SAS 복합표본설계 자료분석 101[[1]](#footnote-1)

**목적과 범위**

우리는 국민건강영양조사(이하 ‘국건영’) 자료를 분석하면서 몇 가지 시행착오를 겪었다. 이 문서는 국건영 자료분석을 시작할 때 필요한 준비적 지식을 요약한다. 문서의 목적은 다음과 같다.

① SAS® OnDemand for Academics 버전(무료 온라인)을 사용하여 연구에 참여할 수 있다.  
② ‘국건영’ 지침서가 권고하는 복합표본설계 요소를 반영하여 자료분석을 수행한다.

국건영 원시자료 분석지침을 따르기 위해 노력했으나 완전한 이해가 부족하여 이 문서와 코드의 오류가 있을 수 있다. 이런 경우 지침과 해석은 국건영 제공 자료를 따르는 것이 원칙이다.

**내용**

[SAS® OnDemand for Academics 조작 기본 3](#_Toc106699102)

[① 등록과 로그인 3](#_Toc106699103)

[② 첫화면 3](#_Toc106699104)

[③ 저장 – 코드 5](#_Toc106699105)

[④ 저장 – 데이터 8](#_Toc106699106)

[⑤ 디렉토리 구조 잡기 10](#_Toc106699107)

[⑥ 프로시저 설명과 예제 – 매뉴얼 참고하기 11](#_Toc106699108)

[사전 지식 요약 - 단순임의표본 자료 분석 12](#_Toc106699109)

[복합표본설계 자료 분석 13](#_Toc106699110)

[① 국민건강영양조사 자료의 특징 13](#_Toc106699111)

[② SAS 복합표본 자료분석 15](#_Toc106699112)

[코드 구조 17](#_Toc106699113)

[① 매크로 17](#_Toc106699114)

[② 주석 17](#_Toc106699115)

[③ prepare-dataset.sas 18](#_Toc106699116)

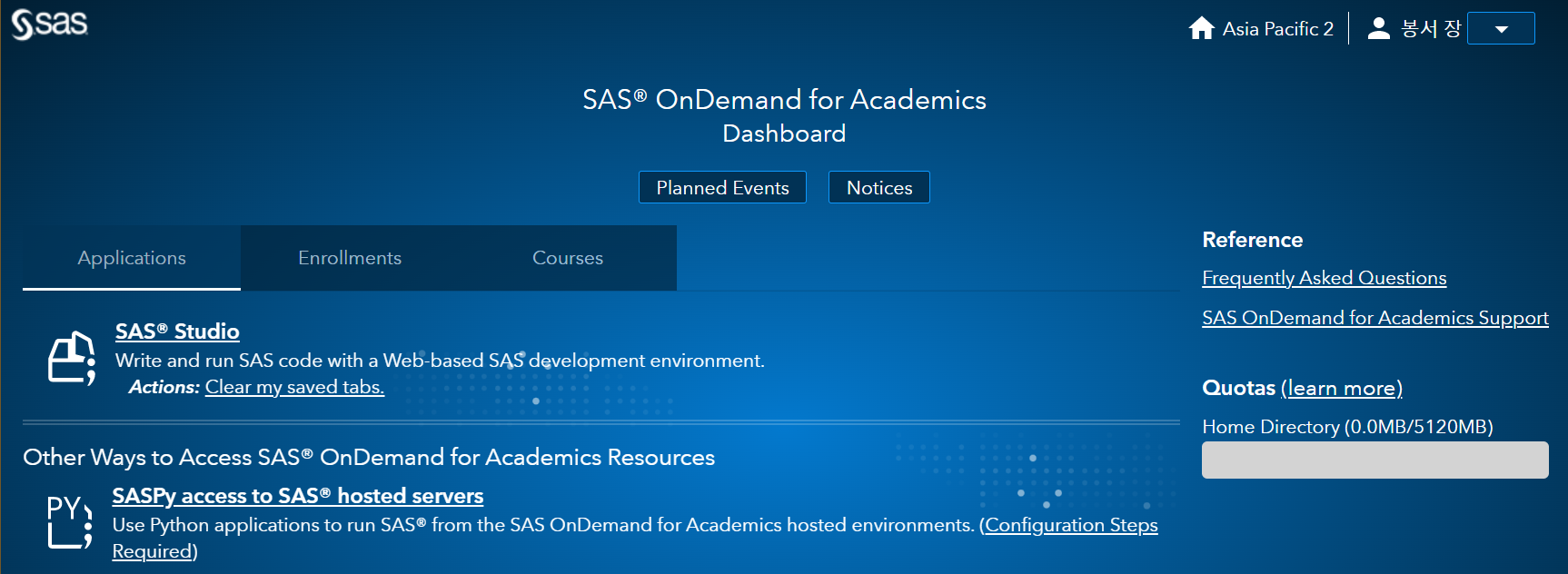
[④ explore-survey.sas 19](#_Toc106699117)

[⑤ analyze-survey.sas 20](#_Toc106699118)

