

11-1790399-000001-01

Korean Genome and Epidemiology Study

가이드북

R편



질병관리청



국립보건연구원



한국인유전체역학조사사업

간행물발간등록번호
11-1790399-000001-01

한국인유전체역학조사사업 (KoGES)
Korean Genome and Epidemiology Study

KoGES 데이터 분석 가이드북

R편



질병관리청



국립보건연구원

KoGES
한국인유전체역학조사사업

들어가는 말

질병관리청 국립보건연구원 유전체역학과에서는 만성질환 발생의 유전, 환경적 위험요인을 규명하기 위하여 한국인유전체역학조사사업(KoGES)을 수행하여 왔으며, 약 23만 5천 명에 대한 역학정보와 인체자원을 수집하여 보건의로 R&D 연구자원으로서 국내 의·과학 연구자들에게 제공하고 있습니다.

본 안내서는 KoGES 수집자료를 이용하고자 하는 연구자의 이해를 높이고자, 교육용 데이터를 이용하여 데이터 편집 및 가공, 예제를 통한 통계 분석 및 결과 해석에 대한 전반적인 내용을 담고 있습니다. 특히, 공개 분양되고 있는 KoGES 자료와 동일한 구조의 교육용 데이터를 이용하여, 통계 프로그램에서 자료 불러오기 및 변수 가공 등 분석 이전의 자료 핸들링 단계에 대한 설명이 강조되어 있으며, 통계 분석의 경우 실제 예제를 소개함으로써 연구자가 쉽게 이해하고, 따라할 수 있도록 구성하였습니다. 또한 통계분석은 오픈 소스인 R(version 3.6.3)을 기반으로 사용자 친화적 통합 개발 환경인 RStudio를 활용하여 진행되었습니다.

본 안내서에서 정의된 질병 및 기타 가공 변수는 해당 분석 예제를 설명하기 위하여 가정한 하나의 예시일 뿐, 연구자의 연구 목적에 따라 적절하게 변환하여 정의하실 수 있습니다. 또한 기반조사와 추적조사 교육용 데이터는 KoGES 자료의 일부를 임의 추출하여 가공 생성한 자료이므로 교육용 목적 이외에 논문 작성을 위한 분석 자료로는 적합하지 않음을 알려드립니다.

연구 목적으로 KoGES 자료를 활용하고자 하는 연구자는 질병관리청 홈페이지(<http://www.kdca.go.kr>) 또는 국립보건연구원 홈페이지(<http://www.nih.go.kr>)에서 자세한 내용을 확인할 수 있습니다.

본 안내서를 통하여 보다 많은 연구자가 한국인유전체역학조사사업 자료를 활용하게 되고, 나아가 국민보건 향상에 도움이 되는 좋은 연구 성과를 거두시기를 기대합니다.

Contents

1장. 한국인유전체역학조사사업(KoGES)

소개

1. 사업 개요	8
2. 사업 수행 체계	9
3. 사업 추진 현황	10
3-1. 세부 코호트	10
3-2. 추적조사 현황	11
4. 조사 항목	12
5. 자료 분량	15

2장. 한국인유전체역학조사사업(KoGES)

역학자료 소개

1. 공개 자료 목록	18
2. 공개 자료 변수 목록(코드북)	19
2-1. 코드북 다운로드	19
2-2. 코드북 구성	20
3. 분량 자료 파일 유형 및 형태	22

3장. 한국인유전체역학조사사업(KoGES)

기반조사 자료 분석하기

1. 기반조사 교육용 데이터 이해하기	26
1-1. 자료 구성	26
1-2. 코드북	27
2. 자료 불러오기	28
2-1. 자료 불러오기	28
2-2. 불러온 자료 확인하기	31
3. 자료 결합하기	33
4. 자료 분석 준비하기	38
4-1. 기본코드 결측치 처리하기	38
4-2. 변수 유형 변환하기	40
5. 자료 분석하기	43
5-1. 분석 대상자 선정	43
5-2. 변수 생성	45
5-3. 빈도 분석	47
5-4. 기술통계	50
5-5. 두 집단 평균 비교	56
5-6. 분산분석	59
5-7. 선형 회귀분석	64
5-8. 로지스틱 회귀분석	70

4장. 한국인유전체역학조사사업(KoGES)

추적조사 자료 분석하기

1. 추적조사 교육용 데이터 이해하기	78
1-1. 자료 구성	78
1-2. 코드북	79
2. 자료 불러오기	80
2-1. 자료 불러오기(CSV 파일)	80
2-2. 불러온 자료 확인하기	81
3. 자료 결합하기	82
4. 자료 분석 준비하기	84
4-1. 기본코드 결측치 처리하기	84
4-2. 변수 유형 변환하기	85
5. 자료 분석하기	87
5-1. 분석 대상자 선정	87
5-2. 변수 생성	88
5-3. 총 관찰인년(Person-years) 산출	94
5-4. 생존함수 추정(생존곡선)	95
5-5. Cox 비례위험모형	97

부록

1. R과 RStudio 설치하기	104
2. RStudio 들어가기	111
3. 분석 목적에 따른 통계분석 방법 요약	115

KoGES 데이터 분석 가이드북
[R편]

Korean Genome and Epidemiology Study

1장.

한국인유전체역학조사사업(KoGES) 소개

1. 사업 개요
2. 사업 수행 체계
3. 사업 추진 현황
4. 조사 항목
5. 자료 분양

1장.

한국인유전체역학조사사업(KoGES) 소개

1. 사업 개요

한국인유전체역학조사사업(Korean Genome and Epidemiology Study, 이하 KoGES)은 질병관리청 국립보건연구원 유전체역학과에서 수행 중인 코호트 사업으로, 한국인에서 흔하게 발생하는 당뇨, 고혈압, 심혈관질환 등 만성질환의 유전 및 환경적인 위험 요인을 밝히고 이들 간의 상호작용을 규명하고자, 전국 50개 이상의 의과대학 및 의료기관의 연구자들이 참여하여 코호트를 구축하였다.

세부 코호트는 40세 이상 성인을 대상으로 구축한 ‘일반인 기반(population-based) 코호트’와 만성질환의 유전-환경 상호작용 위험요인 규명을 위한 ‘유전-환경(gene-environment) 모델 코호트’로 구성된다. 일반인 기반 코호트는 지역사회기반 코호트(안산, 안성), 도시기반 코호트, 농촌기반 코호트로 구성되고, 유전-환경 모델 코호트는 쌍둥이 및 가족기반 코호트, 국내이주자 코호트, 국제협력 코호트로 구성된다.

코호트 참가자의 생활습관, 질병 과거력 등의 조사를 위하여 설문조사와 검진 등을 수행하며, 수집된 인체자원(혈청, 혈장, 뇨, DNA 등)은 향후 연구를 위하여 국립보건연구원 국립중앙인체자원은행에 저장하고 있다. 보관된 자료와 분석된 정보는 심의를 거쳐 한국인 호발성 만성질환을 연구하는 의·과학 연구자들에게 무상으로 제공하여, 국가 보건의료 연구 인프라로서의 중심 역할을 하고 있다.



| 그림 1 | KoGES 코호트 구성

2. 사업 수행 체계

KoGES 조사 수행 기관인 의료기관 및 검진센터에서는 참여자를 모집하여 동의서를 구득한 후 설문 조사와 검진을 통해 자료와 인체자원을 수집하며, 수집된 자료와 인체자원은 질병관리청 국립보건연구원으로 이송된다. 고품질의 자료 수집을 위하여 유전체역학과에서는 조사원 표준화 교육과 CAPI(Computer Assisted Personal Interview) 시스템 등 조사단계의 자료 질 관리 및 수집 후 심층 자료 정제를 수행하고 있다. 또한 인체자원의 수집과 보관의 전 과정에 걸쳐 표준화 프로토콜을 적용하여 고품질의 자료와 인체자원을 연구자들에게 분양할 수 있도록 관리하고 있다.

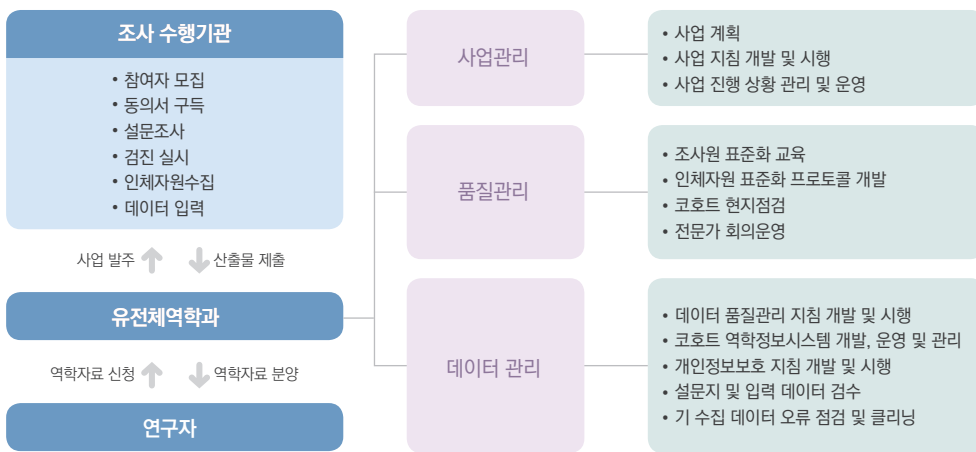
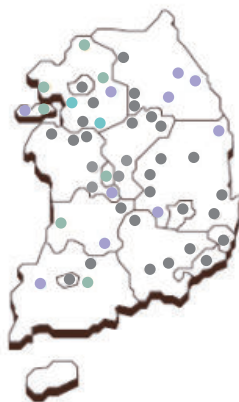


그림 2 | KoGES 조사 수행 체계

KoGES 코호트 및 참여기관



■ 지역사회 기반 코호트 사업 (안산, 안성)	고려대학교 안산병원 아주대학교 임상역학센터		
■ 농촌 기반 코호트 사업	한양대학교 전남대학교 원광대학교 계명대학교	경북대학교 조선대학교 연세대학교	관동대학교 충남대학교
■ 도시 기반 코호트 사업	서울대학교 인제대학교 한인대학교 전북대학교 전남대학교 인제대학교 해운대 백병원 강북삼성병원 인제대학교 상계백병원 울산대학교 병원 삼성상당병원 동국대학교 임산병원 건국대학교 충주병원 단국대학교 병원 이화여대 동독병원	한림대학교 경북대학교 분당서울대학교병원 순천성심병원 녹색병원 강원대학교병원 순천향대학교 천안병원 고신대학교 복음병원 가톨릭대강병원 가톨릭대학교 부산병원 동아대학교병원	전남대학교병원 고려대학교 안산병원 대전중앙의료재단 인천 세브란스병원 대구가톨릭대학교 병원 창원파티마병원 인하대학교병원 경주시 보건소 한림대학교 성심병원 경희대학교
■ 쌍둥이 및 가족 코호트 사업	서울대학교 인제대학교 부산백병원	삼성서울병원	단국대학교병원
■ 국내이주자 코호트 사업	이화여자대학교		
■ 국제협력(국외이민자) 코호트 사업	성균관대학교 길림대학교 고베아사히병원	연변대학교	한양대학교

그림 3 | KoGES 조사 수행 기관

3. 사업 추진 현황

3-1. 세부 코호트

각 세부 코호트의 특성과 참여자 규모는 <표 1>과 같다.

표 1 | KoGES 세부 코호트 요약

세부 사업		시작연도	대상지역	현 황
일반인 기반 코호트	지역사회 기반 코호트	2001년	안산, 안성	<ul style="list-style-type: none"> 안산, 안성에 거주하는 40~69세 남녀 대상으로 모집 기반 참여자수: 10,030명(안산: 5,012명, 안성: 5,018명) 2020년 9차 추적조사 진행 중(2년 간격 추적 실시)
	농촌기반 코호트	2004년	전국 11개 농촌 지역	<ul style="list-style-type: none"> 농촌 지역에 거주하는 40세 이상 남녀 대상으로 모집 기반 참여자수: 28,337명 2007~2016년 6개 지역 중심 추적조사 수행
	도시기반 코호트	2004년	전국 의료기관 중심	<ul style="list-style-type: none"> 도시 지역 의료기관 검진센터에 내원한 40세 이상 남녀를 대상으로 모집 기반 참여자수: 173,208명 2012~2016년 17개 기관 중심 추적조사 수행
유전-환경 모델 코호트	쌍둥이 및 가족 코호트	2005년	서울, 부산	<ul style="list-style-type: none"> 쌍둥이와 그 가족을 대상으로 모집 기반 참여자수: 3,202명 2008~2014년 추적조사 수행
	국제협력 I 코호트	2005년	일본, 중국	<ul style="list-style-type: none"> 일본(고베, 오사카) 및 중국(장춘, 연변) 거주 한국인으로 이주 후 15년 이상인 사람과 자손 및 현지인 대상으로 모집 기반 참여자수: 3,556명 일본(고베, 오사카): 2008~2013년 추적조사 수행 중국(장춘, 연변): 2008~2011년 추적조사 수행
	국내이주자 및 국제협력 II 코호트	2006년 2008년	전국, 베트남, 캄보디아	<ul style="list-style-type: none"> 한국인과의 결혼으로 국내에 이주한 아시아 국가 출신의 여성 및 배우자, 자녀를 대상으로 모집 : (국내이주자) 결혼 이주 여성과 배우자 및 그 자녀 : (국제협력 II) 결혼 이주 여성 현지 가족 및 지역 주민 기반 참여자수: 국내이주자 7,191명, 국제협력 II 4,054명 국내이주자 및 자녀: 2012~2014년 추적조사 수행
기타 코호트	중규모 코호트	2003년	전국	<ul style="list-style-type: none"> 우편설문조사를 통한 건강 관련 설문정보 및 구강상피세포를 이용한 DNA 수집 기반 참여자수: 10,302명
	사상체질 코호트	2004년	전국	<ul style="list-style-type: none"> 40세 이상 남녀 대상으로 사상체질 등 한의학적 이론 적용을 통한 만성질환 발생에 미치는 영향 연구를 위한 코호트 구축 기반 참여자수: 3,062명

3-2. 추적조사 현황

코호트 참여자의 추적조사는 통계청 사망자료, 건강보험공단 수진자료 및 암센터 암등록 자료 등의 자료 연계를 통한 방식과 코호트 참여자를 재 접촉하여 기반조사와 동일하거나 확대된 직접 조사를 수행하는 반복추적조사를 병행하여 진행하고 있다. 지역사회기반 코호트 사업은 기반조사 후 2년 마다 반복추적조사를 수행하여 현재 9차 추적조사가 진행 중이며, 도시기반과 농촌기반, 쌍둥이 및 가족 코호트 등은 1차 이상의 추적조사를 수행하였다.

