



> R의 단점

R은 오픈 소스로써, 매우 만족스런 기능을 제공하지만, 아직 일부 한계가 있음.

(아직까지)불편한 User Interface 및 데이터 핸들링

- GUI가 아닌 **Script** 형태로 초보자의 접근이 용이 하지 않음
- 시각적이지 않은 **User Interface**로 인하여, 다양한 변환이 쉽지 않음
- 데이터 핸들링 부분 및 각종 데이터 종류별 접근이 어려움
- 분석의 변환 과정 및 데이터의 관리, 이동 및 삭제 등이 어려움

시스템화에 많은 장애요소

- 일반 개인 **User** 및 단순 연구 목적의 패키지 성격이 강함.
- 각종 시스템으로의 전이 및 개발이 매우 어려움.→ 근래 상업화를 통해 해결 중
- 유지보수 및 일반적인 체계적인 관리가 어려움.
- 비용을 지출하여, 관리를 하고 싶어도 관리주체 등의 선정 및 발견이 어려움

기타

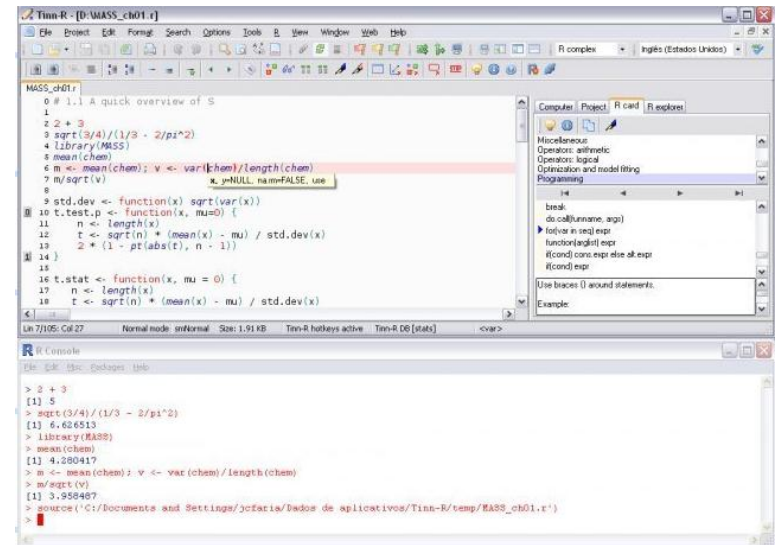
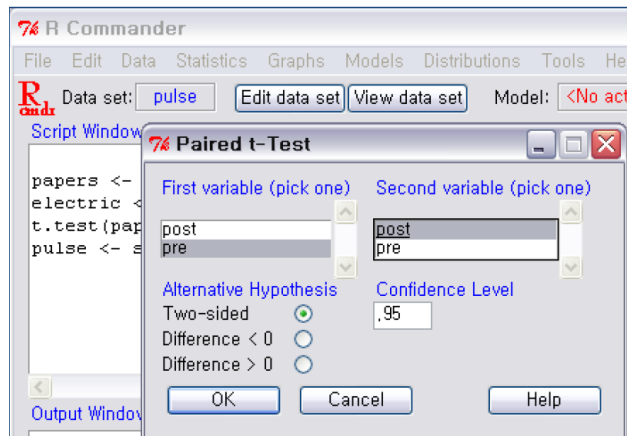
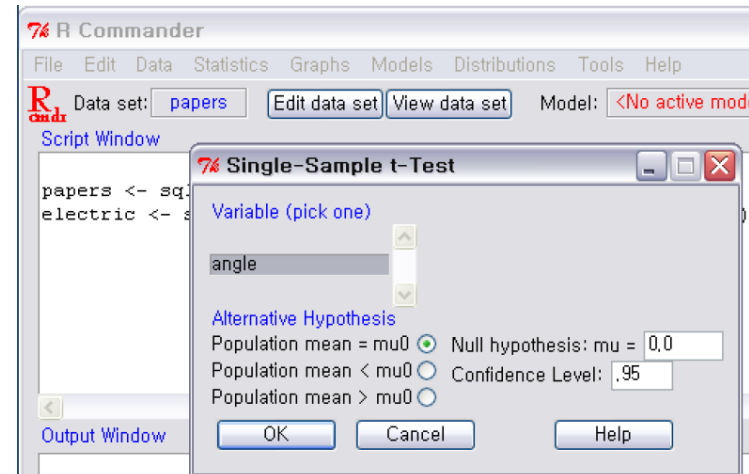
- R 기반 자체가 일반 **User** 중심으로 인한 개발로 인하여, 다양한 **Server** 및 시스템에 최적화 되어 있지 않음.
- 분석결과에 있어, 사용자의 배려가 적어서, 분석 후 사후 해석 등에서 불편함을 초래함.(ex: 로버스트 회귀 후 계수의 **p-value** 출력 없음)
- 분석 기법의 다양한 옵션을 기억하기가 어려워, 특정 옵션만을 사용하여 분석의 폭이 좁을 수 있음.

> R의 한계 및 단점... 그러나

그러나 R의 상용화(유지보수, 기술지원)와 편리한 사용을 위한 R의 진화는 계속...



R+Evolution의 노먼 나이 교수(전자신문 인용)



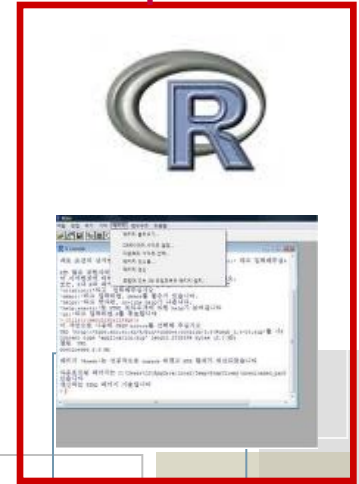
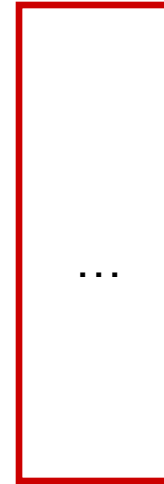
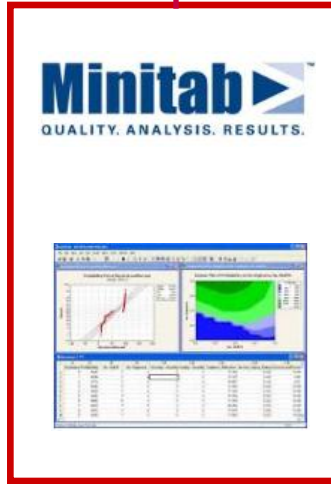
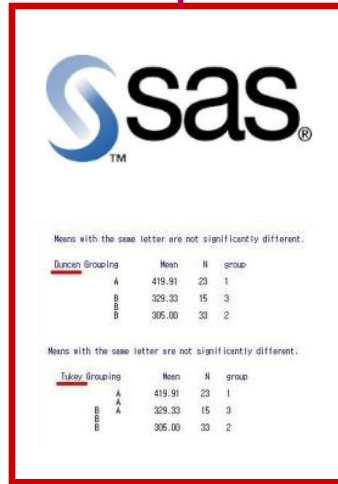
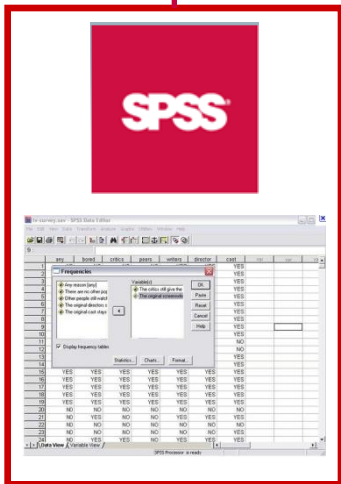
Tinn-R is free, simple but efficient replacement for the basic code editor provided by Rgui.

> 통계분석 패키지의 비교

사람들은 각종 통계 패키지를 계속 비교하고, 우위를 논합니다.

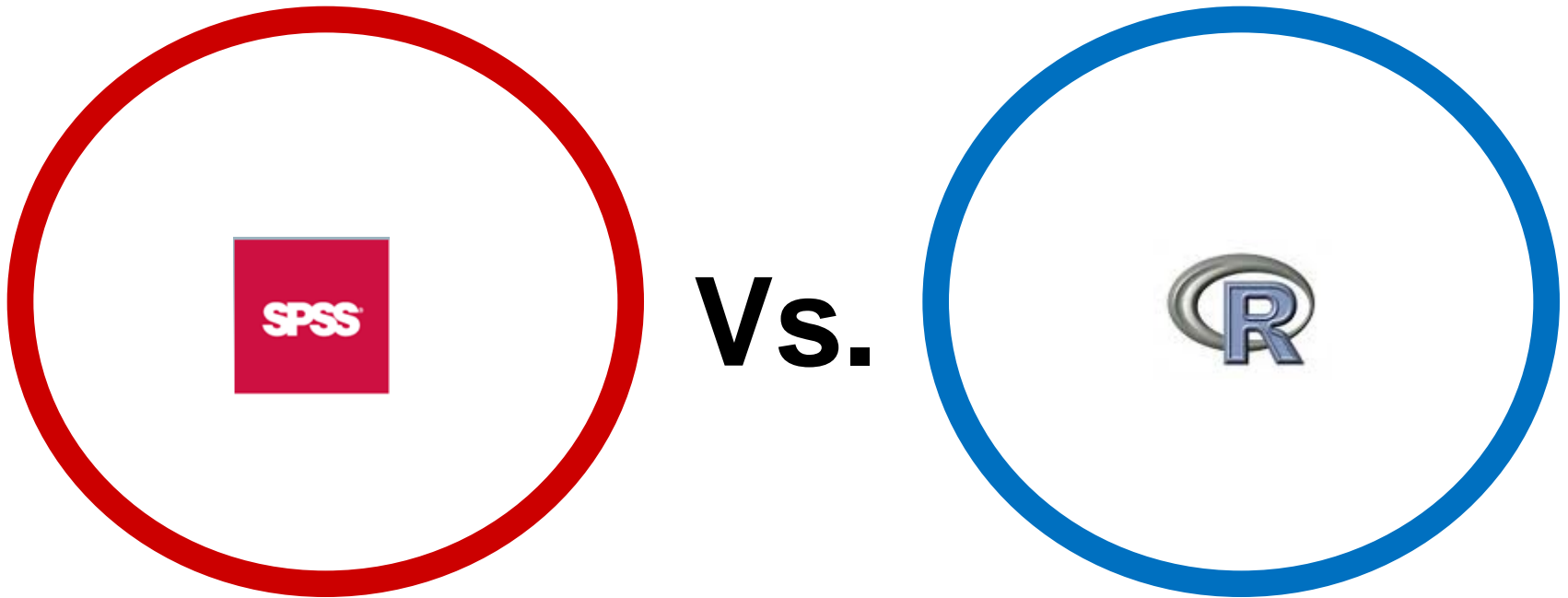


가격? / 성능? /
사용의 편리성? /
회사 내 숙련자의 수?