# SAS® OnDemand for Academics 조작 기본

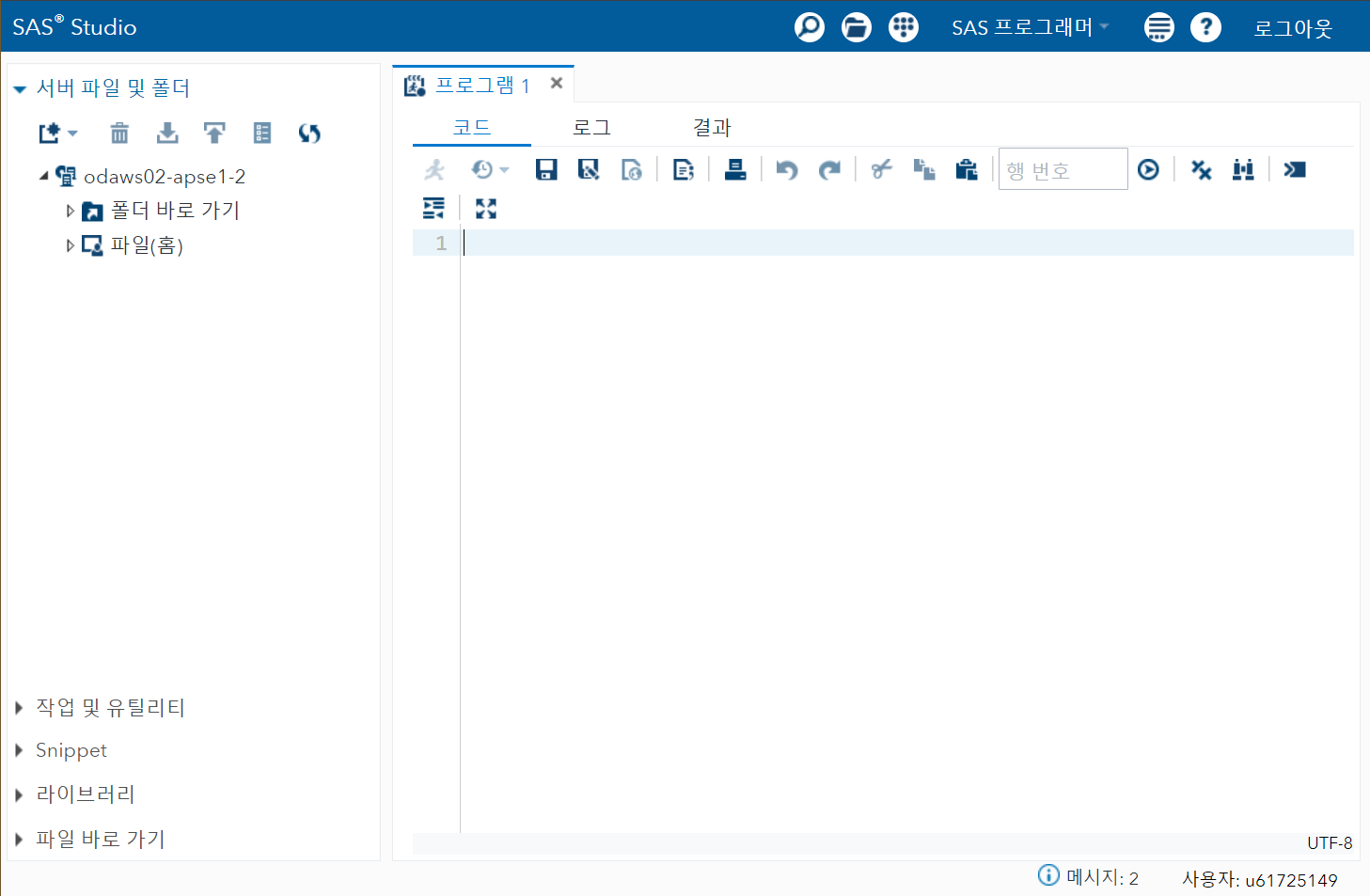
### ① 등록과 로그인

<https://welcome.oda.sas.com/login>에 접속하여 등록한다. 등록한 이메일을 확인하여 활성화, 암호설정을 마치고 로그인한다. 최초로그인 시, 지역을 ‘Asia Pacific’으로 설정한다. 자동으로 로그아웃 되는데, 1~2분 후 다시 로그인하면 다음 화면을 볼 수 있다. SAS® Studio 링크를 클릭한다.



### ② 첫화면

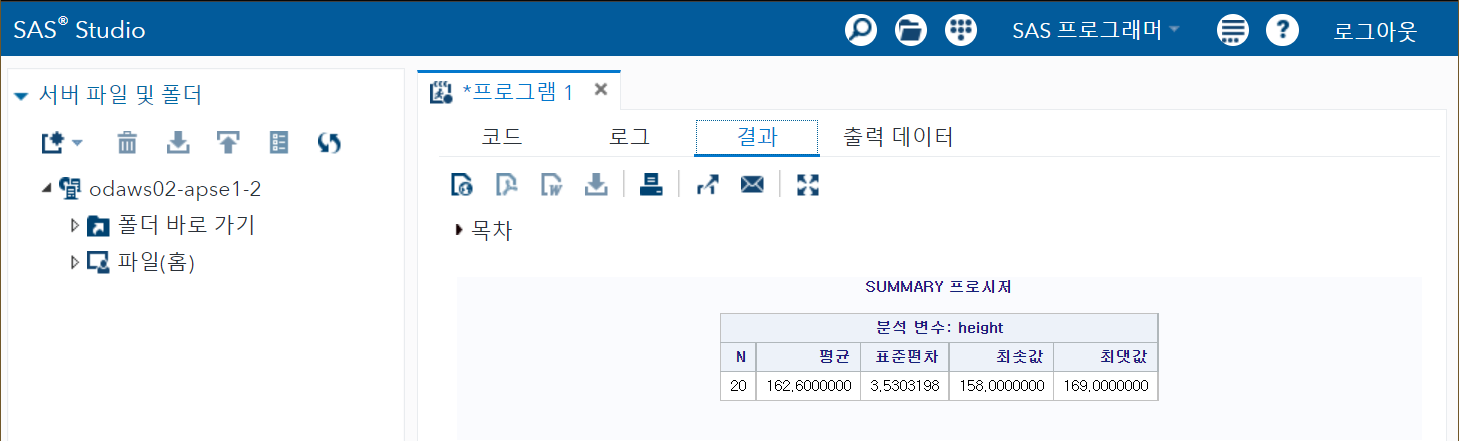
SAS의 첫 화면은 왼쪽에 ‘서버 파일 및 폴더’탭이 열려 있고 오른쪽에 ‘프로그램1’ 탭이 열려있다.



‘프로그램1’ 탭에서 통계분석 프로그램을 작성하고 실행한다. (달리는 사람모양 아이콘)

|  |
| --- |
| DATA example;  INPUT class height @@;  CARDS;  1 169 1 168 1 164 1 165 1 159 1 158 1 158 1 169 1 160 1 162  2 163 2 160 2 160 2 164 2 160 2 161 2 164 2 160 2 161 2 167  ;  RUN;  PROC SUMMARY DATA=example PRINT;  VAR height;  RUN; |

‘프로그램1’ 탭 아래 ‘결과’탭이 추가되었고 SUMMARY 프로시저의 결과를 보여준다.