기반	지역사회 기반	도시기반	농촌기반	쌍둥이 및 가족	국내 이주자	국제협력	
	10,030 (‘01~’02)	173,208 (‘04~’13)	28,337 (‘04~’13)	3,202 (‘05~’13)	7,191 (‘06~’14)	일본 1,063 (‘05~’07,‘11)	중국 2,493 (‘05~’06,‘08)
1차 추적	8,603 (‘03~’04)	65,616 (‘12~’16)	12,463 (‘07~’14)	2,030 (‘08~’14)	1,824 (‘12~’14)	773 (‘08~’09,‘13)	964 (‘08~’11)
2차 추적	7,515 (‘05~’06)		11,399 (‘08~’16)	940 (‘09~’14)		549 (‘10~’11)	
3차 추적	6,688 (‘07~’08)		6,423 (‘11~’16)	165 (‘12~’14)		520 (‘12~’14)	
4차 추적	6,665 (‘09~’10)		1,449 (‘14~’16)				
5차 추적	6,238 (‘11~’12)						
6차 추적	5,906 (‘13~’14)						
7차 추적	6,318 (‘15~’16)						
8차 추적	6,157 (‘17~’18)						
9차 추적	진행 중 (‘19~’20)						

| 그림 4 | KoGES 추적조사 현황

4. 조사 항목

2001년부터 시작된 KoGES는 공통 조사 항목 외에 다양한 지역과 집단을 대상으로, 여러 수행기관에서 참여자를 모집, 조사를 수행하여 코호트별 목적에 따라 조사 항목에 일부 차이가 있다. KoGES 자료는 일반사항, 질병 과거력, 질병 치료 현황, 약물력, 가족력, 음주 및 흡연 등 생활습관, 신체활동 등이 포함된 설문 항목과 혈압 및 맥박 측정, 체성분분석, 혈액 및 소변검사 등 임상검사, 심전도, 흉부 X-ray, 폐기능 검사 등이 포함된 검진 항목으로 구성되어 있다.

| 표 2 | KoGES 조사 항목 - 설문조사

항목		일반인 기반 코호트			유전-환경 모델 코호트		
		지역사회기반	도시기반	농촌기반	쌍둥이 및 가족	국내이주자	국제협력 I
인구사회학		●	●	●	●	●	●
생활습관	흡연	●	●	●	●	●	●
	음주	●	●	●	●	●	●
	신체활동	●	●	●	●	●	●
	체중변화	●	-	-	-	●	△
질환력	과거력	●	●	●	●	●	●
	약물력	●	-	-	●	●	●
	보충제	●	●	●	●	●	△
	수술력	-	-	-	-	●	●
	가족력	●	●	●	●	●	●
	호흡순환기질환	●	-	△	-	●	-
식이조사	식품섭취빈도조사법	●	●	●	●	●	●
	24시간 회상법	-	●	-	-	-	-
	식습관	●	●	●	●	●	●
여성력		●	●	●	●	●	●
사회심리		●	●	●	-	●	●

● 전체 참여자, △ 일부참여자

| 표 3 | KoGES 조사 항목 - 신체계측

항목		일반인 기반 코호트			유전-환경 모델 코호트		
		지역사회기반	도시기반	농촌기반	쌍둥이 및 가족	국내이주자	국제협력 I
신장		●	●	●	●	●	●
체중		●	●	●	●	●	●
허리둘레		●	●	●	●	●	●
엉덩이둘레		●	●	●	●	●	●
혈압		●	●	●	●	●	●
맥박		●	●	●	-	●	●

● 전체 참여자, △ 일부참여자

| 표 4 | KoGES 조사 항목 - 혈액검사

항목	코호트	일반인 기반 코호트			유전-환경 모델 코호트		
		지역사회기반	도시기반	농촌기반	쌍둥이 및 가족	국내이주자	국제협력 I
Glucose (공복)		●	●	●	●	●	●
Glucose (당부하 60분)		●	-	-	-	-	-
Glucose (당부하 120분)		●	-	-	-	△	-
Insulin (공복)		●	-	△	●	-	-
Insulin (60분)		●	-	-	-	-	-
Insulin (120분)		●	-	-	-	-	-
HbA1c		●	△	△	△	-	△
Total protein		●	●	△	●	●	△
Uric acid		-	●	△	●	-	△
Creatinine		●	●	●	●	●	●
BUN		●	●	●	△	●	●
Albumin		●	●	●	●	●	●
Total Bilirubin		●	△	△	△	-	-
AST		●	●	●	●	●	●
ALT		●	●	●	●	●	●
γ-GTP		●	●	●	●	●	●
Total Cholesterol		●	●	●	●	●	●
HDL-Cholesterol		●	●	●	●	●	●
Triglyceride		●	●	●	●	●	●
hs-CRP		●	●	●	●	△	△
W.B.C blood		●	△	●	●	●	●
R.B.C blood		●	△	●	●	●	●
Hemoglobin		●	●	●	●	●	●
Hematocrit		●	△	●	●	●	●
Platelet		●	△	△	●	●	△
평균적혈구용적		-	△	△	●	●	△
평균적혈구혈색소		-	△	△	●	●	△
평균적혈구혈색농도		-	△	△	●	●	△
혈액형		-	-	-	-	●	△
Calcium		●	△	-	●	-	△
Sodium		●	-	-	-	-	△
Potassium		●	-	-	-	-	△
Chloride		●	-	-	-	-	△
Vitamin B12		●	-	-	-	-	-
Folate		●	-	-	-	-	-

● 전체 참여자, △ 일부참여자

표 5 | KoGES 조사 항목 - 소변검사

항목	코호트	일반인 기반 코호트			유전-환경 모델 코호트	
		지역사회기반	도시기반	농촌기반	국내이주자	국제협력 I
PH		●	●	●	●	△
Nitrite		●	-	-	●	△
S.G		●	-	-	●	-
Protein		●	●	●	●	●
Glucose		●	●	●	●	●
Ketone		●	-	●	●	△
Bilirubin		●	-	-	●	●
Blood		●	●	●	●	●
Urobilinogen		●	-	-	●	△
Color		●	-	-	●	-
R.B.C.		●	-	-	●	-
W.B.C.		●	-	-	●	△
E.P cells		●	-	-	●	-
Casts		●	-	-	●	-
Bacteria		●	-	-	●	-
Crystals		●	-	-	●	-
Others		●	-	-	●	-

● 전체 참여자, △ 일부참여자

5. 자료 분양

KoGES를 통해 수집된 역학자료(일반정보, 질병 과거력, 식·생활습관 등의 설문항목과 신체계측 및 임상검사)와 유전체정보(SNP 등) 및 인체자원(DNA, 혈청, 혈장 등)은 대학, 국·공립 및 사립 연구기관 등에 소속되어 연구 업무를 수행하는 연구자를 대상으로 분양하고 있다. 자료 분양은 역학자료만 요청하는 경우와 역학자료와 함께 유전체정보 및 인체자원을 요청하는 경우를 구분하여 두 가지 분양 절차를 운영하고 있다.

역학자료만 활용하고자 하는 연구자는 ‘KoGES 역학자료 분양 서비스’를 통하여 자료 분양 신청을 하면 된다. 자료 신청은 질병보건통합관리시스템(<http://is.kdca.go.kr>)을 통하여 이루어진다. 해당 시스템 이용을 위해서는 사용자 가입이 필요하며, 권한 신청 단계에서 ‘유전체역학-연구자’를 선택하여 연구자 권한 신청을 하면, 담당 부서의 승인 이후 자료 분양 신청이 가능하다. 질병관리청 홈페이지(<http://www.kdca.go.kr>) > 하단 관련 링크 > KoGES 또는 국립보건연구원 홈페이지(<http://www.nih.go.kr>) > 연구·사업 > 한국인유전체역학조사사업에서 분양에 필요한 서류(연구계획서, IRB 승인서 또는 면제서) 및 신청 방법 등을 확인할 수 있다.

자료는 크게 일반공개와 제한공개로 구분하여 제공하고 있다. 일반공개 자료는 분양 승인 후 익명화된 형태로 연구자에게 제공하고 있으며, 제한공개자료는 응답 비율이 낮거나 민감한 변수는 유전체역학과 분석실에 방문하여 분석 후 결과만 반출하도록 운영하고 있다. 마지막으로 KoGES 역학자료와 연계한 통계청 사망원인자료는 유전체 역학과 분석실에 방문하여 분석하거나, 원격분석을 통해 활용하도록 제한적으로 공개하고 있다.

다음으로, 역학자료와 유전체정보 또는 인체자원을 함께 분양 받고자 하는 경우에는 ‘인체자원 분양데스크’(<http://koreabiobank.re.kr>, 1661-9070)를 통하여 원하는 자료와 자원을 신청하면 된다. 이 경우는 월 1회 개최되는 분양심의위원회의 심의를 거치게 되며, 승인 후 분양된 자원은 직접 수령을 원칙으로 한다.

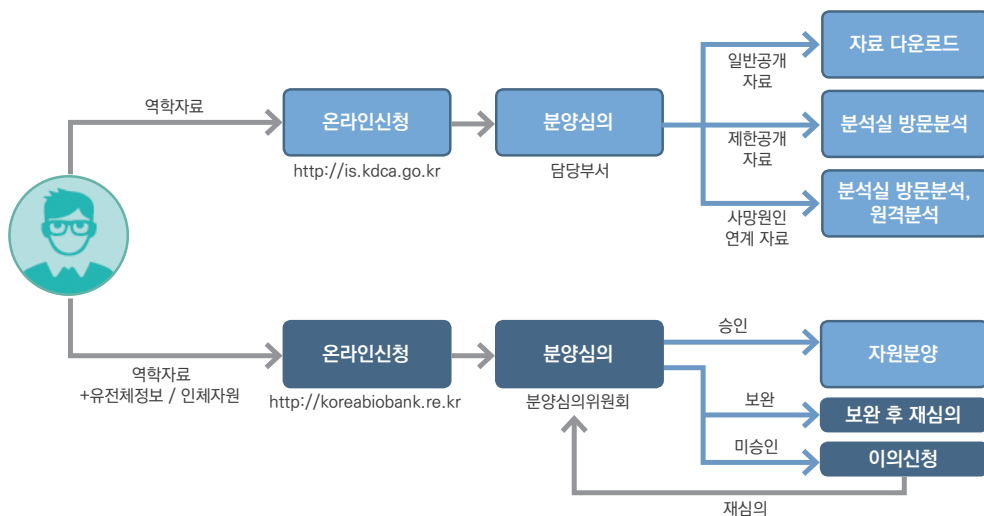


그림 5 | KoGES 자료 분양 체계

KoGES 데이터 분석 가이드북
[R편]

Korean Genome and Epidemiology Study

2장.

한국인유전체역학조사사업(KoGES) 역학자료 소개

1. 공개 자료 목록
2. 공개 자료 변수 목록(코드북)
3. 분양 자료 파일 유형 및 형태

2장.

한국인유전체역학조사사업(KoGES) 역학자료 소개

1. 공개 자료 목록

2020년 9월 기준 분양 가능한 KoGES 역학자료 목록은 아래와 같다. 세부 코호트와 기반/추적 자료에 따라 차이가 있지만 각 자료에는 크게 설문항목(일반정보, 질병력, 식·생활습관 등)과 검사항목(신체계측, 임상검사 등) 변수가 포함되어 있으며, 약 천여 개의 변수들로 구성되어 있다.

표 6 | KoGES 역학자료 공개 목록

2020년 9월 기준			
코호트명	기반/추적	조사년도	참가자수(명)
지역사회기반 (안산, 안성)	기반(1기)	'01 ~'02년	10,030
	1차 추적(2기)	'03 ~'04년	8,603
	2차 추적(3기)	'05 ~'06년	7,515
	3차 추적(4기)	'07 ~'08년	6,688
	4차 추적(5기)	'09 ~'10년	6,665
	5차 추적(6기)	'11 ~'12년	6,238
	6차 추적(7기)	'13 ~'14년	5,906
	7차 추적(8기)	'15 ~'16년	6,318
도시기반	기반	'04 ~'13년	173,208
	1차 추적(CAPI)	'12 ~'16년	65,616
	1차 추적(예비조사)	'07 ~'11년	4,606
농촌기반	기반	'05 ~'11년	28,337
	1차 추적	'07 ~'14년	12,463
	2차 추적	'08 ~'16년	11,399
	3차 추적	'11 ~'16년	6,423
	4차 추적	'14 ~'16년	1,449

코호트명		기반/추적	조사년도	참가자수(명)
쌍둥이 및 가족		기반조사	'05 ~'13년	3,202
		1~3차 추적	'08 ~'14년	2,030
국제협력 I (일본)		기반~3차 추적	'05 ~'13년	1,063
국내이주자(성인)		기반	'06 ~'11년	4,786
		1차 추적	'12 ~'14년	1,105
중규모		기반	'03 ~'04년	10,302
사상체질		기반	'04 ~'06년	3,062
통합자료	KoGES 기반조사	기반 (지역사회, 도시, 농촌)	'01 ~'13년	211,575
	KoGES 반복 추적조사	기반~7차 추적 (지역사회)	'01 ~'16년	10,030

2. 공개 자료 변수 목록(코드북)

2-1. 코드북 다운로드

공개 자료의 코호트별 기반/추적 조사자료에 포함된 변수 목록(코드북)은 질병관리청 또는 국립보건연구원 홈페이지에서 확인이 가능하며, 아래의 순서에 따라 코드북 다운로드가 가능하다.