> SPSS와 R은 경쟁 대상인가?

통계 패키지를 비교할 때, (IBM)SPSS와 R은 경쟁 상대인가?



- 현재 대다수의 경우 R을 SPSS의 대체 패키지로 인식(특히 가격부분)
- 가격 및 일부 기능을 이유로 R로 SPSS를 바꾸고, 2개 간의 비교를 중시
- 그러나 다른 관점에서 바라보자...

> 기업과 조직의 통계 분석 생태계

전체 통계분석 생태계 별로 똑같은 통계 패키지가 필요할 것인가?

고급통계
사용자



통계학 교수, 관련
연구소 연구원, 전문
**Data Miner, Data
Mining Consultant**

일반 데이터 분석가



기타 사회과학
연구원, 일반 기업
마케팅 분석가, 전략
Consultant 등

초급 데이터 분석자



일반기업 업무 관련
데이터 분석가 및
업무 분석가

> 목적이 중요하지 Tool이 중요하지 않다

산을 오르는 것이 목적이지 좋은 피켈(Pickel)을 구매하는 것이 목적은 아니다.

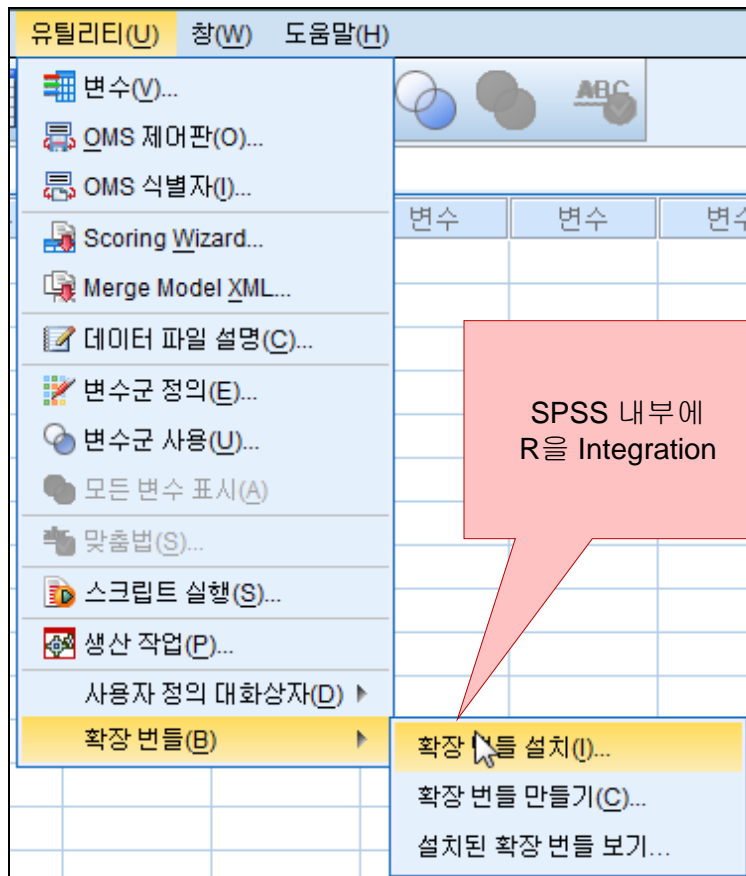


통계 패키지 자체에 유념하기
보다는 현재 업무와 그것을
달성하기 위해 어떤 방법이
좋은 지 고민해야 하고,
반드시 1개만 써야한다는
고정관념은 버리자!!!

> IBM SPSS Stat.과 R과의 결합

SPSS Statistics 17버전 이후 R과의 결합으로 상호 Win-Win

- SPSS Stat. 17버전 이후 19버전까지 R과의 결합을 통해서, SPSS 내부에서 R을 자유롭게 사용하고, R의 단점을 획기적으로 보완하였습니다.



SPSS R-essential이라는 R연동 Middle ware를 통해서 SPSS 내부에서 R의 명령어와 모든 기능을 사용할 수 있도록 개발

SPSS 에서 사용하는 데이터 기반 및 SPSS 자원을 이용한 R 데이터 분석

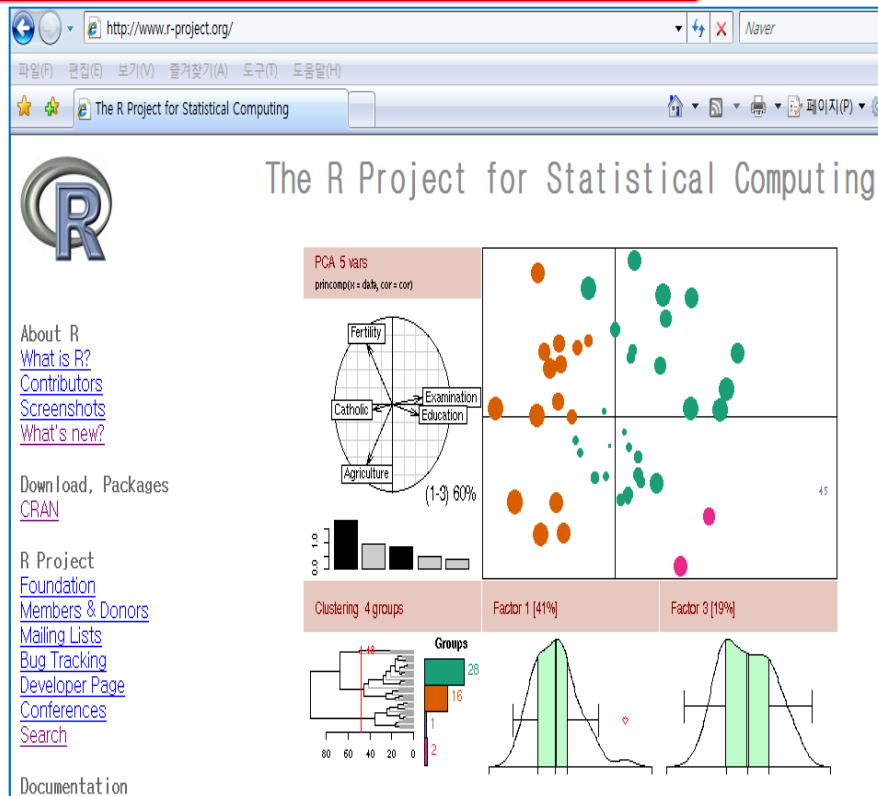
SPSS Stat Server와 Client 모두에서 작동 지원

> 왜 SPSS에서는 R이 필요한가?

가장 최신의... 가장 많은 알고리즘 보유

- 오픈 소스의 강점으로, 매우 희귀한 분석이나 다양한 최신의 알고리즘이 R을 통해서, 자유롭게 공급 배포되고 있음. (일반 상용 통계패키지가 접근이 어려운 점)

R을 무료로 다운로드 하려면?



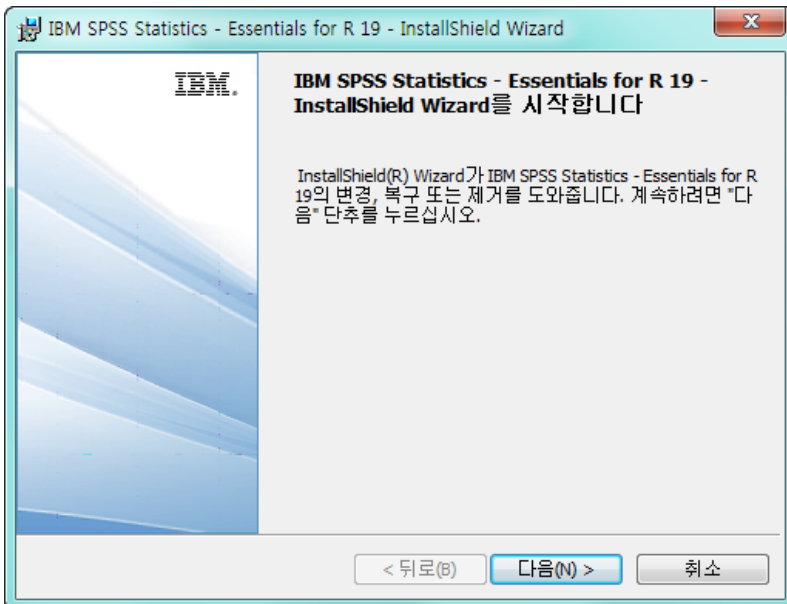
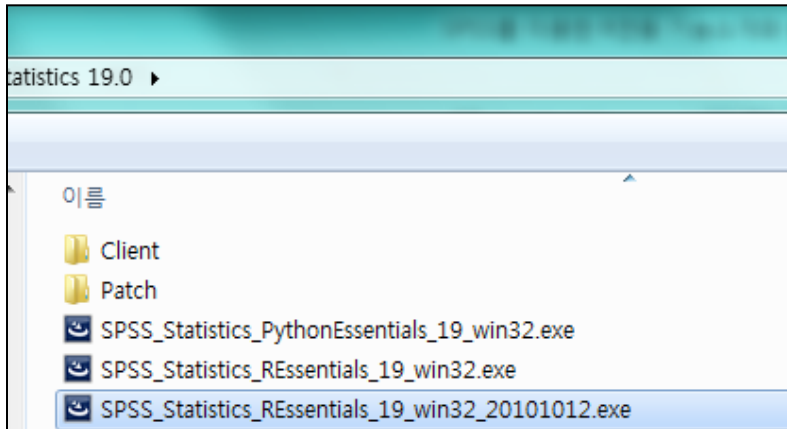
R 공식 홈페이지 : <http://www.r-project.org>

- 이 홈페이지는 CRAN(Comprehensive R Archive Network)에 연결되어 있어, CRAN 사이트에서 R을 내려받아 쉽게 개인 컴퓨터에 설치할 수 있다.