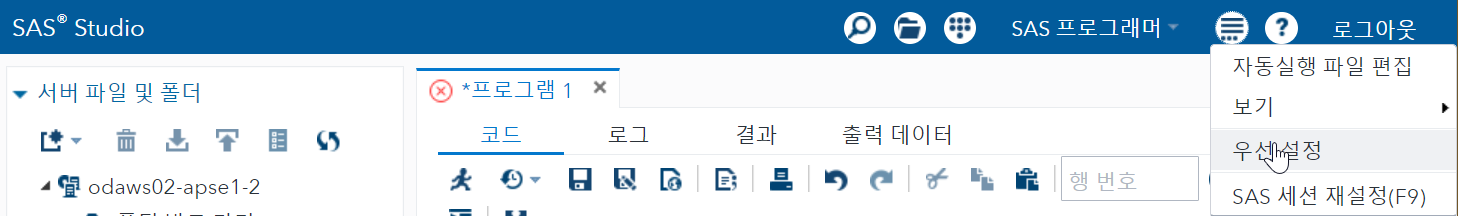
에러가 있다면 ‘프로그램1’탭이 빨간색으로 표시된다. 이 때는 ‘로그’탭에서 에러의 내용을 확인한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

### ③ 저장 – 코드

SAS OnDemand는 일정 시간이 지나면 자동으로 로그아웃 된다. 시간 설정은 왼쪽위의 3단 줄무늬 아이콘을 눌러서 나온 ‘우선설정’ 메뉴에서 할 수 있다.



‘프로그램1’ 탭의 프로그램을 저장하지 않았기 때문(‘프로그램1’ 글자 왼쪽에 ‘\*’표시로 알 수 있음)에 자동 로그아웃이나 임의적인 로그아웃을 하면 내용이 날아간다. 따라서 보존을 원하는 내용은 저장해야 하는데 두가지 방식이 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

◯ ‘서버 파일 및 폴더’에 저장하기 (CTRL-S)

일반적으로 생각하는 파일 저장하기 방법이다. 아이콘이나 CTRL-S를 누르면 ‘서버 파일 및 폴더’에 저장할 수 있다. 위치를 ‘파일(홈) > sasuser.vXX’로 하고, 원하는 파일명을 적고 ‘저장’한다. 왼쪽의 ‘서버 파일 및 폴더’탭에서 저장된 example.sas를 확인할 수 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

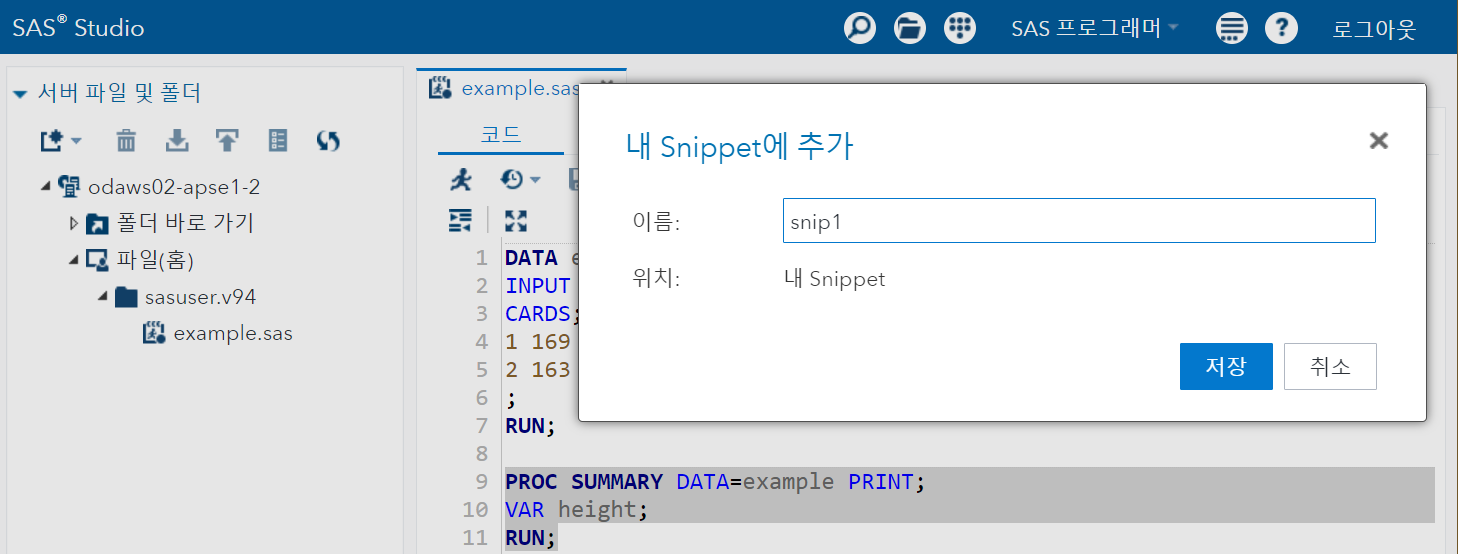
저장한 파일을 다운로드 할 수 있다. ‘서버 파일 및 폴더’탭에서 해당 파일에 오른쪽 클릭을 하면 ‘파일 다운로드’메뉴가 있다. PC에 있는 SAS 프로그램을 업로드할 수 있다. ‘서버 파일 및 폴더’ 탭에서 원하는 위치에서 오른쪽 클릭을 하면 ‘파일 업로드’메뉴가 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

◯ ‘Snippet’에 저장하기 (ALT-A, ALT-I)

Snippet은 완전한 코드는 아니지만 반복적으로 쓸 것 부분을 적어두는 메모장 같은 것이다. 프로그램 중 일부분을 snippet에 저장해 두면 반복적으로 사용하거나 중요한 공식이나 명령문을 공유할 때 편리하다. 원하는 부분을 마우스로 긁어서 선택하고 아이콘이나 ALT-A를 누르면 이름을 지정하여 snippet으로 저장할 수 있다.