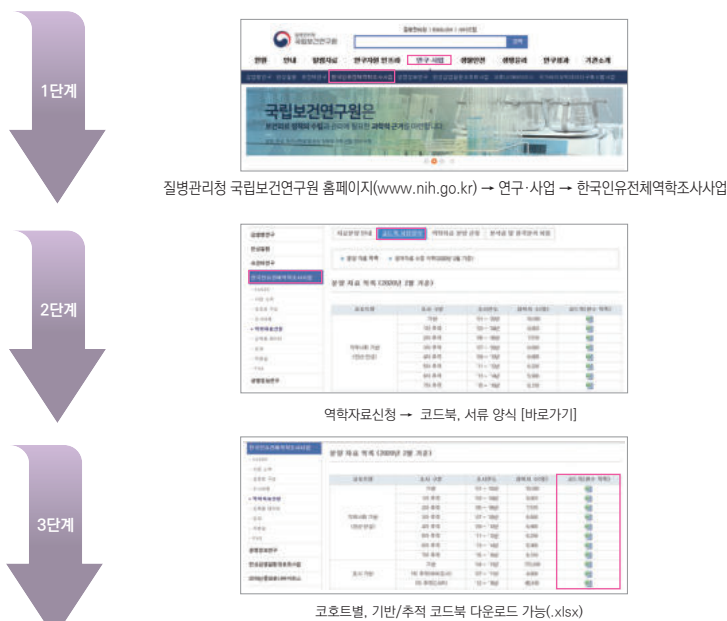


그림 6 | KoGES 조사 수행 체계

2-2. 코드북 구성

코드북에는 각 자료에 포함된 변수의 목록과 변수에 대한 부가적인 설명이 포함되어 있다. 코드북은 크게 변수 내용, 단위 데이터 현황, 설문지에 대한 부분으로 구성된다.

1 공개 여부

공개와 제한공개로 구분되며, 공개 변수란 홈페이지에서 자료를 다운로드하여 장소의 제약 없이 활용 가능한 변수를 의미하며, 제한공개 변수란 참여자의 정보 보호를 위해(응답비율이 낮거나 개인 식별의 우려가 있는 경우) 보안이 갖추어진 환경(유전체역학과 분석실)에서 분석하도록 제한적으로 공개하는 변수를 의미한다.

2 테이블명(국문, 영문)

KoGES 자료는 변수의 특성에 따라 여러 개의 파일(테이블)로 구분하여 제공된다. 각 테이블에는 참여자의 개인 식별 번호인 'ID' 변수(단, 분량자료 마다 개별 생성)가 포함되어 있으며, 필요에 따라 이를 기준으로 자료를 결합하여 사용할 수 있다.

도시거번코호트 기반(04-13년) 역학정보 공개데이터 통합코드북(ver 3.0)																																	설문지		비고																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
변수 내용													단위 데이터 현황													설문지	비고																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																											
영문 변수	공제 여부	데이터명 (국문)	데이터명 (영문)	변수명	변수설명	변수값(코드) 설명	변수유형	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39	Q40	Q41	Q42	Q43	Q44	Q45	Q46	Q47	Q48	Q49	Q50	Q51	Q52	Q53	Q54	Q55	Q56	Q57	Q58	Q59	Q60	Q61	Q62	Q63	Q64	Q65	Q66	Q67	Q68	Q69	Q70	Q71	Q72	Q73	Q74	Q75	Q76	Q77	Q78	Q79	Q80	Q81	Q82	Q83	Q84	Q85	Q86	Q87	Q88	Q89	Q90	Q91	Q92	Q93	Q94	Q95	Q96	Q97	Q98	Q99	Q100	Q101	Q102	Q103	Q104	Q105	Q106	Q107	Q108	Q109	Q110	Q111	Q112	Q113	Q114	Q115	Q116	Q117	Q118	Q119	Q120	Q121	Q122	Q123	Q124	Q125	Q126	Q127	Q128	Q129	Q130	Q131	Q132	Q133	Q134	Q135	Q136	Q137	Q138	Q139	Q140	Q141	Q142	Q143	Q144	Q145	Q146	Q147	Q148	Q149	Q150	Q151	Q152	Q153	Q154	Q155	Q156	Q157	Q158	Q159	Q160	Q161	Q162	Q163	Q164	Q165	Q166	Q167	Q168	Q169	Q170	Q171	Q172	Q173	Q174	Q175	Q176	Q177	Q178	Q179	Q180	Q181	Q182	Q183	Q184	Q185	Q186	Q187	Q188	Q189	Q190	Q191	Q192	Q193	Q194	Q195	Q196	Q197	Q198	Q199	Q200	Q201	Q202	Q203	Q204	Q205	Q206	Q207	Q208	Q209	Q210	Q211	Q212	Q213	Q214	Q215	Q216	Q217	Q218	Q219	Q220	Q221	Q222	Q223	Q224	Q225	Q226	Q227	Q228	Q229	Q230	Q231	Q232	Q233	Q234	Q235	Q236	Q237	Q238	Q239	Q240	Q241	Q242	Q243	Q244	Q245	Q246	Q247	Q248	Q249	Q250	Q251	Q252	Q253	Q254	Q255	Q256	Q257	Q258	Q259	Q260	Q261	Q262	Q263	Q264	Q265	Q266	Q267	Q268	Q269	Q270	Q271	Q272	Q273	Q274	Q275	Q276	Q277	Q278	Q279	Q280	Q281	Q282	Q283	Q284	Q285	Q286	Q287	Q288	Q289	Q290	Q291	Q292	Q293	Q294	Q295	Q296	Q297	Q298	Q299	Q300	Q301	Q302	Q303	Q304	Q305	Q306	Q307	Q308	Q309	Q310	Q311	Q312	Q313	Q314	Q315	Q316	Q317	Q318	Q319	Q320	Q321	Q322	Q323	Q324	Q325	Q326	Q327	Q328	Q329	Q330	Q331	Q332	Q333	Q334	Q335	Q336	Q337	Q338	Q339	Q340	Q341	Q342	Q343	Q344	Q345	Q346	Q347	Q348	Q349	Q350	Q351	Q352	Q353	Q354	Q355	Q356	Q357	Q358	Q359	Q360	Q361	Q362	Q363	Q364	Q365	Q366	Q367	Q368	Q369	Q370	Q371	Q372	Q373	Q374	Q375	Q376	Q377	Q378	Q379	Q380	Q381	Q382	Q383	Q384	Q385	Q386	Q387	Q388	Q389	Q390	Q391	Q392	Q393	Q394	Q395	Q396	Q397	Q398	Q399	Q400	Q401	Q402	Q403	Q404	Q405	Q406	Q407	Q408	Q409	Q410	Q411	Q412	Q413	Q414	Q415	Q416	Q417	Q418	Q419	Q420	Q421	Q422	Q423	Q424	Q425	Q426	Q427	Q428	Q429	Q430	Q431	Q432	Q433	Q434	Q435	Q436	Q437	Q438	Q439	Q440	Q441	Q442	Q443	Q444	Q445	Q446	Q447	Q448	Q449	Q450	Q451	Q452	Q453	Q454	Q455	Q456	Q457	Q458	Q459	Q460	Q461	Q462	Q463	Q464	Q465	Q466	Q467	Q468	Q469	Q470	Q471	Q472	Q473	Q474	Q475	Q476	Q477	Q478	Q479	Q480	Q481	Q482	Q483	Q484	Q485	Q486	Q487	Q488	Q489	Q490	Q491	Q492	Q493	Q494	Q495	Q496	Q497	Q498	Q499	Q500	Q501	Q502	Q503	Q504	Q505	Q506	Q507	Q508	Q509	Q510	Q511	Q512	Q513	Q514	Q515	Q516	Q517	Q518	Q519	Q520	Q521	Q522	Q523	Q524	Q525	Q526	Q527	Q528	Q529	Q530	Q531	Q532	Q533	Q534	Q535	Q536	Q537	Q538	Q539	Q540	Q541	Q542	Q543	Q544	Q545	Q546	Q547	Q548	Q549	Q550	Q551	Q552	Q553	Q554	Q555	Q556	Q557	Q558	Q559	Q560	Q561	Q562	Q563	Q564	Q565	Q566	Q567	Q568	Q569	Q570	Q571	Q572	Q573	Q574	Q575	Q576	Q577	Q578	Q579	Q580	Q581	Q582	Q583	Q584	Q585	Q586	Q587	Q588	Q589	Q590	Q591	Q592	Q593	Q594	Q595	Q596	Q597	Q598	Q599	Q600	Q601	Q602	Q603	Q604	Q605	Q606	Q607	Q608	Q609	Q610	Q611	Q612	Q613	Q614	Q615	Q616	Q617	Q618	Q619	Q620	Q621	Q622	Q623	Q624	Q625	Q626	Q627	Q628	Q629	Q630	Q631	Q632	Q633	Q634	Q635	Q636	Q637	Q638	Q639	Q640	Q641	Q642	Q643	Q644	Q645	Q646	Q647	Q648	Q649	Q650	Q651	Q652	Q653	Q654	Q655	Q656	Q657	Q658	Q659	Q660	Q661	Q662	Q663	Q664	Q665	Q666	Q667	Q668	Q669	Q670	Q671	Q672	Q673	Q674	Q675	Q676	Q677	Q678	Q679	Q680	Q681	Q682	Q683	Q684	Q685	Q686	Q687	Q688	Q689	Q690	Q691	Q692	Q693	Q694	Q695	Q696	Q697	Q698	Q699	Q700	Q701	Q702	Q703	Q704	Q705	Q706	Q707	Q708	Q709	Q710	Q711	Q712	Q713	Q714	Q715	Q716	Q717	Q718	Q719	Q720	Q721	Q722	Q723	Q724	Q725	Q726	Q727	Q728	Q729	Q730	Q731	Q732	Q733	Q734	Q735	Q736	Q737	Q738	Q739	Q740	Q741	Q742	Q743	Q744	Q745	Q746	Q747	Q748	Q749	Q750	Q751	Q752	Q753	Q754	Q755	Q756	Q757	Q758	Q759	Q760	Q761	Q762	Q763	Q764	Q765	Q766	Q767	Q768	Q769	Q770	Q771	Q772	Q773	Q774	Q775	Q776	Q777	Q778	Q779	Q780	Q781	Q782	Q783	Q784	Q785	Q786	Q787	Q788	Q789	Q790	Q791	Q792	Q793	Q794	Q795	Q796	Q797	Q798	Q799	Q800	Q801	Q802	Q803	Q804	Q805	Q806	Q807	Q808	Q809	Q810	Q811	Q812	Q813	Q814	Q815	Q816	Q817	Q818	Q819	Q820	Q821	Q822	Q823	Q824	Q825	Q826	Q827	Q828	Q829	Q830	Q831	Q832	Q833	Q834	Q835	Q836	Q837	Q838	Q839	Q840	Q841	Q842	Q843	Q844	Q845	Q846	Q847	Q848	Q849	Q850	Q851	Q852	Q853	Q854	Q855	Q856	Q857	Q858	Q859	Q860	Q861	Q862	Q863	Q864	Q865	Q866	Q867	Q868	Q869	Q870	Q871	Q872	Q873	Q874	Q875	Q876	Q877	Q878	Q879	Q880	Q881	Q882	Q883	Q884	Q885	Q886	Q887	Q888	Q889	Q890	Q891	Q892	Q893	Q894	Q895	Q896	Q897	Q898	Q899	Q900	Q901	Q902	Q903	Q904	Q905	Q906	Q907	Q908	Q909	Q910	Q911	Q912	Q913	Q914	Q915	Q916	Q917	Q918	Q919	Q920	Q921	Q922	Q923	Q924	Q925	Q926	Q927	Q928	Q929	Q930	Q931	Q932	Q933	Q934	Q935	Q936	Q937	Q938	Q939	Q940	Q941	Q942	Q943	Q944	Q945	Q946	Q947	Q948	Q949	Q950	Q951	Q952	Q953	Q954	Q955	Q956	Q957	Q958	Q959	Q960	Q961	Q962	Q963	Q964	Q965	Q966	Q967	Q968	Q969	Q970	Q971	Q972	Q973	Q974	Q975	Q976	Q977	Q978	Q979	Q980	Q981	Q982	Q983	Q984	Q985	Q986	Q987	Q988	Q989	Q990	Q991	Q992	Q993	Q994	Q995	Q996	Q997	Q998	Q999	Q1000	Q1001	Q1002	Q1003	Q1004	Q1005	Q1006	Q1007	Q1008	Q1009	Q1010	Q1011	Q1012	Q1013	Q1014	Q1015	Q1016	Q1017	Q1018	Q1019	Q1020	Q1021	Q1022	Q1023	Q1024	Q1025	Q1026	Q1027	Q1028	Q1029	Q1030	Q1031	Q1032	Q1033	Q1034	Q1035	Q1036	Q1037	Q1038	Q1039	Q1040	Q1041	Q1042	Q1043	Q1044	Q1045	Q1046	Q1047	Q1048	Q1049	Q1050	Q1051	Q1052	Q1053	Q1054	Q1055	Q1056	Q1057	Q1058	Q1059	Q1060	Q1061	Q1062	Q1063	Q1064	Q1065	Q1066	Q1067	Q1068	Q1069	Q1070	Q1071	Q1072	Q1073	Q1074	Q1075	Q1076	Q1077	Q1078	Q1079	Q1080	Q1081	Q1082	Q1083	Q1084	Q1085	Q1086	Q1087	Q1088	Q1089	Q1090	Q1091	Q1092	Q1093	Q1094	Q1095	Q1096	Q1097	Q1098	Q1099	Q1100	Q1101	Q1102	Q1103	Q1104	Q1105	Q1106	Q1107	Q1108	Q1109	Q1110	Q1111	Q1112	Q1113	Q1114	Q1115	Q1116	Q1117	Q1118	Q1119	Q1120	Q1121	Q1122	Q1123	Q1124	Q1125	Q1126	Q1127	Q1128	Q1129	Q1130	Q1131	Q1132	Q1133	Q1134	Q1135	Q1136	Q1137	Q1138	Q1139	Q1140	Q1141	Q1142	Q1143	Q1144	Q1145	Q1146	Q1147	Q1148	Q1149	Q1150	Q1151	Q1152	Q1153	Q1154	Q1155	Q1156	Q1157	Q1158	Q1159	Q1160	Q1161	Q1162	Q1163	Q1164	Q1165	Q1166	Q1167	Q1168	Q1169	Q1170	Q1171	Q1172	Q1173	Q1174	Q1175	Q1176	Q1177	Q1178	Q1179	Q1180	Q1181	Q1182	Q1183	Q1184	Q1185	Q1186	Q1187	Q1188	Q1189	Q1190	Q1191	Q1192	Q1193	Q1194	Q1195	Q1196	Q1197	Q1198	Q1199	Q1200	Q1201	Q1202	Q1203	Q1204	Q1205	Q1206	Q1207	Q1208	Q1209	Q1210	Q1211	Q1212	Q1213	Q1214	Q1215	Q1216	Q1217	Q1218	Q1219	Q1220	Q1221	Q1222	Q1223	Q1224	Q1225	Q1226	Q1227	Q1228	Q1229	Q1230	Q1231	Q1232	Q1233	Q1234	Q1235	Q1236	Q1237	Q1238	Q1239	Q1240	Q1241	Q1242	Q1243	Q1244	Q1245	Q1246	Q1247	Q1248	Q1249	Q1250	Q1251	Q1252	Q1253	Q1254	Q1255	Q1256	Q1257	Q1258	Q1259	Q1260	Q1261	Q1262	Q1263	Q1264	Q1265	Q1266	Q1267	Q1268	Q1269	Q1270	Q1271	Q1272	Q1273	Q1274	Q1275	Q1276	Q1277	Q1278	Q1279	Q1280	Q1281	Q1282	Q1283	Q1284	Q1285	Q1286	Q1287	Q1288	Q1289	Q1290	Q1291	Q1292	Q1293	Q1294	Q1295	Q1296	Q1297	Q1298	Q1299	Q1300	Q1301	Q1302	Q1303	Q1304	Q1305	Q1306	Q1307	Q1308	Q1309	Q1310	Q1311	Q1312	Q1313	Q1314	Q1315	Q1316	Q1317	Q1318	Q1319	Q1320	Q1321	Q1322	Q1323	Q1324	Q1325	Q1326	Q1327