CRAN>Korea에 해당하는 주소(<http://bibs.snu.ac.kr/R/>)>xp나 vista사용자일 경우(Windows)>base>download R

> SPSS-R Essentials

SPSS Statistics 17버전(권장은 18 이후) 이후 R과의 결합으로 상호 Win-Win 달성



SPSS R-Essentials

- **SPSS**과 **R**을 별도로 설치 후 **SPSS R-Essentials**를 설치하면, **SPSS**와 **R**이 자동 연동
- 간편한 **GUI** 방식의 설치로, 초보자도 쉽게 연동이 가능함.
- 단, 특정 버전의 **R** 패키지만을 지원 (예를 들어 **SPSS 19**버전의 경우 **2.10.1**버전)

SPSS R-Essentials 특징

- **SPSS** 데이터를 이용하여, **R** 분석을 함에 따라 **SPSS**의 기본 성능/자원을 이용(특히 데이터 입출력 파트)
- 별도 비용 없음(**SPSS**와 같이 제공)
- **SPSS**의 기능 동시 사용 가능
- 한글지원

> IBM SPSS Stat. with R의 장점

더욱 더 강력해진 SPSS, 활용도가 높아진 R

막강한 데이터 분석 기능

- 기존 SPSS Stat에서 지원하지 않는 다양한 분석 모듈을 SPSS 내에서 사용이 가능
- R 자체의 장점인 수려한 R 그래프를 그대로 SPSS Stat Output 창에서 활용 가능
- 새로운 분석 기능 추가시 R에서 동일하게 패키지 다운로드 후 그대로 SPSS Stat 내부에서 사용 가능

체계적인 분석 시스템화

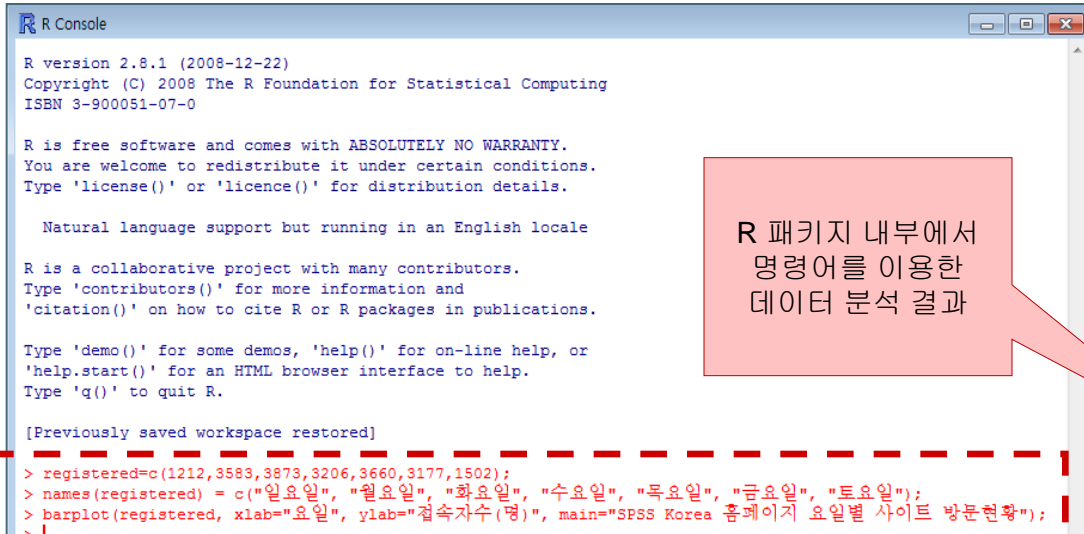
- R의 단점인 부족한 데이터 핸들링 기능을 전부 SPSS에서 핸들링 후 R을 이용한 분석이 가능
- R 분석 결과의 시스템화 및 자동 수행 등 체계적인 시스템으로 개발이 가능함
- SPSS와 함께 유지보수 등을 체계적으로 지원 받을 수 있음.(단, R 자체적인 지원은 제외)

대용량 처리 등 분석의 효율화

- SPSS 자원을 활용하므로, 일반적으로 데이터 로딩 및 적재 등 전반적인 대용량 성능 처리가 향상
- 분석 과정 중 전처리 및 향후 데이터 출력은 모두 SPSS의 기능 사용으로 분석의 효율화
- 기본적으로 SPSS의 편리한 User Interface를 사용하거나 응용할 수 있어 분석 초보자들에게 편리성 제공

> IBM SPSS Stat. with R(1)

기존의 R 명령어를 동일하게 SPSS에서, 사용이 가능함.



```
R Console
R version 2.8.1 (2008-12-22)
Copyright (C) 2008 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

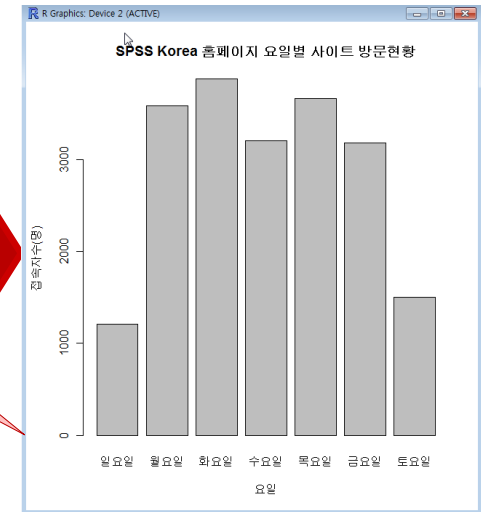
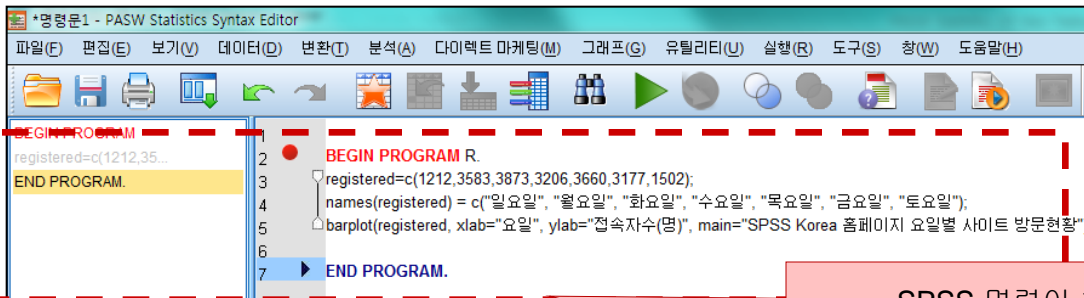
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> registered=c(1212,3583,3873,3206,3660,3177,1502);
> names(registered) = c("일요일", "월요일", "화요일", "수요일", "목요일", "금요일", "토요일");
> barplot(registered, xlab="요일", ylab="접속자수(명)", main="SPSS Korea 홈페이지 요일별 사이트 방문현황");
>
```

R 패키지 내부에서
명령어를 이용한
데이터 분석 결과

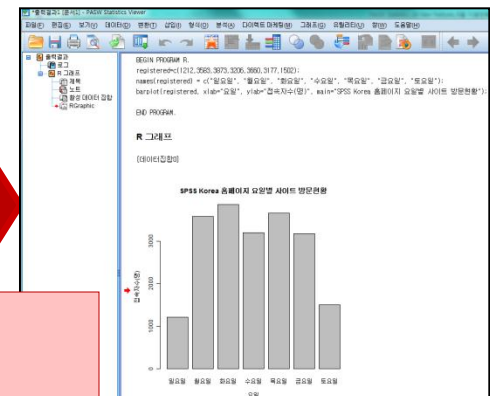
```
*명령문1 - PASW Statistics Syntax Editor
파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래프(G) 유틸리티(U) 실행(R) 도구(S) 창(W) 도움말(H)

BEGIN PROGRAM
registered=c(1212,3583,3873,3206,3660,3177,1502);
END PROGRAM.

2 BEGIN PROGRAM R.
3 registered=c(1212,3583,3873,3206,3660,3177,1502);
4 names(registered) = c("일요일", "월요일", "화요일", "수요일", "목요일", "금요일", "토요일");
5 barplot(registered, xlab="요일", ylab="접속자수(명)", main="SPSS Korea 홈페이지 요일별 사이트 방문현황");
6
7 END PROGRAM.
```

SPSS 명령어 창에서
BEGIN PROGRAM R.

END PROGRAM 사이에 동일한
R 명령어를 입력하면 동일한 수행



> IBM SPSS Stat. with R(2)

간단 시연

The image shows the IBM SPSS Statistics Syntax Editor window with the following code:

```

1
2 BEGIN PROGRAM R.
3 registered=c(1212,3583,3873,3206,3660,3177,1502);
4 names(registered) = c("일요일", "월요일", "화요일", "수요일", "목요일", "금요일", "토요일");
5 barplot(registered, xlab="요일", ylab="접속자수(명)", main="SPSS Korea 홈페이지 요일별 사이트 방문현황");
6 END PROGRAM.
  
```

A red box highlights lines 3 through 5 of the code in the SPSS editor.

The RGui window is also open, showing the R Console with the following output:

```

R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]
  
```

A red box highlights the R console output from the `> registered=c(1212,3583,3873,3206,3660,3177,1502);` line onwards.