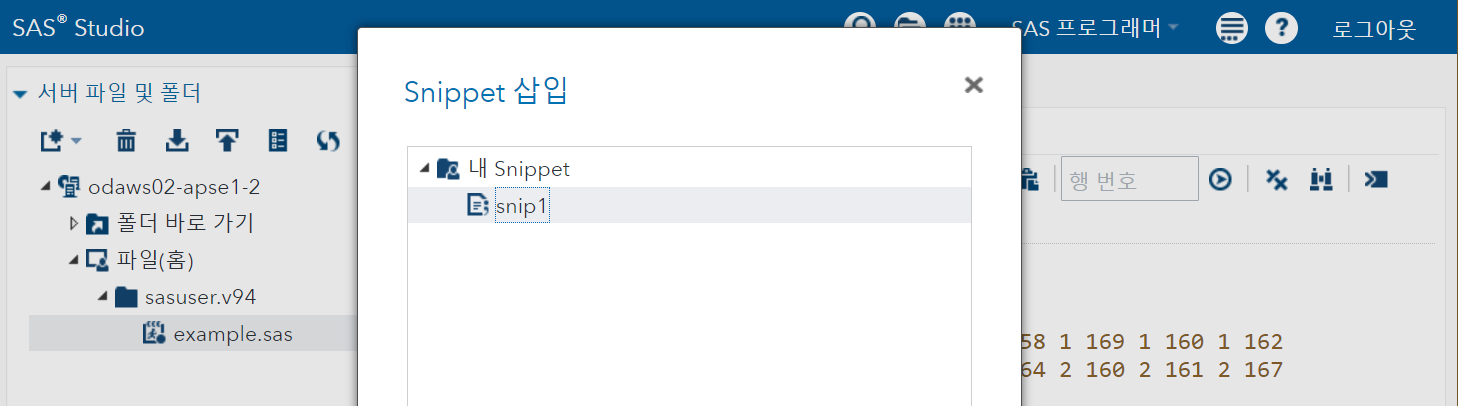


왼쪽에 ‘Snippet’탭을 선택하여 저장한 snip1을 확인할 수 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

SAS 프로그램을 작성하는 중에, 저장해 놓은 snippet을 사용하고 싶으면, ALT-I를 눌러 snippet을 선택한다. 프로그램의 현재 위치에 해당 snippet이 삽입된 것을 확인할 수 있다.



### ④ 저장 – 데이터

SAS는 데이터를 라이브러리 아래 저장한다. 왼쪽의 ‘라이브러리’탭에서 현재 라이브러리 리스트를 제공한다. SAS는 사용자를 위한 라이브러리로 WORK를 기본으로 제공한다. 앞에서 작성한 DATA example이 해당 라이브러리에 저장된 것을 볼 수 있다. 그러나 WORK에 저장된 데이터는 로그아웃 시에 모두 삭제된다. 따라서 보존하고 싶은 데이터는 새로운 라이브러리를 만들어 저장해야 한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

라이브러리를 만드는 순서는 아래와 같다.

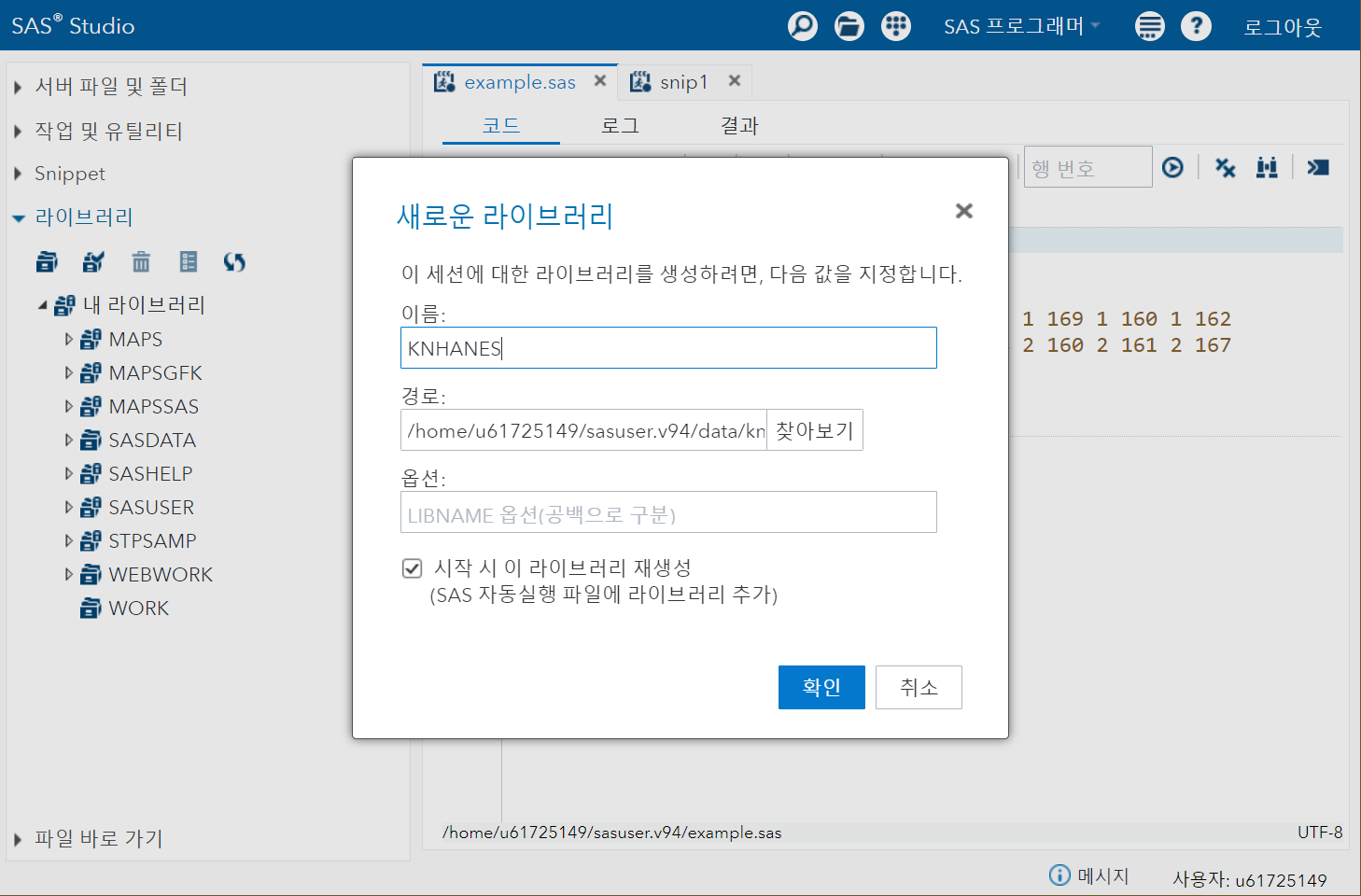
1. ‘서버 파일 및 폴더’탭에서 ‘파일(홈)/sasuser.vXX/data/<라이브러리 이름> 폴더 생성