③ 단위 데이터 현황

KoGES 자료는 장기간에 걸쳐 여러 지역에서 자료가 수집되어, 사업연도 혹은 조사기관에 따라 설문 문항에 차이가 있는 경우가 존재한다. 이에 대한 구분자가 단위 데이터이며, 각 단위 데이터의 변수 조사 유무는 3가지 기호로 구분되어 있다. 「●」는 해당 항목을 조사하여, 데이터가 존재할 경우이며, 「○」는 해당 항목을 조사하지 않았으나, 데이터 가공을 통하여 추가 생성한 경우(가공 변수)이며, 「X」는 해당 항목을 조사하지 않아, 데이터가 없을 경우이다.

[illegible]

그림 8 | KoGES 코드북 - 단위 데이터 현황

④ 설문지





설문 내용만으로 문항 간의 상·하위 관계를 유추하고, 관련 변수를 쉽게 찾을 수 있도록 조사 당시 설문지 내용을 제시하고 있다.

[illegible]

그림 9 | KoGES 코드북 - 설문지

3. 분양 자료 파일 유형 및 형태

KoGES 역학자료는 연구자의 요청에 따라 CSV, ACCESS, SAS DATA SET, EXCEL 파일 유형으로도 분양이 가능하며, 각 파일 유형의 특징은 아래와 같다.

파일 유형	특징
 CSV	용량이 작고, 어떤 통계분석 패키지에도 쉽게 호환이 가능하며, 특히 R에서는 별도의 패키지를 설치하지 않고도, 쉽게 불러오고 저장할 수 있음
 ACCESS	데이터베이스 형식으로 손상 없이 자료를 불러올 수 있음
 SAS DATA SET	통계 프로그램 SAS를 이용할 경우, 별도의 추가 작업 없이 분석 데이터 셋으로 생성 가능하며, R로 데이터를 불러오기 위해서는 특정 패키지 설치가 필요함
 EXCEL	가장 쉽게 접할 수 있는 파일 형식이나, CSV 파일 형식에 비해서 용량이 다소 크며, R로 데이터를 불러오기 위해서는 특정 패키지 설치가 필요함

| 그림 10 | KoGES 분양자료 파일 유형 및 특징

앞서 언급한 것과 같이, KoGES 역학자료는 변수의 특성에 따라 여러 개의 테이블로 구분되어 있다. 예를 들어 지 역사회기반 코호트 기반조사 자료 중 성별과 나이, 음주여부, 흡연여부, 공복혈당, HDL-콜레스테롤 변수를 분 양 신청하였을 경우, 성별과 나이는 기본정보 테이블에, 음주여부와 흡연여부는 생활습관 테이블에, 공복혈당과 HDL-콜레스테롤은 임상검사 테이블에 포함되어, 6개의 변수가 3개의 테이블에 나뉘어 제공되며, 1~n차 추적조 사 자료 역시 변수 특성에 따라 여러 개의 테이블로 나뉘어 제공된다.



| 그림 11 | KoGES 분양자료 형태

한국인유전체역학조사사업 (KoGES)
Korean Genome and Epidemiology Study

KoGES 데이터 분석 가이드북

R편

KoGES 데이터 분석 가이드북
[R편]

Korean Genome and Epidemiology Study

3장.

한국인유전체역학조사사업(KoGES) 기반조사 자료 분석하기

1. 기반조사 교육용 데이터 이해하기
2. 자료 불러오기
3. 자료 결합하기
4. 자료 분석 준비하기
5. 자료 분석하기

3장.

한국인유전체역학조사사업(KoGES) 기반조사 자료 분석하기

1. 기반조사 교육용 데이터 이해하기

1-1. 자료 구성

KoGES 기반조사 교육용 데이터는 대상자 수 10,000명, 3개 테이블, 58개 변수로 구성되어 있으며 자세한 내용은 다음과 같다.

테이블1(base_data1)	테이블2(base_data2)	테이블3(base_data3)
기본정보 및 일반정보 단위데이터, 조사일자, 성별, 만 나이, 월 평균 수입, 결혼 상태	질병 과거력 고혈압/당뇨병/고지혈증 진단여부, 처음 진단받은 나이 가족력 고혈압/당뇨병 진단여부(부, 모) 생활습관 음주 여부, 총 음주 기간, 주류별 평균 음주 횟수 및 1회 음주량, 흡연 여부, 총 흡연 기간, 하루 흡연량, 간접흡연 여부, 규칙적 운동 여부 여성력 초경나이, 폐경 여부, 폐경나이, 임신 경험 여부, 첫 임신 나이	신체계측 SBP, DBP, 맥박수, 허리둘레, 엉덩이둘레, 신장, 체중, BMI 임상검사 HbA1c, Glucose, Creatinine, AST, ALT, Total cholesterol, HDL cholesterol, LDL cholesterol, Triglyceride

그림 12 | KoGES 기반조사 교육용 데이터 구성

1-2. 코드북

KoGES 기반조사 교육용 데이터의 코드북은 공개 자료 코드북과 동일하게 테이블명, 변수명, 변수 설명, 변수값 (코드) 설명, 변수타입, 통합 설문지로 구성되어있다.

KoGES 기반조사 교육용데이터 코드북						
통합 변수				통합 설문지		
NO.	데이터블록명(영문)	데이터블록명	변수명	변수 설명	변수 타입	통합 설문지
1	10000_00001	기본정보	1_01	성별(성: 100001)	문자	1. 성별(성: 100001)
2	10000_00001	기본정보	1_02	나이(나이: 100002)	문자	2. 나이(나이: 100002)
3	10000_00001	기본정보	1_03	직업(직업: 100003)	문자	3. 직업(직업: 100003)
4	10000_00001	기본정보	1_04	학력(학력: 100004)	문자	4. 학력(학력: 100004)
5	10000_00001	기본정보	1_05	연소득(연소득: 100005)	문자	5. 연소득(연소득: 100005)
6	10000_00001	기본정보	1_06	연소득(연소득: 100006)	문자	6. 연소득(연소득: 100006)
7	10000_00001	기본정보	1_07	연소득(연소득: 100007)	문자	7. 연소득(연소득: 100007)
8	10000_00001	기본정보	1_08	연소득(연소득: 100008)	문자	8. 연소득(연소득: 100008)
9	10000_00001	기본정보	1_09	연소득(연소득: 100009)	문자	9. 연소득(연소득: 100009)
10	10000_00001	기본정보	1_10	연소득(연소득: 100010)	문자	10. 연소득(연소득: 100010)
11	10000_00001	기본정보	1_11	연소득(연소득: 100011)	문자	11. 연소득(연소득: 100011)
12	10000_00001	기본정보	1_12	연소득(연소득: 100012)	문자	12. 연소득(연소득: 100012)
13	10000_00001	기본정보	1_13	연소득(연소득: 100013)	문자	13. 연소득(연소득: 100013)
14	10000_00001	기본정보	1_14	연소득(연소득: 100014)	문자	14. 연소득(연소득: 100014)
15	10000_00001	기본정보	1_15	연소득(연소득: 100015)	문자	15. 연소득(연소득: 100015)
16	10000_00001	기본정보	1_16	연소득(연소득: 100016)	문자	16. 연소득(연소득: 100016)
17	10000_00001	기본정보	1_17	연소득(연소득: 100017)	문자	17. 연소득(연소득: 100017)
18	10000_00001	기본정보	1_18	연소득(연소득: 100018)	문자	18. 연소득(연소득: 100018)
19	10000_00001	기본정보	1_19	연소득(연소득: 100019)	문자	19. 연소득(연소득: 100019)
20	10000_00001	기본정보	1_20	연소득(연소득: 100020)	문자	20. 연소득(연소득: 100020)

그림 13 | KoGES 기반조사 교육용데이터 코드북

2. 자료 불러오기

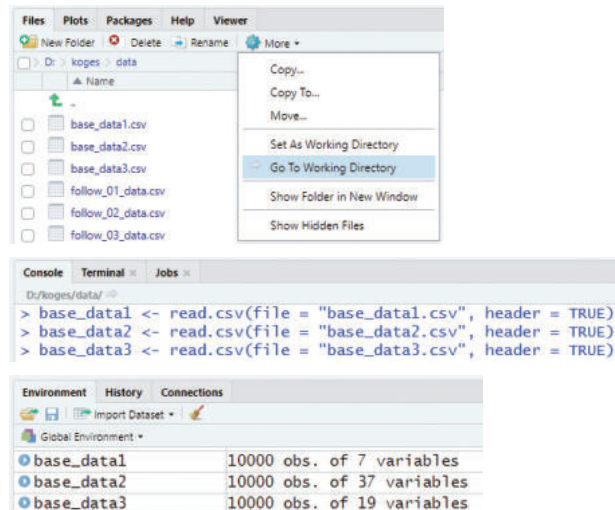
2-1. 자료 불러오기

파일을 불러오기 전, RStudio에서 생성한 데이터 및 분석 결과를 저장하거나 분석할 데이터 파일을 불러올 때 사용하는 폴더를 원하는 경로에 생성하고 이를 워킹 디렉토리로 설정해준다. 현재 설정된 워킹 디렉토리를 알고 싶다면 `getwd()` 함수를 이용하면 되고, 새롭게 설정하고 싶다면 `setwd()` 함수를 이용하면 된다. 본 가이드북에서 사용하는 KoGES 교육용 데이터는 [koges]라는 폴더 하위에 있는 [data] 폴더에 CSV파일 형태로 저장되어 있다. 예를 들어 [koges] 폴더 전체를 D 드라이브에 저장했다면, 데이터가 들어있는 파일의 경로는 "D:/koges/data" 가 되며, `setwd()` 함수를 이용해 다음의 코드 `setwd("D:/koges/data")`를 실행하면 워킹 디렉토리가 지정된다. 이처럼 워킹 디렉토리가 지정되었다면, 그 다음으로 `read.csv()` 함수를 이용해 바로 CSV파일을 R로 불러올 수 있다.

```
# 워킹 디렉토리 ----
getwd()           # 워킹 디렉토리 확인
setwd("D:/koges/data") # 워킹 디렉토리 설정

# CSV 파일 불러오기 ----
base_data1 <- read.csv(file = "base_data1.csv", header = TRUE)
base_data2 <- read.csv(file = "base_data2.csv", header = TRUE)
base_data3 <- read.csv(file = "base_data3.csv", header = TRUE)
```

결과



변수 관련

• 변수 생성 방법

```
x1 <- 1 # '1이라는 데이터를 변수 x1에 할당한다' 라는 의미
x1
x2 <- c(1, 2, 3, 4, 5)
x2
x3 <- 1:5 # :(콜론)은 숫자가 1씩 증가할 때 사용 가능
x3
x4 <- seq(from = 1, to = 5, by = 1) # seq() 함수는 일정한 간격일 때 사용 가능
x4
x5 <- "KoGES" # 데이터가 문자형인 경우 큰따옴표(") 사용
x5
```

• 변수명 규칙

첫 글자는 반드시 영문 또는 마침표(.)
두 번째 글자부터는 영문, 숫자, 밑줄(_) 사용 가능
변수명 중간에 빈칸을 넣을 수 없음
※ R은 대문자와 소문자를 구별함.