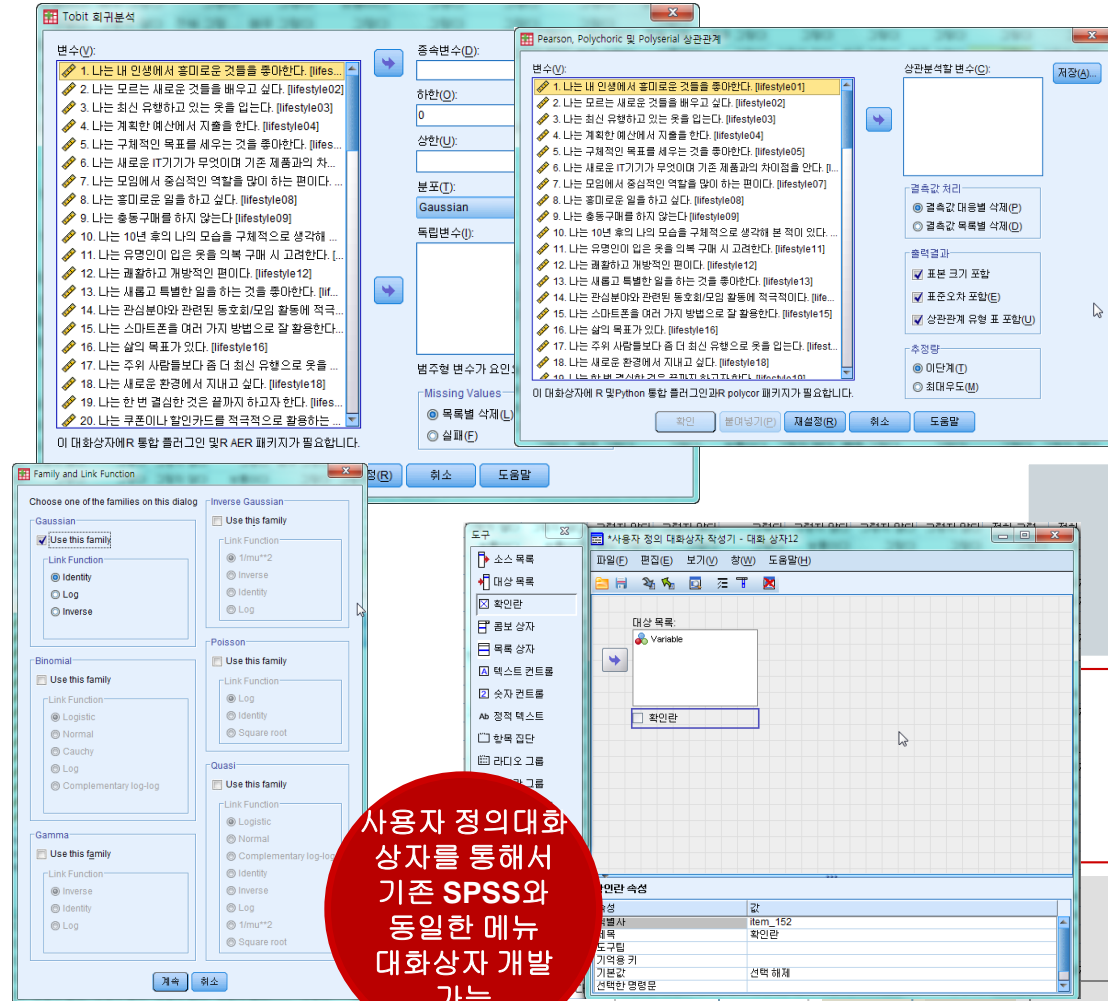
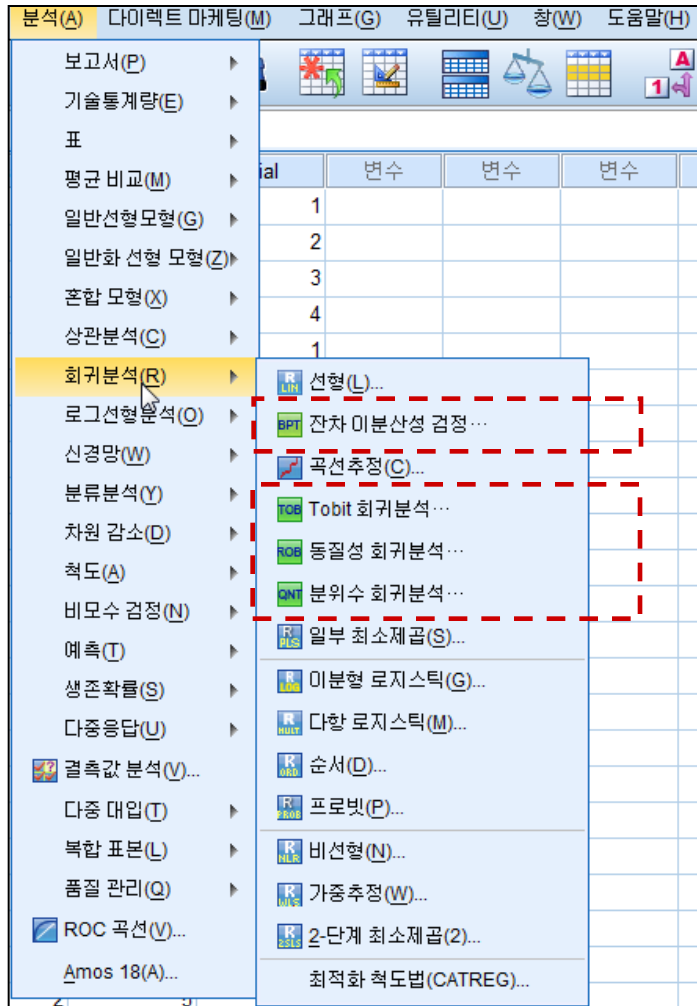
SPSS Syntax 창에서, R 명령어 작업을 수행한다.

R과 동일한 명령어를 SPSS 내에서 사용하면 된다.

> IBM SPSS Stat. with R(3)

SPSS

SPSS-R Extension 대화 상자를 이용하여, R을 좀 더 편리하게 사용!!!



> R-Extension을 이용한 추가 SPSS 분석 모듈

R-Extension 분석 모듈의 종류(1)

R-Extension	잔차 이분산성 검정 (Residual Heteroscedasticity Test)	<ul style="list-style-type: none"> *일명 Breusch-Pagan 검정이라고 함. *회귀분석에서 (잔차의) 이분산성이 존재하는지를 검정하는 방법 → 회귀분석에서는 오차항의 등분산성을 가정함.
	Tobit 회귀분석	<ul style="list-style-type: none"> *종속변수에 상한 또는 하한이 있어서, 데이터가 종도 절단이 될 수 있는 상황에 활용되는 회귀분석 *중단 절단 회귀모형이라고 하기도 함.
	동질성(Robust)회귀분석	<ul style="list-style-type: none"> *회귀분석의 단점인 이상/특이치에 민감할 수 있는, 특성을 보완한 회귀분석 *동질성이라는 해석보다는 우리말로 강건 회귀분석이 더 옳은 표현
	분위수 회귀분석	<ul style="list-style-type: none"> *평균 중심의 추정인 일반 회귀분석과는 달리 중위수 또는 25%, 75%, 5% 등 지정된 분위수 중심으로 추정을 하는 회귀분석 *다양한 비즈니스적인 관점에 따라서 적용이 가능한 회귀분석
	이질성 상관관계 (Heterogenous Correlation)	<ul style="list-style-type: none"> *Polycholic 상관 / Polyserial 상관 등이 있음 *순서 범주형 이변량 자료에 적용되는 상관, 범주형과 연속형간의 상관분석
	상자도표 (Box Plot using R)	<ul style="list-style-type: none"> *데이터 탐색 과정 중 이상치와 사분위수 그리고 평균 등의 파악을 할 수 있는 상자도표 *기존 SPSS에서도 제공

> R-Extension을 이용한 추가 SPSS 분석 모듈

R-Extension 분석 모듈의 종류(2)

Plus pack	Medical Analysis- Bartlett's 등분산 검정 모듈	<p>*의학 및 특히 검정하고자 하는 집단들이 정규분포를 따를 때 활용되는 등분산 검정 모듈</p> <p>*기존 SPSS Statistics에서는 Leven's 등분산 검정 방법만 제공</p>
	Medical Analysis- Poly-K test	<p>*코크란 아미티지 추세검정의 변형된 검정 방법으로, 기존 코크란 아미티지 추세검정에서 노출시간까지 고려한 검정을 통해, 좀 더 정확한 임상 결과에 대한 분석</p>
	Medical Analysis- Page's trend test	<p>*n명의 피험자에게 k개 종류의 순서적 처치가 적용된 실험 자료에서, 순서적 처치들의 간의 차이, 즉, 추세의 유의성을 검정하는 분석 기법</p>



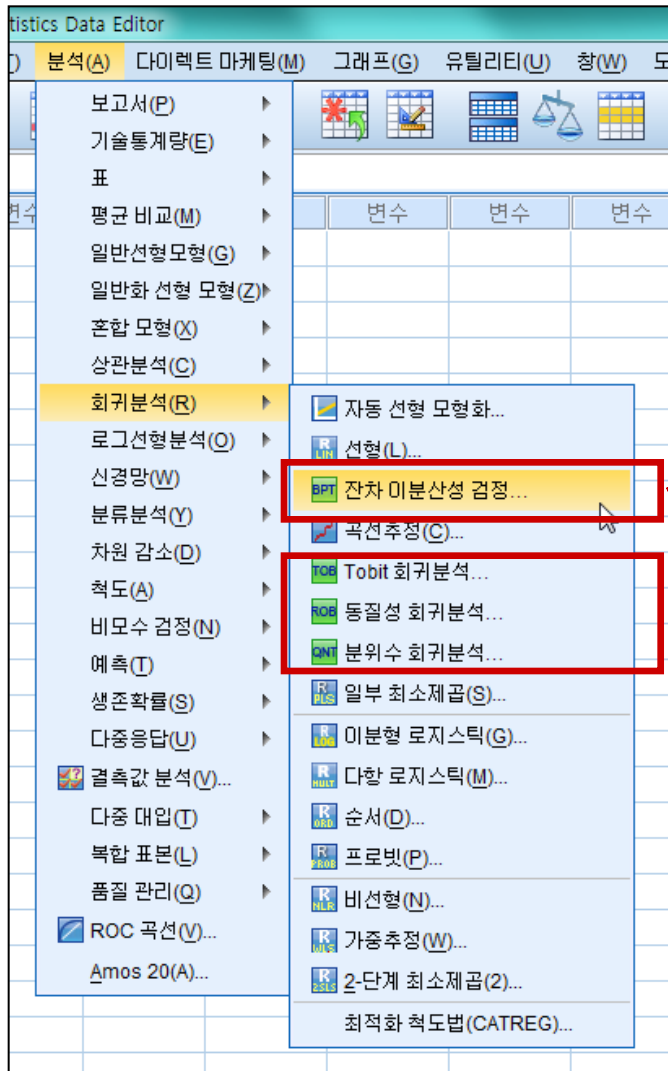
**SPSS에 없는 특수한 통계분석 기법을 R을 이용하여
SPSS와 동일하게 분석이 가능!!!**

Win-Win!!!

**고객은 편리하게 데이터 분석 수행. SPSS 측은
제품의 고객만족도 및 수익 향상!!!**

> SPSS의 R 확장 모듈 예제(1)

기존 SPSS의 분석 방법과 동일한 구조와 Interface를 가진다.



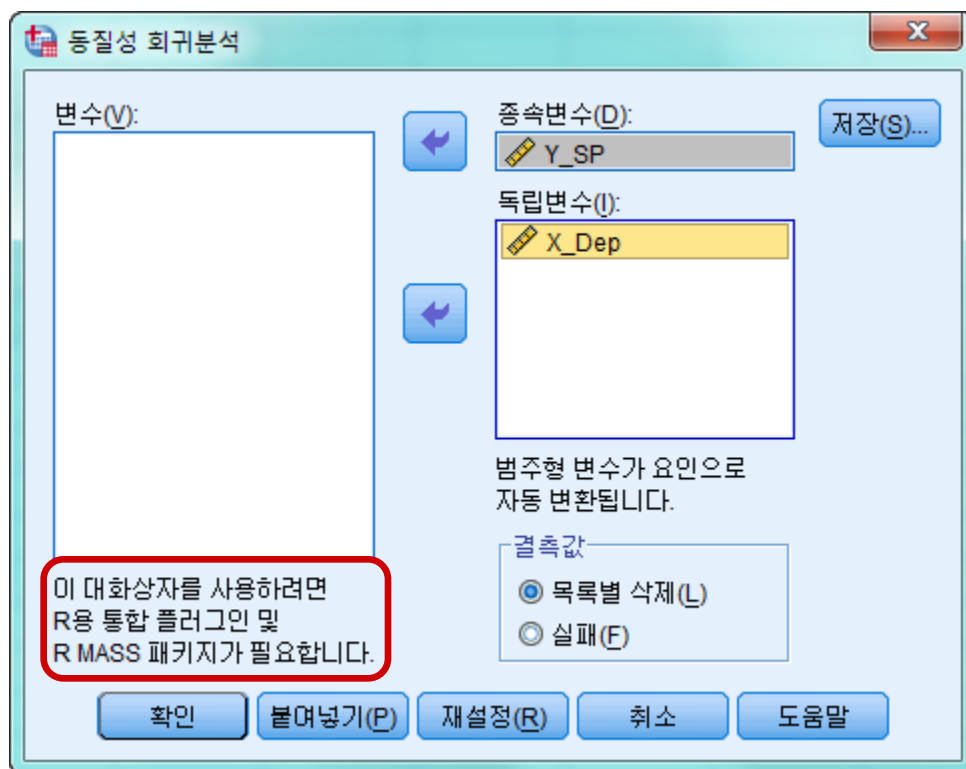
기존 **SPSS** 분석과 동일한 메뉴 방식 분석 특성에 따라 기존 메뉴 구성과 혼합하여 구성이 가능

기존 **SPSS** 분석에서 제공하지 않은 토빗회귀, 로버스트 회귀 등이 들어가져 있음.