텍스트, 테이블이(가) 표시된 사진

자동 생성된 설명

 2. ‘라이브러리’탭에서 아이콘을 눌러 ‘새로운 라이브러리’ 선택

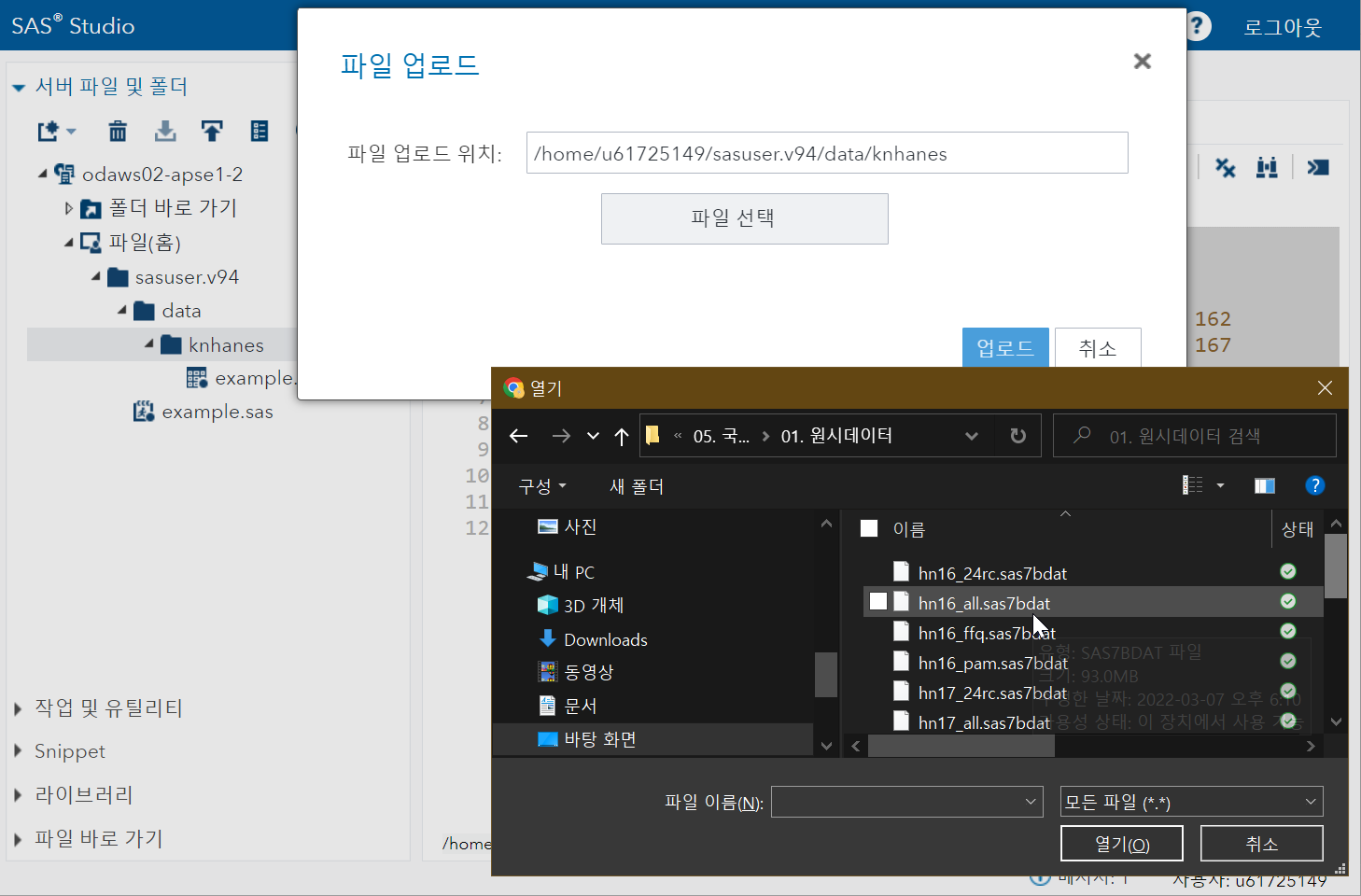
- 위에서 생성한 경로를 사용, 그리고 ‘시작 시 라이브러리 재생성’ 선택



해당 라이브러리에 데이터를 생성하거나 사용할 때는 <라이브러리이름>.<데이터이름> 형식을 쓴다. 앞의 example이 knhanes를 사용하도록 고쳐 쓰면 아래와 같다.

|  |
| --- |
| DATA knhanes.example;  INPUT class height @@;  CARDS;  1 169 1 168 1 164 1 165 1 159 1 158 1 158 1 169 1 160 1 162  2 163 2 160 2 160 2 164 2 160 2 161 2 164 2 160 2 161 2 167  ;  RUN;  PROC SUMMARY DATA=knhanes.example PRINT;  VAR height;  RUN; |

다운받은 국건영 원시데이터를 라이브러리를 저장한 폴더에 업로드하고 분석할 수 있다.



**오른쪽 클릭하여 ‘파일 업로드’ 선택**

### ⑤ 디렉토리 구조 잡기

SAS 분석을 하는데 있어서, 어디에 프로그램을 저장할 것인지, 어디에 데이터를 저장할 것인지, 프로그램 이름은 어떻게 할 것인지, 라이브러리 이름을 어떻게 할 것이지 정해진 것은 없다. (단, 허용된 문자, 길이 제한은 있다.) 그러나 실무실습을 하는 작은 규모의 프로젝트에 대해서 아래와 같이 구조를 잡으면 체계가 있다고 생각하여 그렇게 하였다.