연산자 종류

• 할당 연산자 : 객체(변수, 데이터, 함수)의 이름에 특정 값이나 분석 결과를 저장할 때 사용

<- : 오른쪽의 값을 왼쪽의 이름에 저장
-> : 왼쪽의 값을 오른쪽의 이름에 저장
= : 함수의 인수(argument)를 지정할 때 사용

• 산술 연산자 : 수치에 대한 연산을 하기 위해 사용

더하기(+), 빼기(-), 곱하기(*), 나누기(/), 거듭제곱(**, ^), 몫(%/%), 나머지(%%)

• 비교 연산자 : 값을 비교하여 맞으면 TRUE, 맞지 않으면 FALSE를 반환할 때 사용

크다(>), 작다(<), 크거나 같다(>=), 작거나 같다(<=), 같다(==), 같지 않다(!=), 아니다(!)

• 논리 연산자

AND(&), OR(|)



txt 파일 불러오기

• Usage

```
read.table(<Argument>)
```

• Argument

file = ""	파일의 위치(디렉토리)와 파일명(확장자 포함) 지정※
header = TRUE	첫행이 변수명인지 확인하고 그대로 가져오기
sep = "\t"	구분자 형태 지정, 디폴트는 탭("\t")이며 그 외 공백(" "), 콤마(",",) 등
skip = n	n행까지 제외하고 불러오기
nrows = n	n행까지 불러오기
stringsAsFactors = TRUE	문자형 데이터를 factor로 자동 변경

※ [주의] R에서 디렉토리를 지정할 때 순방향(/) 또는 역방향(\)을 사용함.

excel 파일 불러오기

• Usage

```
readxl::read_excel(<Argument>)*
```

※ read_excel() 함수를 사용하기 위해서는 readxl 패키지의 설치와 로딩이 필요함.

• Argument

path = ""	파일의 위치(디렉토리)와 파일명(확장자 포함) 지정
col_names = TRUE	첫행이 변수명인지 확인하고 그대로 가져오기
sheet = n	n번째 시트를 불러오기

sas 파일 불러오기

• Usage

```
sas7bdat::read.sas7bdat(<Argument>)*
```

※ read.sas7bdat 함수를 사용하기 위해서는 sas7bdat 패키지의 설치와 로딩이 필요함.

• Argument

file = ""	파일의 위치(디렉토리)와 파일명(확장자 포함) 지정
to.data.frame = TRUE	데이터구조를 데이터프레임 형태로 가져오기

2-2. 불러온 자료 확인하기

데이터를 불러와서 가장 먼저 해야할 일은 불러온 데이터 형태와 관측값 수, 변수에 대한 정보(이름, 유형, 값) 등을 확인하는 일이다. R에서 제공하는 다양한 함수들을 이용해 불러온 자료의 정보를 확인할 수 있으며, 자주 사용하는 함수는 다음과 같다.

표 7 | 불러온 자료 확인할 때 유용한 함수

함 수	기 능
str(dataset)	데이터셋 구조(데이터 형태, 관측값(행) 수와 변수(열) 수, 변수별 속성) 확인
class(dataset)	데이터 형태 / 변수별 속성 확인
dim(dataset)	관측값(행) 수와 변수(열) 수 확인
ls(dataset)	데이터셋에 포함된 변수명 확인
View(dataset)	View 창이 별도로 열리며, 엑셀의 시트 형태처럼 되어 있는 데이터셋 전체를 직접 확인하고, 아이콘 클릭을 통해 필터나 정렬 실행할 수 있음
head(dataset, <옵션>)	데이터셋 앞부분 6행만 출력, <옵션> n=10 (앞에서부터 10행까지 출력)
tail(dataset, <옵션>)	데이터셋 뒷부분 6행만 출력, <옵션> n=10 (뒤에서부터 10행까지 출력)

불러온 자료 확인하기 ----

```
str(base_data1) # 데이터셋 구조 확인
View(base_data1) # 데이터셋 전체 새로운 창으로 보기
head(base_data1) # 데이터셋 앞부분 확인
```

결과

```
> str(base_data1) # 데이터셋 구조 확인
'data.frame': 10000 obs. of 7 variables:
 $ t_id      : Factor w/ 10000 levels "K_BASE_00001",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ t_data_class: Factor w/ 33 levels "B01","B02","B03",...: 20 20 14 28 23 30 26 2 22 26 ...
 $ t_edate    : int  200802 200306 201108 201004 200410 201110 201304 201003 200807 200806 ...
 $ t_sex      : int  2 2 2 2 1 2 1 2 1 2 ...
 $ t_age      : int  42 58 60 73 59 54 41 58 51 56 ...
 $ t_income   : int  99999 1 99999 1 2 2 5 3 4 6 ...
 $ t_marry    : int  99999 2 2 5 2 2 2 2 2 2 ...
```

	t_id	t_data_class	t_edate	t_sex	t_age	t_income	t_marry
1	K_BASE_00001	B20	200802	2	42	99999	99999
2	K_BASE_00002	B20	200306	2	58	1	2
3	K_BASE_00003	B14	201108	2	60	99999	2
4	K_BASE_00004	B29	201004	2	73	1	5
5	K_BASE_00005	B23	200410	2	59	2	2

```
> head(base_data1) # 데이터셋 앞부분 확인
  t_id t_data_class t_edate t_sex t_age t_income t_marry
1 K_BASE_00001      B20 200802     2    42   99999   99999
2 K_BASE_00002      B20 200306     2    58     1     2
3 K_BASE_00003      B14 201108     2    60   99999     2
4 K_BASE_00004      B29 201004     2    73     1     5
5 K_BASE_00005      B23 200410     2    59     2     2
6 K_BASE_00006      B31 201110     1    54     2     2
```



데이터 유형

데이터 유형	내용
숫자형	기본적으로 숫자로 되어있는 데이터를 말하고 보통의 표기는 숫자형(numeric)이지만, 정수형(integer), 실수형(double)로도 구분 가능함
문자형	하나의 문자 또는 문자열로 되어있는 데이터를 말하고, 문자형으로 인식하기 위해서는 큰 따옴표("") 또는 작은 따옴표('')로 묶어 주어야 함
논리형	TRUE(참) 또는 FALSE(거짓)으로 이루어진 데이터를 말하고, 주로 데이터값을 비교할 때 사용됨

데이터 형태

- R에서의 객체 중 대표적인 데이터 객체 종류에 대해서 간단하게 소개하고자 한다.

형태	설명	표현 (예)
벡터 (vector)	데이터 분석의 가장 기본 단위로, 한 개 이상의 값(element)과 하나의 열로 되어 있음	v1 <- c(1, 2, 3, 4, 5) v2 <- 1:5 v3 <- seq(from = 1, to = 5, by = 1)
요인 (factor)	벡터와 동일하게 데이터 분석의 가장 기본 단위로, 집단별 데이터 분석을 위해 범주형 자료로 변환해주는 기능을 하며, 변환하면 집단의 순서(levels)의 정보가 제공됨	sex <- c(1, 2, 2, 1) sex_f <- factor(sex, levels = c("1", "2"), labels = c("남자", "여자"))
데이터프레임 (data.frame)	행과 열로 구성된 2차원 구조로 되어 있으며, 하나의 열에는 동일한 데이터 유형을 가지며, 열마다 데이터 유형은 다를 수 있음. R에서의 데이터는 기본적으로 데이터 프레임에 의미함	id <- c(k1, k2, k3, k4, k5) age <- c(10, 20, 30, 40, 50) df1 <- data.frame(id, age)
리스트 (list)	대부분의 R 분석 결과물은 리스트 형태이며, 구성된 값(element)의 사이즈가 달라도 됨	list1 <- list(v1, sex_f, df1)

3. 자료 결합하기

KoGES 역학자료는 변수의 특성에 따라 여러 개의 테이블로 구분되어 있다. 따라서 자료 분석을 위해서는 연구 목적에 따라 하나의 데이터 셋으로 결합(merge)하는 작업이 요구된다.



각각의 개별 테이블에는 공통적으로 참여자의 개인 식별번호인 'ID' 변수가 포함되어 있으며, 이를 기준으로 하나의 자료로 결합할 수 있다. 보통의 통계패키지에서는 결합 전 반드시 'ID' 변수로 정렬하는 작업이 필요하나, R에서는 사전 정렬 없이 바로 테이블 결합이 가능하다.

테이블1 - 기본정보			테이블2 - 생활습관			기본정보 + 생활습관				
ID	SEX	AGE	ID	DRINK	SMOKE	ID	SEX	AGE	DRINK	SMOKE
NO_1	1	55	NO_1	1	2	NO_1	1	55	1	2
NO_2	2	60	NO_2	2	2	NO_2	2	60	2	2
NO_3	2	58	NO_3	2	1	NO_3	2	58	2	1
NO_4	1	63	NO_4	1	1	NO_4	1	63	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

| 그림 14 | KoGES 기반기조 데이터 조인키

예제

기반조사 교육용 데이터를 구성하는 3개 데이터 셋(base_data1, base_data2, base_data3)에 포함된 여러 변수 중 필요한 변수를 선택하여 하나의 데이터 셋(base_data)으로 결합하기

t_id	t_data_class	t_edate	t_sex	t_age	t_id	t_data_class	t_htn	t_htnag	t_id	t_data_class	t_sbp	t_dbp
1	K_BASE_00001	B20	200802	2	42	1	K_BASE_00001	B20	1	77777	115	78
2	K_BASE_00002	B20	200806	2	58	2	K_BASE_00002	B20	2	50	143	87
3	K_BASE_00003	B14	201108	2	60	3	K_BASE_00003	B14	2	50	120	80
4	K_BASE_00004	B29	201004	2	73	4	K_BASE_00004	B29	2	63	104	62
5	K_BASE_00005	B23	200410	2	59	5	K_BASE_00005	B23	1	77777	125	81
6	K_BASE_00006	B31	201110	1	54	6	K_BASE_00006	B31	2	50	130	80
7	K_BASE_00007	B27	201304	2	41	7	K_BASE_00007	B27	1	77777	138	86
8	K_BASE_00008	B02	201003	1	58	8	K_BASE_00008	B02	2	57	137	94
9	K_BASE_00009	B22	200807	2	51	9	K_BASE_00009	B22	2	47	130	83
10	K_BASE_00010	B27	200806	1	56	10	K_BASE_00010	B27	1	77777	128	75

그림 15 | KoGES 기반조사 교육용 데이터 셋

```
# 각 데이터셋에서 필요한 변수만 가져오기 ----
install.packages("dplyr") # 패키지 설치
library(dplyr)           # 패키지 로드
select1 <- base_data1 %>% dplyr::select(t_id, t_sex, t_age)
select2 <- base_data2 %>% dplyr::select(t_id, t_htn, t_dm, t_drink, t_smoke)
select3 <- base_data3 %>% dplyr::select(t_id, t_weight, t_height, t_bmi, t_glu0)
```

```
# 자료 결합 ----
merge1 <- dplyr::left_join(select1, select2)
merge2 <- dplyr::left_join(merge1, select3)
base_data <- merge2
```

```
# 결합된 데이터 확인 ----
str(base_data)
base_data
```

결과

```
> str(base_data)
'data.frame': 10000 obs. of 11 variables:
 $ t_id : Factor w/ 10000 levels "K_BASE_00001",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ t_sex : int 2 2 2 2 2 1 2 1 2 1 ...
 $ t_age : int 42 58 60 73 59 54 41 58 51 56 ...
 $ t_htn : int 1 2 2 2 1 2 1 2 2 1 ...
 $ t_dm : int 1 1 1 1 1 1 1 1 1 1 ...
 $ t_drink : int 1 2 1 1 1 1 3 3 2 1 ...
 $ t_smoke : int 1 1 1 1 1 3 1 1 2 2 ...
 $ t_weight : int 64 64 54 55 60 81 58 72 75 68 ...
 $ t_height : int 164 156 154 143 150 169 168 165 154 170 ...
 $ t_bmi : int 24 26 23 27 27 28 21 26 32 24 ...
 $ t_glu0 : int 68 86 86 108 91 90 86 91 97 87 ...

> base_data
  t_id t_sex t_age t_htn t_dm t_drink t_smoke t_weight t_height t_bmi t_glu0
1 K_BASE_00001 2 42 1 1 1 1 64 164 24 68
2 K_BASE_00002 2 58 2 1 2 1 64 156 26 86
3 K_BASE_00003 2 60 2 1 1 1 54 154 23 86
4 K_BASE_00004 2 73 2 1 1 1 55 143 27 108
5 K_BASE_00005 2 59 1 1 1 1 60 150 27 91
6 K_BASE_00006 1 54 2 1 1 3 81 169 28 90
7 K_BASE_00007 2 41 1 1 3 1 58 168 21 86
8 K_BASE_00008 1 58 2 1 3 1 72 165 26 91
9 K_BASE_00009 2 51 2 1 2 2 75 154 32 97
10 K_BASE_00010 1 56 1 1 1 2 68 170 24 87
```

패키지와 함수

- R에는 데이터 값을 미리 정한 공식에 따라 처리하는 함수가 다양하게 있으며, 이러한 함수를 모아놓은 꾸러미를 패키지라고 한다. 패키지는 R 설치시 자동으로 설치되어 기본적인 통계분석, 그래프 작성, 데이터 처리 등 즉시 사용 가능한 패키지와 자동으로 설치되는 되었지만 사용하려면 R로 불러와야하는 패키지, 그리고 통계분석 목적 및 필요에 따라 따로 설치가 필요한 패키지로 구분된다. 마지막 패키지 종류의 경우 패키지 사용을 위해서는 먼저 특정 패키지를 설치(install)하고, 설치한 패키지를 현재의 작업환경으로 로딩(load)하는 과정을 거쳐야 한다.

```
install.packages("패키지명") # 패키지 설치
library(패키지명)           # 패키지 로드
```

- R에는 많은 종류의 패키지가 있는데, 만약 어떤 패키지를 사용해야 할지 모르는 경우 분야별 패키지 목록이 정리되어 있는 CRAN 웹 사이트(<http://cran.r-project.org/web/views>)를 참고하면 된다.
- 설치한 패키지 확인 : `installed.packages()`
- 설치한 패키지 업데이트 : `update.packages("패키지명")`
- 컴퓨터간 동일한 패키지 환경 설정하기 : library 파일을 복사한 후 기존 설치된 library 폴더와 대체
- 특정 패키지내 데이터셋 이용하기 : `data(package="패키지명")`
- 특정 패키지내 함수 지정

함수를 사용하다 보면 오류 메시지가 뜰때가 있는데, 이는 다른 패키지의 동일한 이름을 가진 함수 사용으로 인해 발생한 것으로, 사용하고자 하는 함수가 속한 패키지를 명시적으로 지정해주면 해결이 된다. 예를 들어 자주 사용하는 함수 중 `filter()`라는 함수는 {stats}라는 패키지와 {dplyr} 패키지에서 동일한 이름으로 사용되는 함수이다. 따라서 dplyr 패키지의 `filter()` 함수를 사용하고자 한다면, `dplyr::filter()`로 해당 패키지를 지정해주는 방법을 추천한다.