> SPSS의 R 확장 모듈 예제(2)

로버스트(또는 동질성)회귀

일반적인 회귀분석에서 최소제곱법을 통한 추정값을 산정하는 경우 이상치 및 특이값에 따른 영향을 많이 받을 수 있는 구조. 이런 경우 이상치 등에 덜 민감하고, 대다수의 값에 근접할 수 있도록 안정적인 구조의 회귀분석을 의미.



→ 동질성 회귀분석

[데이터집합5] J:₩국문자료₩기타자료₩08.PTW2011

계수

	값	표준 오차	t 값
(절편)	68.515	4.647	14.744
X_Dep	.877	.209	4.205

```
rlm(formula = Y_SP ~ X_Dep, data = dta, na.action = na.exclude, method = "MM", model = FALSE)
잔차 표준오차: 4.91761
자유도: 19
```

동일한 형태의 대화상자/
그리고 기존 **SPSS**의 결과와
동일한 형태의 **Output** 출력

> SPSS의 R 확장 모듈 예제(3)

토빗(Tobit) 회귀

토빗 회귀분석의 경우 종도절단 회귀모형(**censored regression model**)이라고도 하며, 특정한 절단점을 기준으로 종속변수 y 가 절단점 기준을 벗어난 경우 이를 절단점으로 대체하고 회귀분석을 하는 것임. 단, 절단점으로 대체한 경우에도 x (설명변수) 측정되어질 수 있으므로, 좀 더 많은 설명변수를 이용하여 정확하고 효율적인 회귀분석을 하는 것이 특징임.

Tobit 회귀분석

변수(V):
remark

종속변수(D):
pur_cnt

저장(S)...

하한(L):
0

상한(U):

분포(T):
Gaussian

독립변수(I):
age

범주형 변수가 요인으로 자동 변환됩니다.

결측값:
☒ 목록별 삭제(L)
☐ 실패(F)

이 대화상자를 사용하려면 R용 통합 플러그인 및 R AER 패키지가 필요합니다.

확인 불여넣기(P) 재설정(R) 취소 도움말

→ Tobit 회귀분석

[데이터집합2] J:\₩국문자료₩기타자료₩08_PTW20111028_R_User_Conference₩tobit_

계수

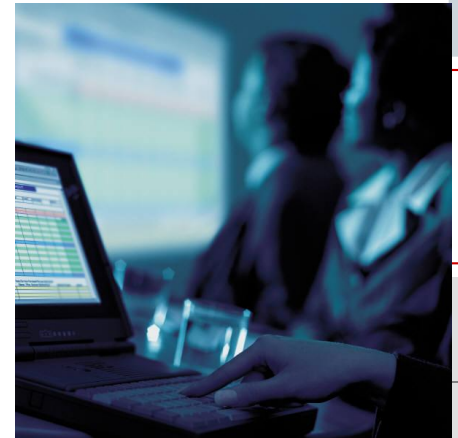
	계수	표준 오차	z 값	유의확률
(절편)	-21.447	8.590	-2.497	.013
age	.618	.175	3.524	.000
로그(척도)	1.706	.197	8.666	.000

하한: 0, 상한: 없음
 tobit(formula = pur_cnt ~ age, left = 0, right = Inf, dist = "gaussian",
 data = dta, na.action = na.exclude)
 척도: 5.5077
 잔차 자유도: 17
 로그 우도: -56.983 자유도: 3
 Wald 통계량: 12.421 자유도: 1

> SPSS의 R 확장 모듈 예제(4)

SPSS-R 확장 모듈 시연

앞서 ppt로 소개한 것의 실제 SPSS에서의 구동 모습을 보시겠습니다.



> SPSS의 R 확장 모듈의 반응

어느 사회과학 계열의 교수님의 말씀

“평소에 **SPSS**를 이용하여, 회귀분석이나, 각종 **TEST**를 수행을 하여, 논문 및 연구보고서를 만들었는데,...

새로운 연구의 경우 **Tobit** 회귀를 써야 했어요... 그러나 **SPSS**에서 분석 기능이 지원이 되지 않아, 다른 통계 패키지를 배우고, 습득하는데, 고생이 많았습니다. **SPSS**쓰다 다른 패키지 갔다가... 헛갈리기도 하고...

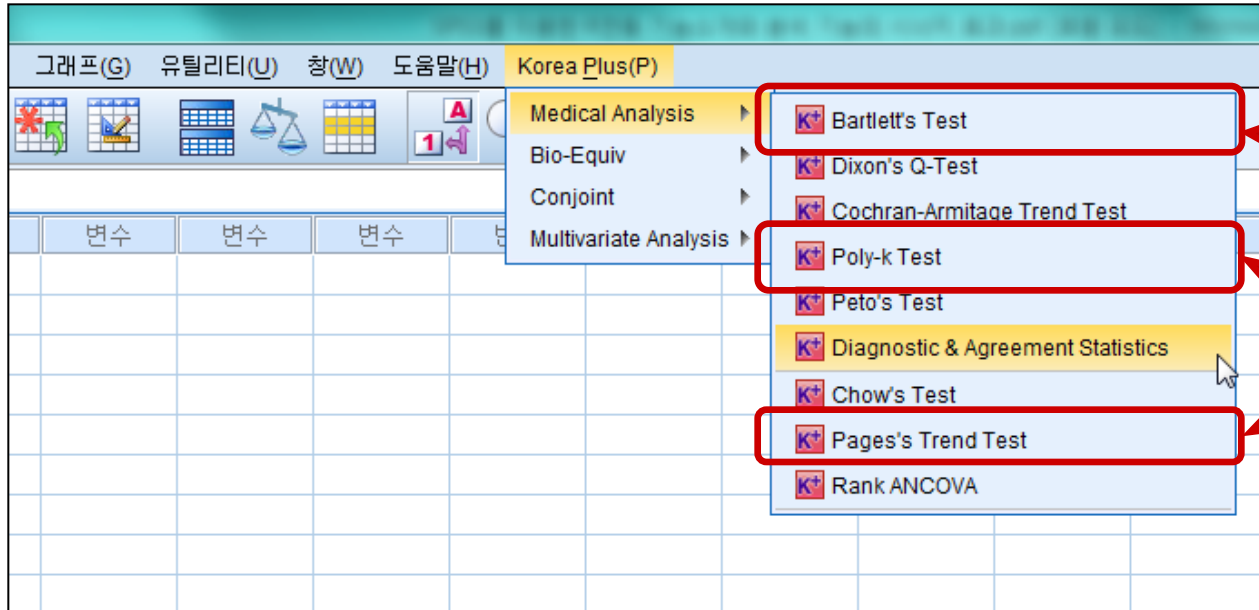
그러나 이번 확장 모듈로 간단히 해결이 되었습니다!!!”



겉보기엔 단순히 **SPSS User**들에게, 편리한 대화상자 하나 추가 되었지만, 실제로는 **SPSS** 전체 만족도를 높여주는 계기가 됨!!!

> SPSS의 R 확장 모듈의 자체 상품화

SPSS와 R의 확장을 통해 새로운 상품을 개발할 수도 있다!!!



총 9가지 모듈 중

3가지가 R의 모듈을
이용한 개발!!!

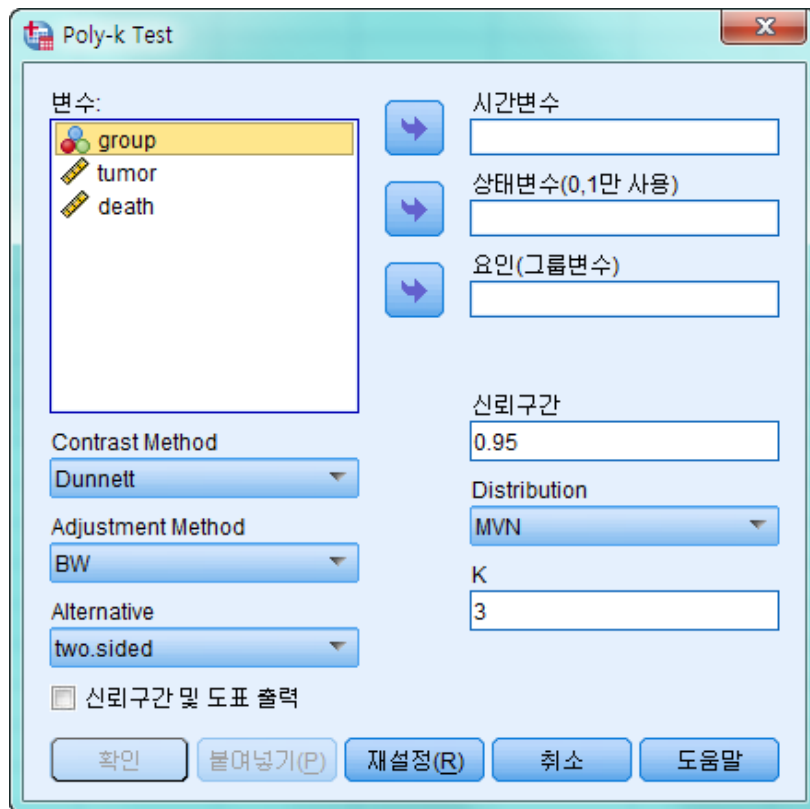
SPSS Korea Plus Medical Analysis

- 의학, 약학, 생명과학, 화학 등의 분야는 통계분석 기법이 매우 특이함 (일반적으로 잘 사용이 안 되는 분석이 많음)
- **SPSS**의 전체적인 매출 신장을 위해서, 기존 **SPSS**에서는 존재하지 않고 의학분야에서 많이 사용되는 **Medical Analysis Pack**을 개발
- 전체 총 9개의 분석 모듈 중 3개가 R의 알고리즘을 이용하여 개발
- R 2.10.1 버전을 이용(**SPSS 19**버전 기준)

> SPSS Korea Medical Plus(1)

Poly-k test

Korea Plus Module의 하나로, 추세가 있는 수준과 집단 간의 유의성 검정에 있어서, 노출된 관측 시간에 따라서 유의성을 파악하기 위한 의학 및 생명과학의 검정 방법임.
Cochran-Armitage test의 변형으로 알려져 있다.