텍스트이(가) 표시된 사진

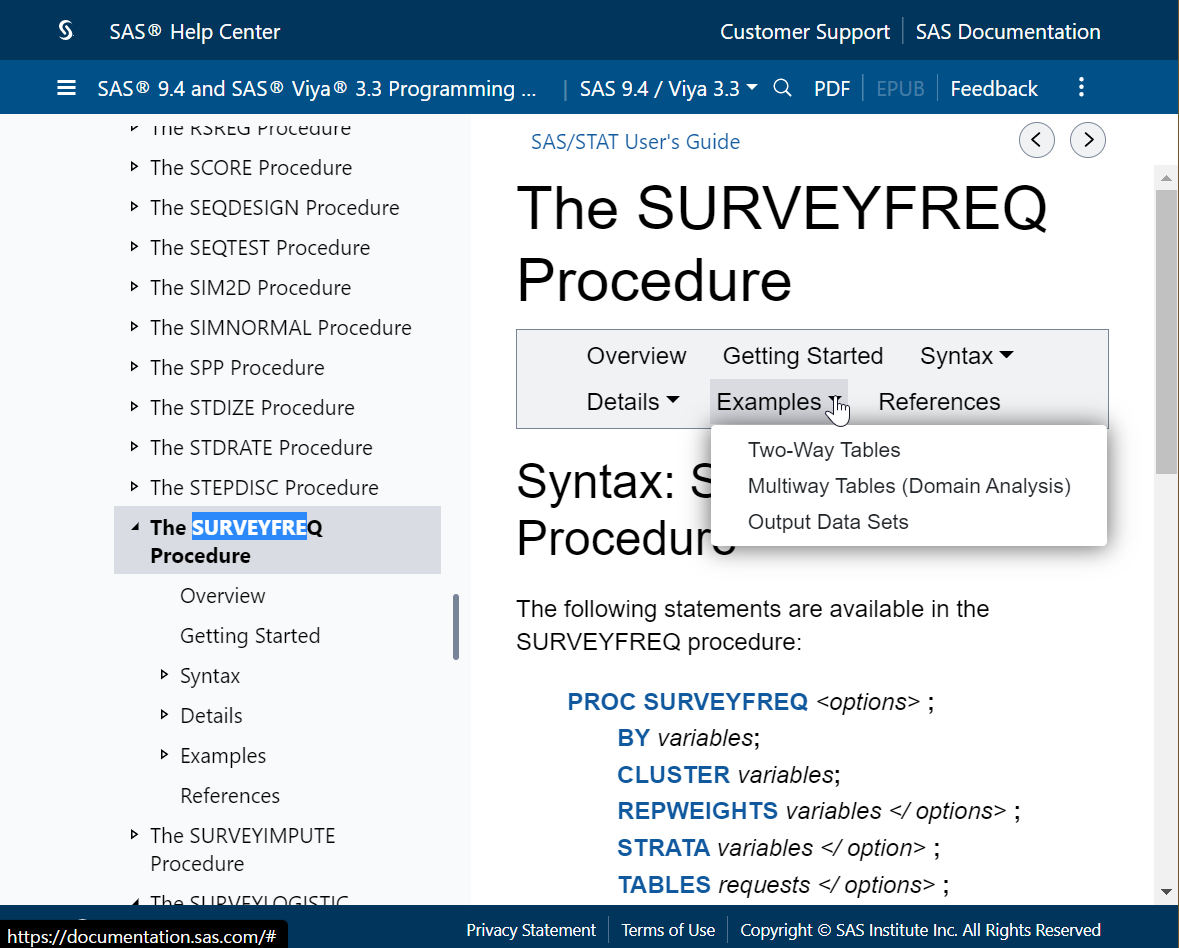
자동 생성된 설명

앞의 예제와 다른 점은 ‘파일(홈)/sasuser.v94’ 디렉토리를 사용하지 않고 ‘파일(홈)’ 디렉토리를 바로 사용한 점이다.

앞의 예제와 같은 점은 코드를 ‘code’ 디렉토리 아래에 저장하였고 데이터를 ‘data’ 디렉토리 아래에 저장하였다는 점이다.

### ⑥ 프로시저 설명과 예제 – 매뉴얼 참고하기

SAS 통계분석을 하다보면 프로시저에 대해 궁금하거나 도움이 될 만한 예제를 찾아야 할 때가 많다. 구글 검색을 통해 원하는 정보를 얻을 수 없다면 SAS® Help Center 페이지를 참고할 수 있다. 여기서 각종 온라인 매뉴얼과 개별 프로시저에 대한 예제를 찾을 수 있다.



# 사전 지식 요약 - 단순임의표본 자료 분석

통계분석에 관한 검색이나 책을 통해 아래 개념을 이해하고 있으면 좋다. 차이는 무엇인지, 언제 사용하는지, 검정 통계량은 무엇인지 그리고 검정•분석방법의 결과로부터 무엇을 얻을 수 있는지 알고 있으면 충분하다.

① p-value

② T-검정(t-test)

③ 카이제곱검정(χ² test)

④ 분산분석(ANOVA)

④ 공분산분석(ANCOVA)

⑤ 회귀분석(Regression analysis)

⑥ 로지스틱 회귀분석(Logistic regression analysis)

# 복합표본설계 자료 분석[[2]](#footnote-2)

### ① 국민건강영양조사 자료의 특징

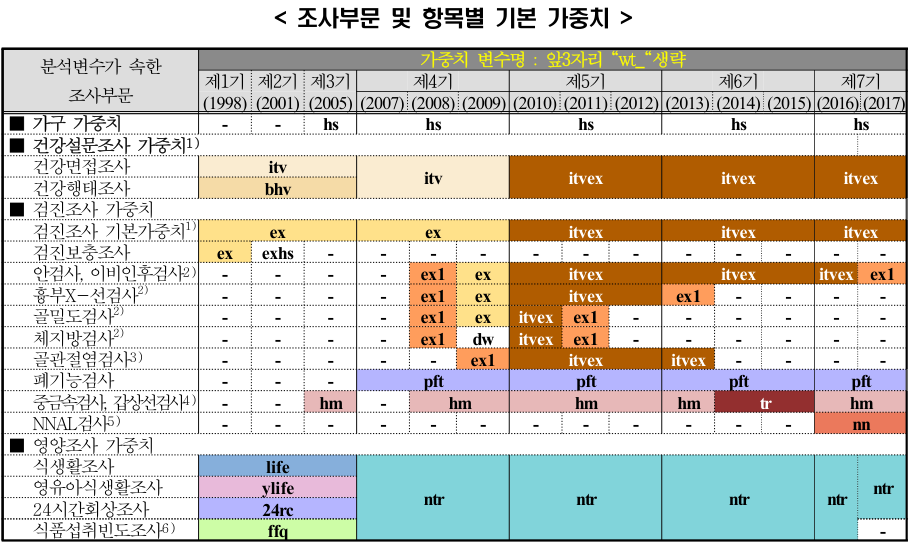
연구조사에서 Simple random sampling(SRS)은 현실적으로 불가능하다. 예를 들면, 한국인 고혈압환자를 대표하는 표본을 얻기도 어렵고 샘플이 적으면 편향이 발생할 수 있다. 국건영 데이터는 한국인 전체 집단에 대한 대표성을 가지고 있다. 그러나 표본추출설계는 SRS을 사용하지 않고 다단계층화집락추출법(2단계집락추출+층화)을 사용한 복합표본설계이다.