- R에서 자주 사용하는 몇 가지의 패키지를 요약하면 다음과 같다.

구분	패키지	기능
데이터가공	dplyr	데이터 전처리 작업
	reshape2	데이터 레이아웃 변환 (long type ↔ wide type)
	psych	기술통계량 산출
시각화	ggplot2	자료의 시각화
	lattice	격자 그래프
모델링	stats	선형 모형(lm), 일반화 선형 모형(로지스틱, 포아송 회귀 포함, glm)
	survival	생존분석
통합분석	tidyverse	강력한 데이터 조작, 관리(시계열 포함), 시각화 분석 제공 tidyverse는 8개의 패키지가 묶여 있음 (ggplot2, tibble, tidyr, readr, purrr, dplyr, stringr, forcats)



dplyr 패키지 – 파이프 연산자

- 파이프 연산자(`%>%`)는 함수의 결과값을 별도로 저장하지 않고도 여러 개의 함수들을 연결해 연산하는 기능을 가지고 있다. 따라서 파이프 연산자를 사용하면 훨씬 가독성 있고 직관적인 코드를 작성할 수 있다. 형식은 `[dataset %>% 조건]`으로 파이프 연산자 기준으로 앞쪽에 있는 결과가 뒤에 있는 함수에 반영되어 연산이 된다.

dplyr 패키지 – 주요 함수

- dplyr 패키지는 데이터 전처리 작업을 할 수 있는 함수들을 제공하는 패키지로, 미리 주요 함수들의 기능과 사용법을 익혀두면 유용하게 쓰인다.

- filter()** : 조건에 맞는 행 추출 (형식 : `filter(조건)`)

```
# 남성만 추출
data1 <- base_data %>% filter(t_sex == 1)
```

- select()** : 변수(열) 추출 (형식 : `select(추출할 변수1, ...)`)

```
# 당뇨병 관련 변수(당뇨병 의사진단, 공복혈당)만 추출
data2 <- base_data %>% select(t_dm, t_glu0)
* 특정 변수를 제외하고 나머지를 불러오고 싶은 경우 제외할 변수만 써주고 변수명 앞에 마이너스 부호(-)를 붙여주면 됨
```

- arrange()** : 정렬 (형식 : `arrange(정렬할 변수1, ...)`)

```
# 비만도가 높은 순으로 정렬
data3 <- base_data %>% arrange(desc(t_bmi)) # 내림차순
```

- mutate()** : 새로운 변수 추가 (형식 : `mutate(새로운 변수 = 조건1, ...)`)

```
# 비만도 변수 계산하여 생성
data4 <- base_data %>% mutate(bmi = t_weight / ((t_height / 100) ^ 2))
```

- group_by()** : 그룹 생성 (형식 : `group_by(그룹 변수1, ...)`)

```
# 단독으로 쓰이기보다는 summarize() 함수 등과 함께 잘 쓰임
```

- summarize()** : 데이터 요약 (형식 : `summarize(저장할 변수 = 통계함수(요약할 변수))`)

```
# 성별 평균 공복혈당
data5 <- base_data %>% group_by(t_sex) %>% summarize(mean_glu0 = mean(t_glu0, na.rm = TRUE))
```

자료 가로 결합

• 가로 결합 함수

여러 자료의 가로 결합은 “변수 추가”로 생각하면 이해가 쉽다. 가로 결합을 하기 위해서는 `merge()` 함수를 사용하거나 mutating join 방법 중 해당 함수를 사용하면 된다. 만약 데이터셋이 2개 이상일 경우, 2개씩 순차적으로 결합해야 한다는 점을 주의해야 하며, 이로 인해 코드가 다소 길어질 수 있다. 또한 mutating join 방법은 dplyr 패키지 설치 및 로드가 먼저 선행되어야 실행됨을 주의해야 한다.

구분	방법 1	방법 2
inner join	<code>merge(A, B, by = "key")</code>	<code>inner_join(A, B)</code>
full outer join	<code>merge(A, B, by = "key", all = TRUE)</code>	<code>full_join(A, B)</code>
left outer join	<code>merge(A, B, by = "key", all.x = TRUE)</code>	<code>left_join(A, B)</code>
right outer join	<code>merge(A, B, by = "key", all.y = TRUE)</code>	<code>right_join(A, B)</code>

자료 세로 결합

• 세로 결합 함수 : `bind_rows(A, B, ...)`

여러 자료의 세로 결합은 “대상자 추가”로 생각하면 이해가 쉽다. 데이터셋의 열이 동일하지 않아도 되지만, 결합하고자 하는 변수의 변수명은 동일해야 하며 행 기준으로 결합된다. 함수 실행 전 dplyr 패키지 설치 및 로드 여부 확인하고 결합한다.

기타, 알아두면 좋아요

• `all_equal()` 함수 : `all_equal(A, B)`

A, B 두 개의 데이터프레임의 데이터가 동일한지 아닌지를 확인해주는 함수로 함수 실행 전 dplyr 패키지 설치 및 로드 여부를 확인하고 실행한다.

• `rm()` 함수 : `rm(A)`

A 삭제

4. 자료 분석 준비하기

4-1. 기본코드 결측치 처리하기

KoGES 역학자료는 설문 문항의 ‘미상/무응답’, ‘설문 문항 간의 상·하위 관계(해당없음)’, ‘해당변수의 조사유무(조사안함)’, ‘반복추적조사 통합자료의 경우, 추적조사 참여유무(추적조사 미참여)’ 등을 구분하기 위하여, 기본 코드가 적용되어 있으며, 구체적인 기본 코드는 다음과 같다.

표 8 | KoGES 기본코드의 종류와 정의

구분	코드명	코드	코드 정의		
결측	미상/무응답	99999	Null값(missing value) 또는 조사항목 상의 미상/무응답 값		
	해당없음	77777	조사항목에 대해 응답의 대상이 아닌 경우 예)		
			변수명	변수설명	변수값(코드)
			HTN	고혈압 과거력 - 진단여부 (1=아니오, 2=예)	1
	HTNAG	고혈압 과거력 - 처음 진단 나이	77777		
	조사안함	66666	특정 조사단위에 조사되지 않은 항목의 경우		
추적조사 미참여		55555	반복추적조사 통합자료에서 해당 차수의 조사에 참여하지 않은 경우		

‘음주 여부(t_drink)’와 ‘신장(t_height)’ 변수를 빈도 분석 및 평균 분석하면, 아래와 같이 미상/무응답에 대한 코드 값(‘99999’)이 포함된 결과가 출력이 되는 것을 확인 할 수 있다. 따라서 분석 전 기본코드를 결측치로 처리해주는 작업이 선행되어야 한다.

음주 여부 변수 빈도분석	신장 변수 평균분석
<pre>> descr::freq(base_data\$t_drink) base_data\$t_drink Frequency Percent 1 5015 50.15 2 478 4.78 3 4446 44.46 99999 61 0.61 Total 10000 100.00</pre>	<pre>> summary(base_data\$t_height) Min. 1st Qu. Median Mean 3rd Qu. Max. 117.0 154.0 159.0 619.3 166.0 99999.0</pre>
음주 여부(t_drink, 범주형) : 1=비음주, 2=과거 음주, 3=현재 음주	신장(t_height, 연속형) : ()cm

예제

기반조사 교육용 데이터에 포함된 변수의 기본코드 결측치 처리하기

방법 I. 변수별 기본코드 결측치 처리

```
# 기본코드 결측치 처리 (변수별) ----
base_data_null <- base_data
base_data_null$t_sex <- ifelse(base_data_null$t_sex %in% c(66666, 77777, 99999), NA, base_data_null$t_sex)
base_data_null$t_age <- ifelse(base_data_null$t_age %in% c(66666, 77777, 99999), NA, base_data_null$t_age)
base_data_null$t_htn <- ifelse(base_data_null$t_htn %in% c(66666, 77777, 99999), NA, base_data_null$t_htn)
:
base_data_null$t_glu0 <- ifelse(base_data_null$t_glu0 %in% c(66666, 77777, 99999), NA, base_data_null$t_glu0)
```

방법 II. 모든 변수 일괄 결측치 처리

```
# 기본코드 결측치 처리 (일괄) ----
base_data_null <- base_data
base_data_null[base_data_null == 66666 | base_data_null == 77777 | base_data_null == 99999] <- NA
```

《방법 II》를 이용하여 변수별 기본코드 코드(66666, 77777, 99999)를 결측치(NA)로 처리한 후, ‘음주 여부 (t_drink)’와 ‘신장(t_height)’ 변수를 빈도 분석 및 평균 분석하면, 각 분석에서 결측치를 제외한 자료의 분석 결과를 얻을 수 있다.

	기본코드 결측치 처리 전	기본코드 결측치 처리 후
음주 여부 변수 빈도분석	<pre>> desc::freq(base_data\$t_drink) base_data\$t_drink Frequency Percent 1 5015 50.15 2 478 4.78 3 4446 44.46 99999 61 0.61 Total 10000 100.00</pre>	<pre>> desc::freq(base_data_null\$t_drink) base_data_null\$t_drink Frequency Percent Valid Percent 1 5015 50.15 50.458 2 478 4.78 4.809 3 4446 44.46 44.733 NA's 61 0.61 Total 10000 100.00 100.000 > round(desc::freq(base_data_null\$t_drink), digits = 2) base_data_null\$t_drink Frequency Percent Valid Percent 1 5015 50.15 50.46 2 478 4.78 4.81 3 4446 44.46 44.73 NA's 61 0.61 Total 10000 100.00 100.00</pre>
	음주 여부(t_drink, 범주형) : 1=비음주, 2=과거 음주, 3=현재 음주	
신장 변수 평균분석	<pre>> summary(base_data\$t_height) Min. 1st Qu. Median Mean 3rd Qu. Max. 117.0 154.0 159.0 619.3 166.0 99999.0</pre>	<pre>> summary(base_data_null\$t_height) Min. 1st Qu. Median Mean 3rd Qu. Max. NA's 117 154 159 160 166 190 46</pre>
	신장(t_height, 연속형) : ()cm	

4-2. 변수 유형 변환하기

R에서 다루는 데이터 유형은 숫자형, 문자형, 논리형 등이 있다. 숫자형(numeric)은 예를 들어 몸무게(실수, double)나 맥박수(정수, integer)처럼 숫자 크기 자체에 의미가 있는 자료를 말하고, 문자형(character)은 문자가 포함된 자료를 말하며, 논리형(logical)은 참(TRUE) 또는 거짓(FALSE)으로 분류되는 자료를 말한다. 통계적으로는 숫자형 변수를 '연속형 변수 또는 이산형 변수' 라고 표현하고, 문자형 또는 숫자형인데 숫자 크기 자체보다는 분류에 의미를 둔 변수를 '범주형 변수'로 분류한다. R에서는 이러한 범주형 변수를 집단으로 인식하기 위해서는 추가적인 유형 변환 작업(요인형으로 변환)이 필요하며, 변환된 자료는 수준(levels)이라는 범주(집단)의 정보를 추가적으로 제공한다. R에서 제공하는 함수 기능을 올바르게 적용하기 위해서는 그에 맞는 데이터 유형으로 변환해 사용해야 함을 유의하도록 하자.

아래의 <그림 16>는 제공된 KoGES 기반조사 교육용 데이터를 R로 불러와 결합한 데이터 셋을 가지고 str() 함수를 이용해 데이터 구조를 확인한 결과이다. 각각의 변수는 R이 인식하는 변수 유형으로 불러와져 있다.

```

> str(base_data)
'data.frame': 10000 obs. of 11 variables:
 $ t_id      : Factor w/ 10000 levels "K_BASE_00001",...: 1 2 3 4 5 6 7 8 9 10 ...
 $ t_sex     : int  2 2 2 2 2 1 2 1 2 1 ...
 $ t_age     : int  42 58 60 73 59 54 41 58 51 56 ...
 $ t_htn     : int  1 2 2 2 1 2 1 2 2 1 ...
 $ t_dm      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ t_drink   : int  1 2 1 1 1 1 3 3 2 1 ...
 $ t_smoke   : int  1 1 1 1 3 1 1 2 2 ...
 $ t_weight  : int  64 64 54 55 60 81 58 72 75 68 ...
 $ t_height  : int  164 156 154 143 150 169 168 165 154 170 ...
 $ t_bmi     : int  24 26 23 27 27 28 21 26 32 24 ...
 $ t_glu0    : int  68 86 86 108 91 90 86 91 97 87 ...
  