→ Poly-k-adjustment

[DATA.1]

Sample Estimates

Sample estimate, using ploy-3-adjustment

	0	1	2	3
x	1.000	5.000	8.000	9.000
n	50.000	50.000	50.000	50.000
adjusted n	41.405	32.698	38.744	40.311
adjusted estimate	.024	.153	.206	.223

Contrast matrix of Multiple Comparisons

Multiple Comparisons of Means: Dunnett Contrasts

	0	1	2	3
1 - 0	-1.000	1.000	.000	.000
2 - 0	-1.000	.000	1.000	.000
3 - 0	-1.000	.000	.000	1.000

Union-Intersection test

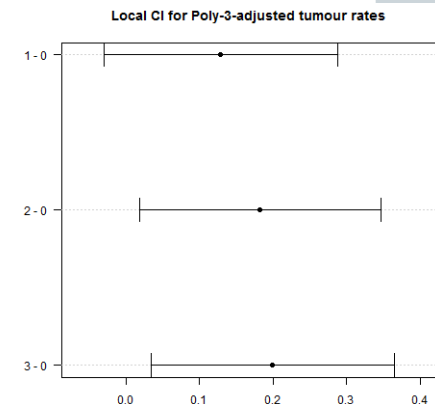
Union-Intersection test using BW variance estimator

	estimate	teststat	p.val.adj
1 - 0	.129	1.936	.149
2 - 0	.182	2.649	.024
3 - 0	.199	2.869	.012

Confidence Intervals

Simultaneous 95 percent confidence intervals using BW variance estimators

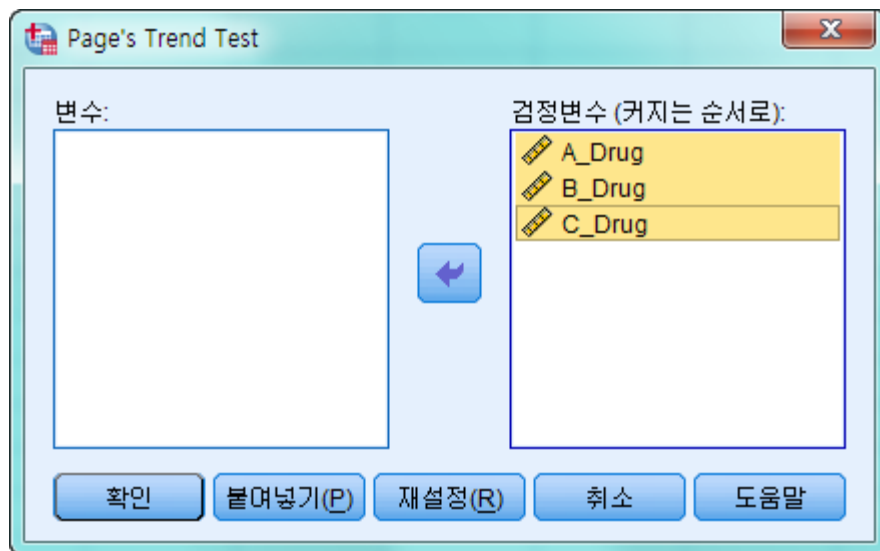
	estimate	lower	upper
1 - 0	.129	-.030	.287
2 - 0	.182	.018	.347
3 - 0	.199	.034	.365



> SPSS Korea Medical Plus(2)

Page's Trend Test

Korea Plus Module의 하나로, n 개 종류의 처치(아래의 예에서는 3가지 약의 종류)를 순서적으로 처치(약 순서)했을 때의 증상(추세)의 변화량에 대한 유의성을 살펴보는 검정방법으로, 특히 약학 임상 실험 및 각종 물리치료 등의 순서에 따른 병의 호전도 파악에 좋은 연구 방법 중 하나



기술통계

[데이터집합5] J:\₩국문자료₩기타자료₩08_PT₩20111028_R_User_Co

기술통계량

	N	최소값	최대값	평균	표준편차
A_Drug	8	65.00	79.00	71.6250	5.31675
B_Drug	8	66.00	78.00	72.5000	4.65986
C_Drug	8	65.00	82.00	75.0000	5.60612
유효수 (목록별)	8				

→ Page's Trend Test

[데이터집합5] J:\₩국문자료₩기타자료₩08_PT₩20111028_R_User_Co

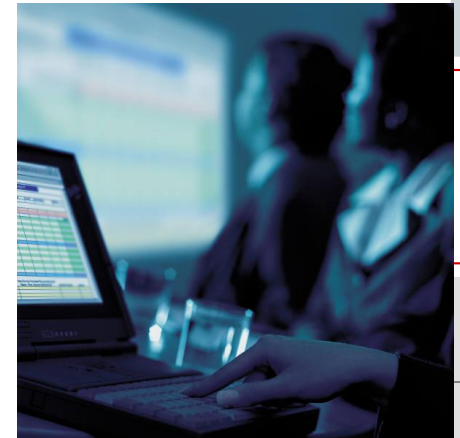
검정결과

	L	유의확률
Page's trend test	104.500	<=.05

> SPSS Korea Medical Plus(3)

SPSS Korea Medical Plus

앞서 ppt로 소개한 것의 실제 SPSS에서의 구동 모습을 보시겠습니다.



> SPSS Korea Medical Plus(4)

개발 기간과 실제 성과

개발기간/인원

*1차 6개모듈: 2주

*2차 3개모듈: 1주

*개발인원: 컨설턴트 2명,
기술개발자: 1.5명

→ 적은 수의 인원으로
간단히 개발 가능



*자사 7개의 패키지 모듈
중 3번째로 가장 많이
판매됨.(SPSS Medical
Pack)

*병원/의료기관 판매
전년 대비 약 17% 증가

> SPSS Korea Medical Plus(5)

SPSS

대만으로의 확대 보급

Welcome to SPSS TAIWAN | 加入我的最愛 | 客服專區 | 回首頁 | Facebook

XISHU 台灣析數資訊股份有限公司
XISHU Software, TAIWAN

Business Partner IBM

產品介紹 優惠方案 產品購買 教育訓練 研討會 關於台灣析數 臺銀採購

> 首頁 / 產品介紹 / 台灣析數統計分析加值模組 / SPSS醫學分析模組

> SPSS醫學分析模組

由台灣析數資訊獨家開發的醫學分析模組，是全國唯一可快速執行臨床試驗的進階分析模組。我們依據研究人員在使用數據分析軟體的使用狀況後，為貼近研究人員的需求，整合其功能與方法，讓您在分析更多數據時，能更快取得研究結論。

More functionality, improved graphics and an easier to use interface
IBM SPSS Statistics 19

功能特色 分析方法詳述 安裝需求

Bartlett's Test

Variable(s):

Dependent Variable: GRP

Factor: VAL

OK Paste Reset Cancel Help

Bartlett Test
假設為常態分佈，可使用Bartlett Test執行變異數同質性檢定並確保ANOVA等分析結果的可靠性。

Dixon's Q-Test

Variable(s):

Target: 變數 [VAR]

OK Paste Reset Cancel Help

Dixon's Q-Test
在極小樣本情況下，可以輕鬆執行離群值檢定，並支援依照不同信心水準90%、95%、99%，可以一次掌握是否有離群值。

Cochran-Armitage Trend Test

Variable(s):

Rows (k categories): 劑量 [dose]

Columns (2 categories): 反應 [response]

Weight cases: 權重 [weight]

Poly-k Test

Variable(s):

Time Variable: 時間 [time]

Status (0 1): 狀態 [status]

Grouping Variable: 群組 [group]

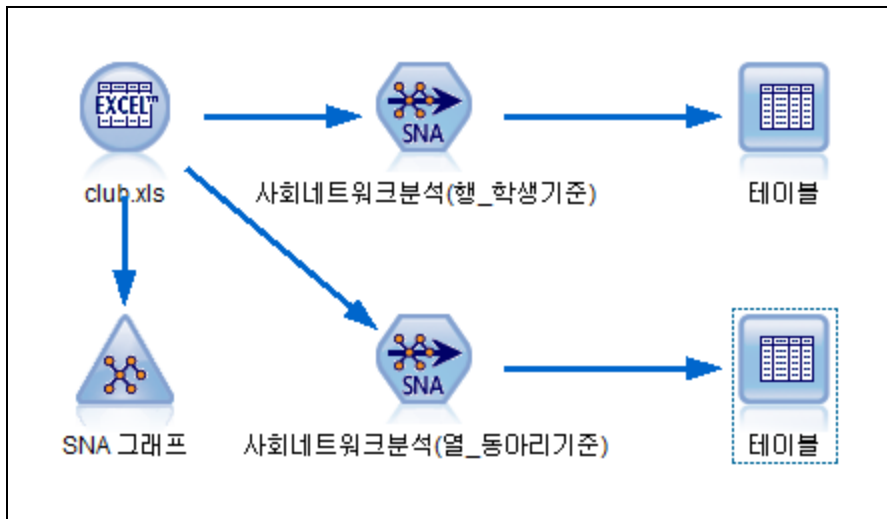
Contrast Method: Contrast

Confidence Level: 0.95

***Medical Analysis는**
현재 대만에서도 병원 및
관련 기관 판매에 큰 역할
을 하고 있음

> SPSS Korea SPSS Modeler SNA 모듈

SPSS Modeler에 R의 Social Networks Analysis 기능을 탑재하여 새로운 기능을 추가



테이블 (5개 필드, 15개 레코드)

	CASE_ID	DEG_VAL	BET_VAL	CLO_VAL	EIG_VAL
1	김수박	0.714	0.075	0.857	0.331
2	김사과	0.429	0.037	0.714	0.166
3	김말기	0.643	0.031	0.821	0.320
4	김자몽	0.500	0.028	0.750	0.228
5	김포도	0.714	0.075	0.857	0.322
6	김향외	0.643	0.031	0.821	0.320
7	김살구	0.714	0.054	0.857	0.339
8	김체리	0.500	0.014	0.750	0.262
9	김레몬	0.500	0.028	0.750	0.228
10	김메론	0.357	0.019	0.679	0.170
11	김금귤	0.429	0.019	0.714	0.182
12	김키위	0.429	0.026	0.714	0.167
13	김리치	0.571	0.032	0.786	0.279
14	김매실	0.500	0.021	0.750	0.242
15	김거봉	0.500	0.037	0.750	0.203

확인

사회네트워크분석(행_학생기준)

미리보기(P)

설정 | 그래프 | 지정 | 주석

행렬 유형: ☐ 인접행렬(n x n) ☒ 2부 네트워크(n x m)

2부 네트워크 분석기준: ☒ 행 노드 ☐ 열 노드

방향 유무: ☐ 방향 없음 ☒ 방향 있음

가중 네트워크: ☐

분석 프로그램: ☒ R sna ☐ Pajek

시점(From):

종점(To):

가중치:

분석 지표 선택

연결정도 중심성: ☒ On ☐ Off

중개 중심성: ☒ On ☐ Off

근접 중심성: ☒ On ☐ Off

고유벡터 중심성: ☒ On ☐ Off

부그룹 번호: ☐ On ☒ Off

확인 취소 적용(A) 재설정(R)

> SPSS와 R의 동시 분석 작업(1)

통계(데이터)분석자들은 SPSS와 R을 동시에 한 프로그램에서 이용하며, 서로의 장점을 취할 수 있다.