국건영의 표본추출 기본단위(psu – primary sampling unit)는 인구주택총조사의 조사구이다. 인구주택총조사를 모집단으로 하여 층화(지역, 주거유형 – 아파트/일반, 동/읍면)의 비율에 따른 추출 조사구수를 정하고[[3]](#footnote-3) 해당 조사구에서 가구와 가구원을 선정하여 조사한다.

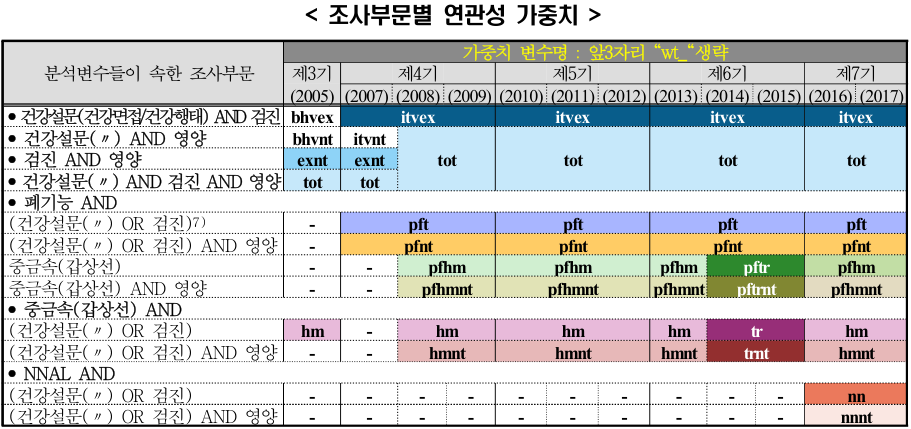
텍스트이(가) 표시된 사진

자동 생성된 설명

조사가 끝나면 설계가중치, 응답률가중치, 기본가중치를 구하여 적용하고 시도•성•연령별 가중치를 사후보정한다. 가중치는 조사부문(가구, 건강면접조사, 검진조사, 영양조사)마다 따로 있고, 각 부문별로 가중치의 합은 우리나라 인구수이다. 즉, 가중치는 데이터 1개가 얼마나 많은 인구를 대표하는지를 의미한다. 따라서, 국건영 자료로 연구할 때, 조사부문에 맞는 가중치를 사용하여야 한다. 예로, 제7기(2016~2018)의 건강면접조사의 경우 wt\_itvex를 사용한다. (사실, 기수내 3년 자료를 통합하므로 가구수 비율로 가중치를 보정이 필요하다. 가중치보정한 자료 통합은 이 문서의 ‘코드 구조’항목, 혹은 각주1를 참조할 수 있다.)



또, 다른 조사부문 사이의 연관성 연구를 위한 가중치가 별도로 존재한다.



국민건강영양조사의 표본설계에 대해 더 자세한 연구는 각주[[4]](#footnote-4)를 참고할 수 있다.

### ② SAS 복합표본 자료분석

국민건강영양조사 자료를 분석하는 경우에는 복합표본설계 정보를 고려한 분석방법을 사용해야 하는데, 이를 고려하지 않는 경우(=단순임의표본분석) 추정치(평균, 유병률, 오즈비 등) 및 이들의 분산(표준오차) 추정치에서 편향된 결과를 얻을 수 있다. 따라서 국민건강영양조사 자료분석 시에는 복합표본 자료분석 방법을 사용하여야 하며, SAS에서 복합표본 자료분석은 SURVEY 프로시저를 이용하여 수행한다.1

“국민건강영양조사 분석지침서(SAS). 질병관리본부. 2019”의 “04 복합표본 자료분석”은 SURVEY 프로시저를 이용한 분석방법과 예제를 제공한다.

테이블이(가) 표시된 사진

자동 생성된 설명

단순임의표본 분석 프로시저와 복합표본 분석을 위한 SURVEY 프로시저는 1대 1로 대응하지 않는다. 단순임의표본 분석의 경우 T-검정, ANOVA, ANCOVA, 회귀 분석을 할 때, 각기 다른 프로시저를 사용한다. 그러나 복합표본설 자료를 분설할 때는, SURVEYREG 프로시저 하나를 사용하여 T-검정, 회귀분석, 공분산분석, 회귀분석을 모두 수행한다. 로지스틱 회귀분석의 경우, 단순임의표본에서와 같이 별도의 프로시저인 SURVEYLOGISTIC를 사용하고, 빈도분석과 카이검정을 할 때, SURVEYFREQ 프로시저를 사용한다.

***국민건강영양조사 분석지침서(SAS)를 전체 읽어볼 것을 권한다.***

# 코드 구조

### ① 매크로

프로그래밍 언어에서 매크로는 반복적으로 사용할 명령문을 더 간단한 문자열에 저장해 두는 것이다. 매크로를 정의하고 나면, 프로그램을 작성할 때 해당 문자열을 사용할 수 있다. 실행 시, 매크로가 있는 자리에 원래 명령문이 치환되어 들어간다.

SAS에서 매크로를 사용하면 번거로운 일을 피해갈 수 있다. 다음은 SAS에서 매크로 변수를 사용하여 치환하는 예이다. 분석방법과 데이터는 동일하지만, 다양하게 가공한 여러 데이터를 분석할 때 DATA와 STUDY\_DOMAIN 값만 수정하는 방식으로 활용할 수 있다.

|  |  |
| --- | --- |
| 실행 전(원본 프로그램) | 실행 시점(SAS 내부에서 자동변환) |
| %LET DATA=data\_by\_imj;  %LET STUDY\_DOMAIN=score\_yes;  PROC SURVEYFREQ DATA=&DATA NOMCAR;  STRATA kstrata;  CLUSTER psu;  WEIGHT wt\_ex\_pool;  TABLES &STUDY\_DOMAIN\*…  …  RUN;  PROC SURVEYLOGISTIC DATA=&DATA NOMCAR;  STRATA kstrata;  CLUSTER psu;  WEIGHT wt\_ex\_pool;  DOMAIN &STUDY\_DOMAIN;  …  RUN; | PROC SURVEYFREQ DATA data\_by\_imj NOMCAR;  STRATA kstrata;  CLUSTER psu;  WEIGHT wt\_ex\_pool;  DOMAIN score\_yes;  …  RUN;  PROC SURVEYLOGISTIC DATA data\_by\_imj NOMCAR;  STRATA kstrata;  CLUSTER psu;  WEIGHT wt\_ex\_pool;  DOMAIN score\_yes;  …  RUN; |