```

그림 16 | R로 불러와 결합된 KoGES 기반조사 교육용 데이터 변수 유형

만약 신장(t_height)의 변수가 숫자형이 아닌 문자형인 경우, 이를 숫자형으로 변환하지 않고 문자형인 상태로 평균 신장을 산출하고자 한다면, 콘솔 창에 평균값이 아닌 다음과 같은 결과가 뜨는 것을 확인할 수 있다. 따라서 자료 분석에 앞서 올바른 분석과 결과 해석을 위해 각각의 변수 유형을 파악하고, 필요 시 변수의 유형을 변환해 주는 작업이 필요하다.

```

> base_data_null$t_height <- as.character(base_data_null$t_height)
> summary(base_data_null$t_height)
  Length      Class      Mode 
10000 character character 
>
  
```

그림 17 | 문자형 변수의 기술통계량을 구한 사례

다음의 예제를 통해 기반조사 교육용 데이터에 포함된 변수 중 변수 유형 변환이 필요한 변수는 무엇이 있으며, 어떻게 변환하는지 살펴보도록 하자.

예제

기반조사 교육용 데이터 변수 유형 변환

```
# 변수 유형 변환 ----
base_data_type <- base_data_null
base_data_type$t_id <- as.character(base_data_type$t_id) # 요인형을 문자형으로 변환
base_data_type$t_sex <- as.factor(base_data_type$t_sex) # 숫자형을 요인형으로 변환
base_data_type$t_htn <- as.factor(base_data_type$t_htn)
base_data_type$t_dm <- as.factor(base_data_type$t_dm)
base_data_type$t_drink <- as.factor(base_data_type$t_drink)
base_data_type$t_smoke <- as.factor(base_data_type$t_smoke)

# 변수 유형 확인 ----
sapply(base_data_null, class) # 변경 전 모든 변수 유형 확인
sapply(base_data_type, class) # 변경 후 모든 변수 유형 확인
```

결과

```
> sapply(base_data_null, class) # 변수유형 변경 전
  t_id t_sex t_age t_htn t_dm t_drink t_smoke t_weight t_height t_bmi t_glu0
"factor" "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer" "integer"
> sapply(base_data_type, class) # 변수유형 변경 후
  t_id t_sex t_age t_htn t_dm t_drink t_smoke t_weight t_height t_bmi t_glu0
"character" "factor" "integer" "factor" "factor" "factor" "factor" "integer" "integer" "integer" "integer"
> |
```



데이터(변수) 유형

• 데이터 유형별 확인 및 전환 함수

유형 종류	유형 확인 함수	유형 전환 함수	함수 설명
숫자형	is.numeric()	as.numeric()	숫자형으로 변환
	is.integer()	as.integer()	정수형으로 변환
	is.double()	as.double()	실수형으로 변환
문자형	is.character()	as.character()	문자형으로 변환
논리형	is.logical()	as.logical()	논리형으로 변환

※ R에서의 데이터 유형은 통계적으로는 보통 숫자형은 연속형/이산형으로 구분되며, 숫자형이든 문자형이든 집단으로의 구분이 의미가 있는 경우 이를 범주형(요인형)으로 구분한다.

dplyr 패키지 - 주요 함수

• 변수 유형 파악을 위한 함수

```
# class() 함수 : 데이터셋의 지정한 변수 유형 파악
class(base_data_type$t_sex) # 데이터셋 변수 중 성별 변수의 유형 파악

# sapply() 함수 : 데이터셋의 모든 변수에 함수를 동시에 적용할 수 있는 함수
sapply(base_data_type, class) # 데이터셋의 모든 변수 유형을 파악
```

5. 자료 분석하기

5-1. 분석 대상자 선정

자료 분석에 앞서, 분석에 포함할 연구 대상자를 선정하여야 한다. 일반적으로 의학 및 보건학 연구 논문에는 연구 대상자 흐름도를 통해 원자료에 포함된 전체 대상자 중 특정 기준에 따라 분석에 포함된 대상자의 수를 나타낸 그림을 확인 할 수 있다.

예를 들어, '비만한 사람에서 당뇨병 위험이 높아질까?'라는 연구 주제에 따라 분석을 하기 위해서는 당뇨병과 비만을 정의하는 변수가 필수적이다. 먼저 두 지표를 정의하는데 사용되는 변수가 결측인 경우, 이를 제외한 후 해당 분석을 진행해 보자.

예제

당뇨병 또는 비만 관련 변수가 결측인 대상자 제외하기

I. 변수 설명

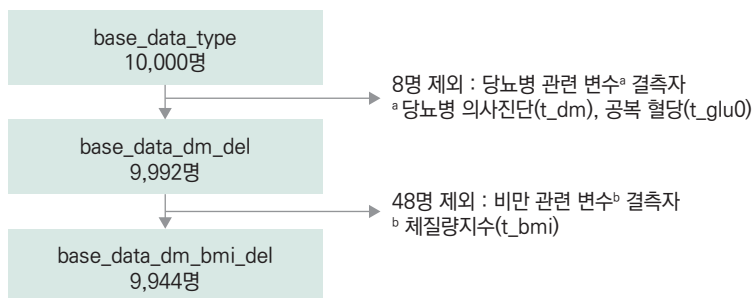
변수명	변수 설명	변수값 설명	변수 유형
t_dm	당뇨병 의사진단	1=아니오, 2=예	범주형
t_glu0	공복 혈당	() mg/dL	연속형
t_bmi	체질량지수(BMI)	() kg/m ²	연속형

II. R을 이용한 분석 대상자 선정

```
# 당뇨병 관련 변수가 모두 결측인 대상자 제외 ----
base_data_dm_del <- base_data_type %>% dplyr::filter(! (is.na(t_dm) & is.na(t_glu0)))
dim(base_data_dm_del)      # 대상자 수 확인 : 9,992명

# 비만 관련 변수가 결측인 대상자 제외 ----
base_data_dm_bmi_del <- base_data_dm_del %>% dplyr::filter(! (is.na(t_bmi)))
dim(base_data_dm_bmi_del)  # 9,944명
```

결과





결측치 관련 함수

- 결측치 확인

`is.na(base_data_type)` # 결측치 있을 때 함수의 리턴값은 TRUE

- 결측값이 하나라도 있는 행은 모두 제거하고 부르기

`na.omit(base_data_type)`

- 데이터셋 모든 변수별 결측치 개수 확인하기

`colSums(is.na(base_data_type))`

5-2. 변수 생성

원 자료에 포함된 여러 변수들을 조합하여 하나의 변수를 만들거나, 연속형 변수를 특정 값을 기준으로 나누어 범주형 변수로 생성하는 등 분석에 앞서 새로운 변수의 생성이 필요한 경우가 있다. 이러한 경우 dplyr 패키지의 함수들을 이용하여 연구 목적에 맞는 새로운 변수를 생성할 수 있다.

아래 기준에 따라 체질량지수(t_bmi) 변수를 이용하여 비만도 변수(bmi_gr)를, 당뇨병 의사진단(t_dm)과 공복 혈당(t_glu0) 변수를 이용하여 당뇨병 여부 변수(dm)를 생성해보자.

비만도 ^a 변수 생성 기준	당뇨병 여부 ^b 변수 생성 기준
<ul style="list-style-type: none"> · 저 체 중 : 체질량지수(BMI) < 18.5 kg/m² · 정 상 체 중 : 체질량지수(BMI) 18.5 ~ 23 kg/m² · 과 체 중 : 체질량지수(BMI) 23 ~ 25 kg/m² · 비 만 : 체질량지수(BMI) ≥ 25 kg/m² 	<ul style="list-style-type: none"> · 당 뇨 병 : 공복혈당 ≥ 126 mg/dL 이상 또는 과거 의사로 부터 당뇨병을 진단 받은 적이 있는 경우 (당뇨병 의사진단 '예') · 정 상 : 당뇨병에 해당하지 않은 모든 경우
^a 세계보건기구(WHO) 아시아-태평양 지역 기준	^b Report or a WHO/IDF consultation (2006) 기준

예제

비만도 및 당뇨병 여부 변수 생성하기

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
t_bmi	체질량지수(BMI)	() kg/m ²	연속형
t_glu0	혈당	() mg/dL	연속형
t_dm	당뇨병 의사진단	1=아니오, 2=예	범주형

II. R을 이용한 새로운 변수 생성

```
# 비만도 변수(bmi_gr) 생성 ----
base_data_bmi <- base_data_dm_bmi_del %>%
  dplyr::mutate(bmi_gr = ifelse(t_bmi < 18.5, 1,
                              ifelse(t_bmi >= 18.5 & t_bmi < 23, 2,
                                    ifelse(t_bmi >= 23 & t_bmi < 25, 3,
                                          ifelse(t_bmi >= 25, 4, NA))))))
base_data_bmi$bmi_gr <- as.factor(base_data_bmi$bmi_gr) # 범주형

# 당뇨병 변수(dm) 생성 ----
base_data_bmi_dm <- base_data_bmi %>% mutate(dm = ifelse(t_dm == 2 | t_glu0 >= 126, 1, 0))
base_data_bmi_dm$dm[is.na(base_data_bmi_dm$dm)] <- 0 # NA <- 0
base_data_bmi_dm$dm <- as.factor(base_data_bmi_dm$dm) # 범주형

# 최종 데이터셋(base_data_final) 저장 및 확인 ----
base_data_final <- base_data_bmi_dm
save(base_data_final, file = "base_data_final.RData") # RData 저장(메모리 -> 하드)
str(base_data_final)
head(base_data_final)
```

결과

```
Console Terminal Jobs
D:/kges/data/
> str(base_data_final)
'data.frame': 9944 obs. of 13 variables:
 $ t_id : chr "K_BASE_00001" "K_BASE_00002" "K_BASE_00003" "K_BASE_00004" ...
 $ t_sex : Factor w/ 2 levels "1","2": 2 2 2 2 2 1 2 1 2 1 ...
 $ t_age : int 42 58 60 73 59 54 41 58 51 56 ...
 $ t_htn : Factor w/ 2 levels "1","2": 1 2 2 2 1 2 1 2 2 1 ...
 $ t_dm : Factor w/ 2 levels "1","2": 1 1 1 1 1 1 1 1 1 1 ...
 $ t_drink : Factor w/ 3 levels "1","2","3": 1 2 1 1 1 1 3 3 2 1 ...
 $ t_smoke : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 3 1 1 2 2 ...
 $ t_weight: int 64 64 54 55 60 81 58 72 75 68 ...
 $ t_height: int 164 156 154 143 150 169 168 165 154 170 ...
 $ t_bmi : int 24 26 23 27 27 28 21 26 32 24 ...
 $ t_glu0 : int 68 86 86 108 91 90 86 91 97 87 ...
 $ bmi_gr : Factor w/ 4 levels "1","2","3","4": 3 4 3 4 4 2 4 4 3 ...
 $ dm : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
> head(base_data_final)
  t_id t_sex t_age t_htn t_dm t_drink t_smoke t_weight t_height t_bmi t_glu0 bmi_gr dm
1 K_BASE_00001 2 42 1 1 1 1 64 164 24 68 3 0
2 K_BASE_00002 2 58 2 1 2 1 64 156 26 86 4 0
3 K_BASE_00003 2 60 2 1 1 1 54 154 23 86 3 0
4 K_BASE_00004 2 73 2 1 1 1 55 143 27 108 4 0
5 K_BASE_00005 2 59 1 1 1 1 60 150 27 91 4 0
6 K_BASE_00006 1 54 2 1 1 3 81 169 28 90 4 0
```


5-3. 빈도 분석

일반량 질적 자료 분석 중 하나로 명목이나 순위 척도 값을 가지는 범주형 변수에 대하여, 각 변수 값에 대한 빈도와 백분율 등을 구하는 빈도 분석에 대해 알아보도록 하자. R에서는 descr 패키지의 freq() 함수를 이용하면 빈도표를 구할 수 있으며, 기본적으로 막대그래프가 같이 출력된다.

예제

비만도 및 당뇨병 여부 빈도 분석하기 - 빈도표

I. 변수 설명

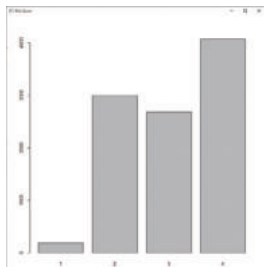
변수명	변수 설명	변수값 설명	변수 유형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
dm	당뇨병 유병 여부	0=정상, 1=당뇨병	범주형

II. R을 이용한 통계분석

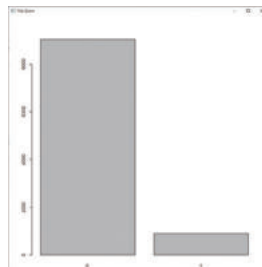
```
# 빈도 분석 ----
install.packages("descr")
library(descr)
table.bmi_gr <- descr::freq(base_data_final$bmi_gr) # bmi_gr 빈도표
round(table.bmi_gr, digits = 2)
table.dm <- descr::freq(base_data_final$dm) # dm 빈도표
round(table.dm, digits = 2)
```

결과

```
> round(table.bmi_gr, digits = 2)
base_data_final$bmi_gr
Frequency Percent
1          191     1.92
2       2999    30.16
3       2683    26.98
4       4071    40.94
Total     9944   100.00
>
```



```
> round(table.dm, digits = 2)
base_data_final$dm
Frequency Percent
0          9043    90.94
1           901     9.06
Total     9944   100.00
>
```



결과 해석

- 분석대상자 총 9,944명 중 저체중(bmi_gr=1)은 191명(1.92%), 정상체중(bmi_gr=2)은 2,999명(30.16%), 과체중(bmi_gr=3)은 2,683명(26.98%), 비만(bmi_gr=4)은 4,071명(40.94%)이다.
- 당뇨병 유병자(dm=1)는 901명으로 전체 대상자 중 9.06%이고, 정상(dm=0)은 9,043명으로 90.94%이다.