GLMM_예제.sav [데이터집합1] - IBM SPSS Statistics Data Editor

파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다이렉트 마케팅(M) 그래픽(G) 유틸리티(U) 창(W) 도움말(H) Korea Plus(P)

새 파일(N) ▶
열기(O) ▶
데이터베이스 열기(B) ▶
테스트 데이터 열기(O)...
닫기(C) Ctrl+F4
저장(S) Ctrl+S
다른 이름으로 저장(A)...
모든 데이터 저장(L)
데이터베이스로 내보내기(T)...
파일을 열기 전용으로 표시(K)
데이터 파일 이름변경(M)...
데이터 파일 정보 표시(I) ▶
데이터 캐쉬(H)...
변수 정보 수집
프로세서 중단(E) Ctrl+Period
서버 전환(W)...
리포지토리(Y) ▶
인쇄 미리보기(V)
인쇄(P)... Ctrl+P
최근에 사용한 데이터(Y) ▶
최근에 사용한 파일(F) ▶
종료(X)

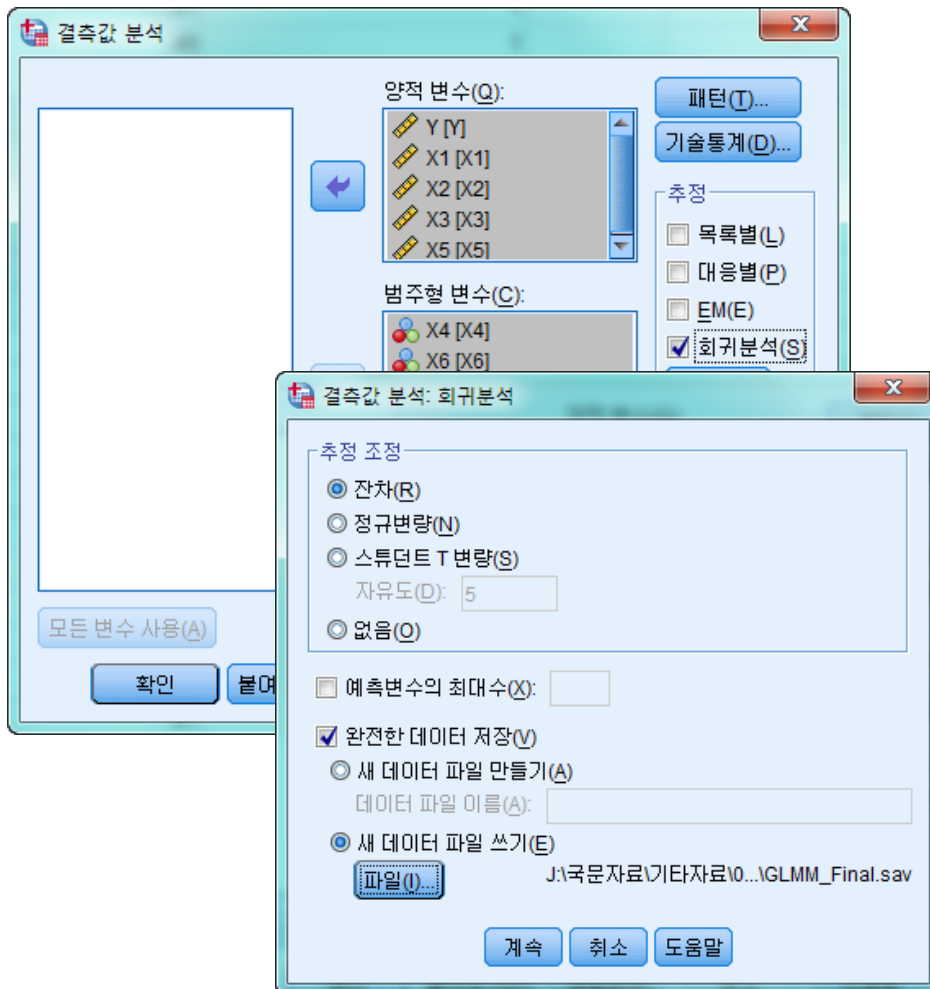
데이터(A)
명령문(S)...
출력결과(O)...
스크립트(C)...

		X3	X4	X5	X6	
22		2.50	3.15	2.00	.53	
23	.47	1.30	1.43	1.00	1.09	
24	1.25	1.56	1.00	2.31	.08	
25	-.69					
26	.88	3.47	2.80	4.21	2.00	-.30
27	1.41	2.11	1.87	3.15	1.00	.36
28	2.29	3.08	3.13	3.70	1.00	-.30
29	2.17	2.31	1.89	3.02	2.00	.41
30	1.03	3.00	2.82	4.15	1.00	.49
31	-.69	3.02	2.86	4.01	2.00	.08
32	3.00	4.01	3.75	5.66	2.00	-.30
33	-.69	2.88	2.69	3.06	2.00	.18
34	-.69	3.46	3.43	5.65	2.00	.74
35	.26	4.22	3.76	4.77	2.00	.46
36	1.03	3.57	3.18	4.46	1.00	.48
37	-.69	4.30	3.84	4.80	1.00	.93
		4.12	3.67	4.93	2.00	-.30
		3.86	3.65	4.94	2.00	-.30
		3.70	3.15	4.70	2.00	.18
		1.51	1.43	2.00	2.00	1.03
		3.01	2.67	3.16	2.00	.20
		-1.10	-1.07	.45	1.00	.54
		4.97	4.45	4.69	2.00	-.30
		4.11	3.45	5.07	2.00	.38
		-.68	-1.05	2.54	1.00	.61
		.76	.71	1.96	1.00	1.00
		1.19	.93	1.41	2.00	.94
		2.44	2.10	2.28	1.00	.45
		4.38	3.71	4.73	1.00	-.30
		4.18	3.83	5.03	1.00	1.30
		3.60	3.46	4.26	2.00	-.30
		3.95	3.87	5.14	2.00	-.30
		3.54	3.44	4.64	2.00	.11
		2.86	2.67	3.08	1.00	.45
		5.08	4.96	4.80	2.00	-.30

Spread Sheet 구조와
메뉴 방식의 SPSS를
이용하여 데이터
열기
→ 데이터 구조
한 눈에...

> SPSS와 R의 동시 분석 작업(2)

각종 결측치 보강 등에 대해서, 사용자가 한글로 된 메뉴를 이용하여 편리하게 핸들링



GLMM_Final.sav [데이터집합3] - IBM SPSS Statistics Data Editor

	Y	X1	X2	X3
1	1.22	2.66	2.50	3.15
2	2.50	1.29	1.30	1.43
3	.18	1.56	1.00	2.31
4	-.69	1.51	.93	4.77
5	-.69	3.47	2.80	4.21
6	.83	2.11	1.87	3.15
7	-.69	3.08	3.13	3.70
8	.96	2.31	1.89	3.02
9	1.13	3.00	2.82	4.15
10	.18	3.02	2.86	4.01
11	-.69	4.01	3.75	5.66
12	.41	2.88	2.69	3.06
13	1.70	3.46	3.43	5.65
14	1.06	4.22	3.76	4.77
15	1.10	3.57	3.18	4.46
16	2.15	4.30	3.84	4.80
17	-.69	4.12	3.67	4.93
18	-.69	4.12	1.43	2.31
19	-.69	3.86	3.65	4.94
20	.41	3.70	3.15	4.70
21	2.37	1.51	1.43	2.00
22	-.69	1.56	2.10	3.02
23	.47	3.01	2.67	3.16
24	1.25	-1.10	-1.07	.45
25	-.69	4.97	4.45	4.69
26	.88	4.11	3.45	5.07
27	1.41	-.68	-1.05	2.54
28	2.29	.76	.71	1.96

> SPSS와 R의 동시 분석 작업(3)

만약 SPSS에 없는 분석 기법인 경우, R에서 해당 분석 기법을 이용하여 분석 (GLMM → SPSS 19에 신 기능이지만, 여기서는 예제 상...)

```

13
14 *R을 이용한 Mixed Model.
15 BEGIN PROGRAM R.
16 casedata <- spssdata.GetDataFromSPSS()
17 library(lme4)
18 result<-fixef(lmer (Y ~ X1 + X2 + (1|X3) , data = casedata))
19 result1<-data.frame(result)
20 library(foreign)
21 write.foreign(df=result1, datafile="D:\\test.txt", codefile="d:\\test.sps", package="SPSS")
22 END PROGRAM.
23
24 *GLMM의 계수값 읽어오기.
25 DATA LIST FILE= "j:\\test.txt" free (" ,")
26 / result .
27
28 VARIABLE LABELS
29 result "result"
30 .
31
32 EXECUTE.
  
```

이 부분이 spss 데이터 가져오기 명령어. 나머지는 R 명령어와 동일

요구된 패키지 Matrix를 로드중입니다
요구된 패키지 lattice를 로드중입니다

다음의 패키지를 추가합니다: 'lme4'

The following object(s) are masked from package:'stats' :

AIC
Linear mixed model fit by REML
Formula: Y ~ X1 + X2 + (1 | X3)
Data: casedata
AIC BIC logLik deviance REMLdev
159.5 168.8 -74.73 145.2 149.5
Random effects:
Groups Name Variance Std.Dev.
X3 (Intercept) 0.17431 0.41751
Residual 1.11451 1.05570
Number of obs: 48, groups: X3, 44

Fixed effects:
Estimate Std. Error t value
(Intercept) 1.5391 0.3758 4.096
X1 -0.5303 0.3952 -1.342
X2 0.2733 0.4154 0.658