### ② 주석

프로그래밍 언어에서 주석(comment)은 프로그램의 설명을 적거나 일부 코드를 실행하지 않도록 잠시 막아 두는 데 쓰인다. SAS에서 두가지 주석처리 방법을 사용할 수 있다.

|  |  |
| --- | --- |
| 코드 | 설명 |
| /\*  /\*  \* 이것은 그냥 설명입니다.  \* 이 방식은 여러 줄에 걸쳐 설명을  \* 적을 수 있는 장점이 있습니다  \*/  PROC SURVEYLOGISTIC DATA=BlahBlah;  …  RUN;  %MACRO DO\_SOMETHING;  PROC SURVEYLOGISTIC DATA=BlahBlah;  …  RUN;  %MEND;  %MACRO DO\_ANOTHERTHING;  PROC SURVEYLOGISTIC DATA=BlahBlah;  …  RUN;  %MEND;  %DO\_SOMETHING;  \*%DO\_ANOTHERTHING; | /\* ~ \*/ 는 여러 줄에 걸쳐 주석을 쓸 때 좋다.  이 분석은 여기서 실행된다.  이 분석은 %DO\_SOMETHING이 나타나는 자리에 치환되어 그 자리에서 실행된다.  이 분석은 %DO\_ANOTHERTHING이 나타나는 자리에 치환되어 그 자리에서 실행된다.  DO\_SOMETHING 매크로가 치환되어 실행된다.  DO\_ANOTHERTHING은 \* ~ ;로 주석 처리되었기 때문에 치환되지 않고 실행되지 않는다. |

### ③ prepare-dataset.sas

이후 분석에 필요한 데이터를 준비한다. 이 파일에서 아래의 매크로를 정의한다.

|  |  |
| --- | --- |
| 매크로 | 요약 |
| LIBRARY | 라이브러리 이름  아래 DATA\_\* 항목들은 LIBRARY 아래에 위치한다. |
| DATA\_RAW | 원시자료 혹은 원시자료의 결합자료 |
| DATA\_KEEP | DATA\_RAW에서 필요한 변수만 KEEP한 자료 |
| DATA\_STUDY\_SELECTION | DATA\_KEEP에서 연구대상만 표시한 자료 |
| DATA\_STUDY\_READY | DATA\_STUDY\_SELECT에 필요한 변수를 생성한 자료  (참조: 생성변수, “02 연구설계.docs”) |
| DO\_DATA\_RAW | DATA\_RAW를 생성하는 데이터 단계 실행 |
| DO\_DATA\_KEEP | DATA\_KEEP을 생성하는 데이터 단계 실행 |
| DO\_STUDY\_SELECTION | DATA\_STUDY\_SELECTION을 생성하는 데이터 단계 실행 |
| DO\_STUDY\_READY | DATA\_STUDY\_READY를 생성하는 데이터 단계 실행 |

LIBRARY, DATA\_\* 매크로를 사용하면 여러 연구자가 라이브러리 이름과 데이터 이름을 원하는 데로 정할 수 있다. 그리고, 다른 연구자의 데이터를 받아서 분석을 실행할 때, 라이브러리와 데이터 이름만 수정하여 응용할 수 있기를 의도하였다.

DO\_\* 매크로를 사용하여 원하는 데이터 단계를 선택적으로 실행할 수 있다. DATA\_RAW는 연구 중에 수정할 일이 거의 없고, 연구가 진행되면 DATA\_KEPP, DATA\_STUDY\_SELECTION 등도 수정할 일일 거의 없어진다. 그리고, 원시데이터가 특정 환경에서만 처리 가능할 때, 그 환경에서 해당 데이터 단계를 실행하고, 다른 환경에서는 결과 데이터를 받아서 그 후의 데이터 단계와 분석을 문제없이 실행할 수 있기를 의도하였다.

국민건강영양조사 원시자료 내의 한글은 EUC-KR로 인코딩되어 있다. PC버전 SAS에서는 잘 불러올 수 있지만, UTF-8 인코딩만 지원하는 온라인 버전 SAS OnDemand에서는 온전히 불러올 수 없다. 따라서 PC버전에서 데이터 호환성 문제를 해결한 파일을 만들고 온라인 버전에서 공유하는 방법을 쓰기로 하였다. PC버전에서 DO\_DATA\_RAW, DO\_DATA\_KEEP 데이터 스텝을 실행하면 한글을 포함하는 변수를 제거하므로 생성된 데이터파일을 공유하여 같이 연구할 수 있었다.

### ④ explore-survey.sas

앞에 정의한 매크로 중 DATA\_STUDY\_READY를 사용하므로 아래와 같이 시작하도록 하였다. 또, 여러 분석 변수에 대해 동일한 분석방법을 진행할 수 있다고 가정하였기 때문에 연구대상 선택 변수를 지정하는 매크로를 추가하였다.

|  |
| --- |
| %LET LIBRARY=KNHANES;  %LET DATA\_STUDY\_READY=&LIBRARY..HN16\_17\_18\_STUDY\_READY;  %LET STUDY\_IN=phq\_in; /\* phq\_in, eq5d\_in or bp7\_in \*/  %LET STUDY\_OUTCOME=phq; /\* phq, LQ\_5EQL, BP7 \*/  PROC SURVEYFREQ DATA=& DATA\_STUDY\_READY NOMCAR;  STRATA kstrata;  CLUSTER psu;  WEIGHT wt\_ex\_pool;  TABLES &STUDY\_IN\*&STUDY\_OUTCOME\* …  …  RUN; |

### ⑤ analyze-survey.sas

explore-survey.sas 항목과 동일

1. 검토와 개선을 위해 작성되었습니다. 모든 종류의 수정과 재배포를 허용합니다. [↑](#footnote-ref-1)
2. 국민건강영양조사 원시자료 분석지침서(SAS). 질병관리본부. 2019. [↑](#footnote-ref-2)
3. “2005년 인구주택총조사의 조사구 설정 방법(이건 등. 2006)”에 따르면 주사구는 층화내용을 포함. [↑](#footnote-ref-3)
4. 제4기 국민건강•영양조사를 위한 순환표본 설계연구. 이계오, 박진우. 조사연구(2007. 9). 8권 2호. [↑](#footnote-ref-4)