Correlation of Fixed Effects:
(Intr) X1
X1 -0.322
X2 0.059 -0.955

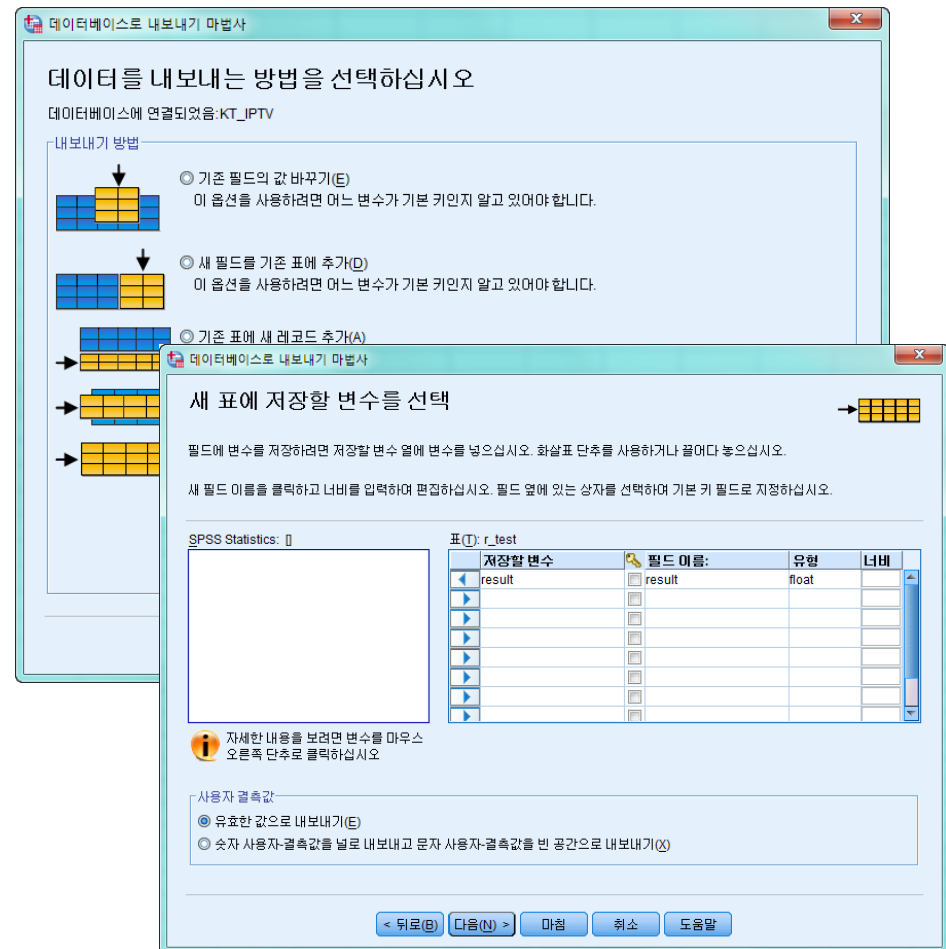
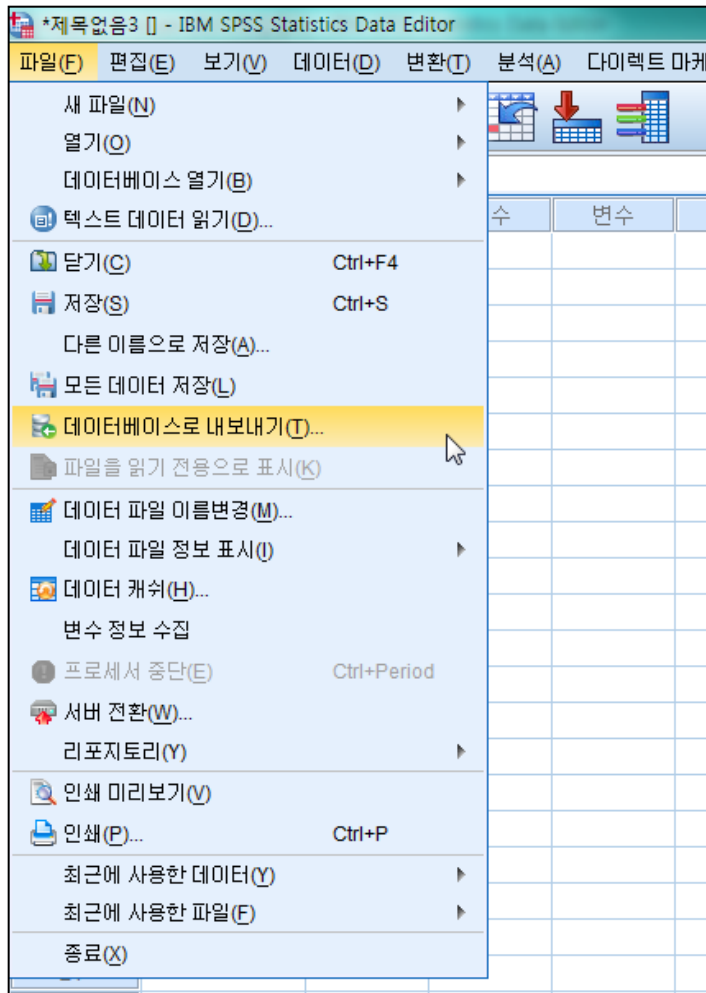
*제목없음3 - IBM SPSS Statistics Data Editor

	result	변수	변수
1	1.54		
2	-.53		
3	.27		
4			

R을 이용하여 분석하고,
그 결과를 다시 SPSS
데이터 파일로 저장

> SPSS와 R의 동시 분석 작업(4)

SPSS의 편리한 GUI DB 저장 기능을 이용하여, 결과 값을 Database Table로 저장



> SPSS와 R의 동시 분석 작업(5)

이 모든 과정을 하나의 Syntax로 만들어, 배치(batch)파일화 시킬 수 있고, 쉽게 적용

```

*GLMM_Syntax_수정.sps - IBM SPSS Statistics Syntax Editor
파일(F) 편집(E) 보기(V) 데이터(D) 변환(T) 분석(A) 다차원 마케팅(M) 그래프(G) 사용자 정의(C) 유틸리티(U) 실행(R) 도구
*데이터 불러오기.
GET
DATASET NAME
*결과값 분석 및 보완.
MVA
BEGIN PROGRAM
END PROGRAM.
*R을 이용한 Mixed Model...
BEGIN PROGRAM
END PROGRAM.
*GLMM의 계수값 읽어오기.
DATA LIST
VARIABLE LABELS
EXECUTE.
SAVE TRANSLATE
1
2
3 GET
4 FILE=J:\국문자료\기타자료\08_PT20111028_R_User_Conference\GLMM_예제.sav.
5 DATASET NAME 데이터집합2 WINDOW=FRONT.
6
7 *결과값 분석 및 보완.
8 MVA VARIABLES=Y X1 X2 X3 X5 X4 X6
9 /MAXCAT=25
10 /CATEGORICAL=X4 X6
11 /REGRESSION(TOLERANCE=0.001 FLIMIT=4.0 ADDTYPE=RESIDUAL
12 OUTFILE=J:\국문자료\기타자료\08_PT20111028_R_User_Conference\GLMM_Final.sav).
13
14 BEGIN PROGRAM R.
15 casedata <- spssdata.GetDataFromSPSS()
16 library(lme4)
17 lmer(Y ~ X1 + X2 + (1|X3), data = casedata)
18 END PROGRAM.
19
20 *R을 이용한 Mixed Model.
21 BEGIN PROGRAM R.
22 casedata <- spssdata.GetDataFromSPSS()
23 library(lme4)
24 result<-fixef(lmer(Y ~ X1 + X2 + (1|X3), data = casedata))
25 result1<-data.frame(result)
26 library(foreign)
27 write.foreign(df=result1, datafile="D:\test.txt", codefile="d:\test.sps", package="SPSS")
28 END PROGRAM.
29
30 *GLMM의 계수값 읽어오기.
31 DATA LIST FILE= "j:\test.txt" free (" ,")
32 / result .
33
34 VARIABLE LABELS
35 result "result"
36 EXECUTE.
37
38
39 SAVE TRANSLATE /TYPE=ODBC
40 /CONNECT=DSN=KT_IPTV;UID=;Trusted_Connection=Yes;APP=IBM SPSS '+
41 'Statistics;WSID=HOH-HP;DATABASE=TEST
42 /ENCRYPTED
43 /MISSING=IGNORE
44 /SQL='CREATE TABLE [R_TEST]([result] float)'
45 /REPLACE
46 /TABLE='SPSS_TEMP'
47 /KEEP=result
48 /SQL='INSERT INTO [R_TEST]([result]) SELECT [result] FROM [SPSS_TEMP]'
49 /SQL='DROP TABLE [SPSS_TEMP]'
  
```

```

spssb -f D:\SPSS_MINING\Batch_2\Syntax\PRED_EXEC_A2.SPS > D:\SPSS_MINING\Batch_2\Log\PRED_EXEC_A1.txt
spssb -f D:\SPSS_MINING\Batch_2\Syntax\PRED_EXEC_A2.SPS > D:\SPSS_MINING\Batch_2\Log\PRED_EXEC_A2.txt
spssb -f D:\SPSS_MINING\Batch_2\Syntax\PRED_EXEC_A2.SPS > D:\SPSS_MINING\Batch_2\Log\PRED_EXEC_A3.txt
spssb -f D:\SPSS_MINING\Batch_2\Syntax\PRED_EXEC_A2.SPS > D:\SPSS_MINING\Batch_2\Log\PRED_EXEC_A4.txt
spssb -f D:\SPSS_MINING\Batch_2\Syntax\PRED_EXEC_A2.SPS > D:\SPSS_MINING\Batch_2\Log\PRED_EXEC_A5.txt
spssb -f D:\SPSS_MINING\Batch_2\Syntax\PRED_EXEC_A2.SPS > D:\SPSS_MINING\Batch_2\Log\PRED_EXEC_A6.txt
spssb -f D:\SPSS_MINING\Batch_2\Syntax\PRED_EXEC_A2.SPS > D:\SPSS_MINING\Batch_2\Log\PRED_EXEC_A7.txt
spssb -f D:\SPSS_MINING\Batch_2\Syntax\PRED_EXEC_A2.SPS > D:\SPSS_MINING\Batch_2\Log\PRED_EXEC_A8.txt
  
```



SPSS_Batch_A2
MS-DOS 일괄 파일
1KB

일반 XXXX.bat 파일
형태로 만듦

*SPSS의 경우 Syntax 파일을 간단히 저장하고, SPSS Batch로 만드는 과정이 매우 간단함.

*IT의 초보자들도 바로 배치 파일을 만들고, 이를 스케줄러 등에 탑재하여, 시스템화 시킬 수 있음.

*이로써, 통계 모형을 이용한 예측적 모델 시스템 개발이 완료됨.

> SPSS와 R의 동시 분석 작업(6)

SPSS 내에서 R을 동시 분석하는 경우의 시너지(synergy)

1

전체적인 데이터 탐색 및 파악이 용이하여, 순조롭게 분석 진행이 가능함.(SPSS Spread sheet / 메타 데이터 작성 기능 등)

2

간단한 데이터 핸들링 및 데이터 변환/조작 등이 매우 용이 하여, R로 만든 모델 또한 즉각적인 수정 후 결과 도출이 가능함.

3

간략한 분석 및 SPSS가 지원하는 분석에 대해서는 손쉽게 분석이 진행되고, 여러 모델에 대한 동시 평가가 매우 편리하여, 최종 모델 선택 및 Hybrid 모델로의 전환이 용이함.

4

SPSS와 R을 이용하여, 전체적인 흐름 Process를 만들고, 최종적으로 SPSS Batch 시스템을 이용하여, 시스템화가 간단함.

> 맺음말

데이터 분석의 시장 그리고 데이터 분석 인력을 넓히자!!!



Q&A