범주형 변수에 대하여 별도의 레이블(label) 없이 분석을 수행하면, 비만도 변수(bmi_gr)의 경우 1, 2, 3, 4와 같이 범주의 값이 출력이 된다. 이때 factor() 함수를 이용하여 비만도 변수에 대한 레이블을 1=저체중, 2=정상체중, 3=과체중, 4=비만으로 지정한 후 분석을 실시하면 각 범주의 값에 대한 설명으로 출력된다.

예제

비만도 및 당뇨병 여부 빈도 분석하기 - 빈도표(변수 레이블 지정)

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
dm	당뇨병 유병 여부	0=정상, 1=당뇨병	범주형

II. R을 이용한 통계분석

```
# 레이블 후 빈도 분석 ----
base_data_final$bmi_gr <- factor(base_data_final$bmi_gr,
                                levels = c("1", "2", "3", "4"),
                                labels = c("저체중", "정상체중", "과체중", "비만"))
base_data_final$dm <- factor(base_data_final$dm,
                             levels = c("0", "1"),
                             labels = c("정상", "당뇨"))

table.bmi_gr <- descr::freq(base_data_final$bmi_gr) # bmi_gr 빈도표
round(table.bmi_gr, digits = 2)
table.dm <- descr::freq(base_data_final$dm) # dm 빈도표
round(table.dm, digits = 2)
```

결과

변수 레이블을 지정한 경우

```
> round(table.bmi_gr, digits = 2)
base_data_final$bmi_gr
      Frequency Percent
저체중      191      1.92
정상체중  2999     30.16
과체중     2683     26.98
비만       4071     40.94
Total     9944    100.00

> round(table.dm, digits = 2)
base_data_final$dm
      Frequency Percent
정상      9043     90.94
당뇨       901      9.06
Total     9944    100.00
```

참고

변수 레이블을 지정하지 않은 경우

```
> round(table.bmi_gr, digits = 2)
base_data_final$bmi_gr
      Frequency Percent
1           191      1.92
2          2999     30.16
3           2683     26.98
4          4071     40.94
Total       9944    100.00

> round(table.dm, digits = 2)
base_data_final$dm
      Frequency Percent
0           9043     90.94
1            901      9.06
Total       9944    100.00
```

R에서는 두 개의 범주형 변수에 대하여 CrossTable() 함수를 통해 교차표(cross table) 작성이 가능하며, 각 그룹의 빈도 차이(비율 차이) 분석을 통해 변수간의 관련성을 분석할 수 있다. 대표적인 방법이 카이제곱 검정이며, chisq.test() 함수를 이용한다.

예제

성별에 따른 비만도 빈도 분석하기 - 교차표, 카이제곱 검정

I. 가설 설정

- 귀무가설: 성별에 따라 비만도(저체중/정상체중/과체중/비만)에 차이가 없다(독립이다).
- 대립가설: 성별에 따라 비만도(저체중/정상체중/과체중/비만)에 차이가 있다(독립이 아니다).

II. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
t_sex	성별	1=남자, 2=여자	범주형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형

III. R을 이용한 통계분석

```
# 교차표, CrossTable() 함수 이용 ----
install.packages("gmodels") # 패키지 설치
library(gmodels)           # 패키지 로드
base_data_final$t_sex <- factor(base_data_final$t_sex, levels = c("1", "2"), labels = c("남자", "여자"))
gmodels::CrossTable(base_data_final$bmi_gr, base_data_final$t_sex)

# 카이제곱 검정 ----
chisq.test(base_data_final$bmi_gr, base_data_final$t_sex)
```

결과

base_data_final\$bmi_gr	base_data_final\$t_sex		Row Total
	남자	여자	
저체중	47 5.449 0.246 0.014 0.005 0.014 0.002 0.014 191	144 2.875 0.754 0.022 0.014 0.014 0.002 0.014	
정상체중	840 36.963 0.280 0.245 0.084 0.217 0.280 0.245 2999	2159 19.498 0.720 0.332 0.217 0.217 0.332 0.332	
과체중	923 0.013 0.344 0.269 0.093 0.270 0.270 0.270 2683	1760 0.007 0.656 0.270 0.177 0.270 0.270 0.270	
비만	1624 33.850 0.399 0.423 0.163 0.409 0.409 0.409 4071	2447 17.856 0.601 0.316 0.246 0.316 0.316 0.316	
Column Total	3434 0.345 0.655 0.655 0.655 0.655 0.655 0.655 9944	6510 0.655 0.655 0.655 0.655 0.655 0.655 0.655	

Cell Contents

N
Chi-square contribution
N / Row Total
N / Col Total
N / Table Total

```
> chisq.test(base_data_final$bmi_gr, base_data_final$t_sex)

Pearson's Chi-squared test

data: base_data_final$bmi_gr and base_data_final$t_sex
X-squared = 116.51, df = 3, p-value < 2.2e-16
```

결과 해석

- 남자 중 저체중에 해당하는 비율은 1.4%, 정상체중은 24.5%, 과체중은 26.9%, 비만은 47.3%이며, 여자 중 저체중 비율은 2.2%, 정상체중은 33.2%, 과체중은 27.0%, 비만은 37.6%이다.
- χ^2 검정통계량 값은 116.51이며, 유의확률은 유의수준 0.05보다 현저히 작아 성별에 따라 비만도에 차이가 없다는 귀무가설을 기각할 수 있다.

5-4. 기술통계

양적 자료를 분석하는 방법으로 기술통계량을 통하여 자료가 가지는 중요한 특징을 찾아낼 수 있다. 여기서 기술 통계량이란 요약통계량이라고도 하며, 특정한 연산을 통하여 얻어진 수치로 자료의 중심 / 퍼짐정도 / 분포 모양 등을 알려준다. R에서는 기본적으로 summary() 함수를 사용하면 6가지의 기술통계량(최솟값 / 1사분위수 / 중위수 / 평균 / 3사분위수 / 최댓값)을 제공해주며, psych 패키지의 describe() 함수를 사용하면 summary() 함수 보다 좀 더 다양한 정보를 제공해준다. 또한 양적 자료의 특징은 기술통계량 같은 수치를 통한 확인 방법 이외에 히스토그램(hist() 함수)이나 상자그림(boxplot() 함수) 등의 그래프를 통해서도 확인할 수 있다.

예제

분석 대상자의 연령, 체질량지수 특징 파악하기

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
t_age	연령	만 ()세	연속형
t_bmi	체질량지수(BMI)	() kg/m ²	연속형

II. R을 이용한 통계분석

```
# 기술통계 ----
summary(base_data_final$t_age)

install.packages("psych")      # 패키지 설치
library(psych)                 # 패키지 로드
psych::describe(base_data_final$t_age)

summary(base_data_final$t_bmi)
psych::describe(base_data_final$t_bmi)

# 그래프 ----
par(mfrow=c(1,2))
hist(base_data_final$t_age, main = "히스토그램", xlab = "t_age")
boxplot(base_data_final$t_age, main = "상자 그림", xlab = "t_age")

hist(base_data_final$t_bmi, main = "히스토그램", xlab = "t_bmi")
boxplot(base_data_final$t_bmi, main = "상자 그림", xlab = "t_bmi")
```

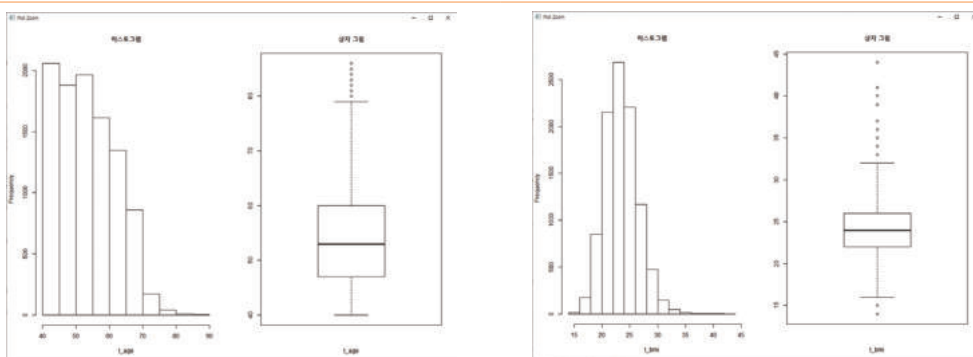
결과

```
> summary(base_data_final$t_age)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  40.0   47.0   53.0   53.7   60.0   86.0

> psych::describe(base_data_final$t_age)
  vars   n mean  sd median trimmed  mad min max range skew kurtosis  se
x1      1 9944 53.7 8.72   53   53.4 10.38  40  86  46  0.3  -0.67 0.09

> summary(base_data_final$t_bmi)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.00  22.00  24.00  24.02  26.00  44.00

> psych::describe(base_data_final$t_bmi)
  vars   n mean  sd median trimmed  mad min max range skew kurtosis  se
x1      1 9944 24.02 2.98   24   23.9 2.97  14  44  30 0.52   1.07 0.03
```



결과 해석

- 분석 대상자 총 9,944명의 연령(t_age)의 기본적 특성과 분포를 확인한 결과, 평균은 53.7(세)이고 중위수는 53.0(세)이며, 1사분위수 47.0(세)와 3사분위수 60.0(세) 사이에 자료의 절반이 분포하고 있고, 최솟값은 40.0(세), 최댓값은 86.0(세)이다.
- 체질량지수(BMI, t_bmi)의 경우도 분석 대상자 9,944명은 동일하며, 평균은 24.02(kg/m²), 중위수 24.0(kg/m²)이며, 1사분위수 22.0(kg/m²)과 3사분위수 26.0(kg/m²) 사이에 자료의 절반이 분포하고 있고, 최솟값은 14.0(kg/m²), 최댓값은 44.0(kg/m²)이다. 또한 체질량지수(BMI)는 오른쪽으로 꼬리가 긴(왼쪽으로 치우친)분포로 이상치(outliers)로 판단되는 값들도 주로 오른쪽에 있는 것을 상자그림을 통해 확인할 수 있다.



기술통계

수집된 자료를 특정한 연산을 통해 정리, 요약한 값을 기술통계량이라고 하며, 기술통계량 값을 개별적으로 구하고 싶을 때는 다음의 함수를 사용하면 된다.

• 함수 소개

함수	기능	설 명
mean()	평균	모든 데이터 합을 개수로 나눈 값, 이상치에 민감
	절사평균	정렬된 데이터의 양쪽의 일부 자료를 제거한 후 구한 평균으로 이상치에 덜 민감 <옵션> trim=0.05 (5% 절사평균)
median()	중위수	정렬된 데이터의 가운데 값, 이상치에 덜 민감
min()	최솟값	정렬된 데이터의 가장 작은 값
max()	최댓값	정렬된 데이터의 가장 큰 값
range()	범위	최댓값 - 최솟값
quantile()	분위수	Q1 정렬된 데이터의 하위 25% 지점 값, probs=0.25
		Q2 정렬된 데이터의 하위 50% 지점 값, probs=0.5
		Q3 정렬된 데이터의 하위 75% 지점 값, probs=0.75
var()	분산	데이터 분포가 평균으로부터 퍼진 정도
sd()	표준편차	분산의 제곱근, 데이터의 퍼진 정도
skew()	왜도	데이터 분포의 비대칭성 정도 >0이면 오른쪽으로 꼬리가 긴(왼쪽으로 치우친) 분포
kurtosi()	첨도	데이터 분포의 꼬리가 두터운 정도 >0이면 정규분포보다 꼬리가 두터움

• 기술통계량 구하기

함수[패키지]	예 제
summary{base}	# min, Q1, median, mean, Q2, max 정보 제공 summary(base_data_final\$t_bmi)
describe{psych}	# summary() 함수보다 왜도나 첨도 등의 조금 더 많은 정보 제공 describe(base_data_final\$t_bmi)
summarise{dplyr}	base_data_final %>% select(t_bmi) %>% summarise(mean_bmi = mean(t_bmi, na.rm = TRUE))

• 그룹별 기술통계량 구하기

함수[패키지]	예 제
summaryBy{doBy}	summaryBy(t_bmi ~ t_sex, data = base_data_final, FUN = c(mean, sd,...))
describeBy{psych}	describeBy(base_data_final\$t_bmi, base_data_final\$t_sex)
group_by{dplyr}, summarise{dplyr}	base_data_final %>% select(t_bmi, t_sex) %>% group_by(t_sex) %>% summarise(mean_bmi = mean(t_bmi, na.rm = TRUE))

상자 그림

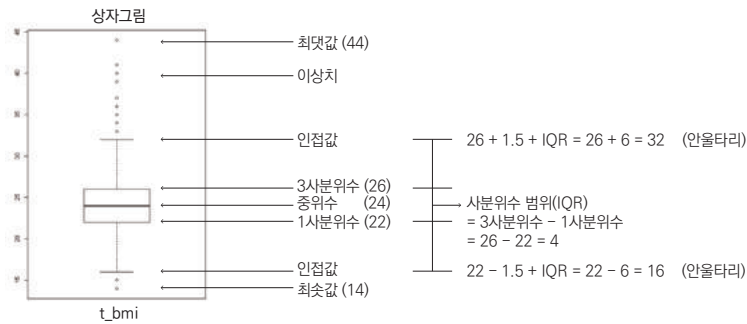
상자 그림은 자료의 분포 형태를 파악할 뿐만 아니라 이상치를 판단하는 기준으로 사용되며, 특히 여러 그룹의 자료를 비교할 경우, 효율적으로 비교할 수 있다.

• 상자 그림 형태

예제

체질량지수(t_bmi)의 상자 그림 결과

`boxplot(base_data_final$t_bmi, main = "상자 그림", xlab = "t_bmi")` 적용 결과



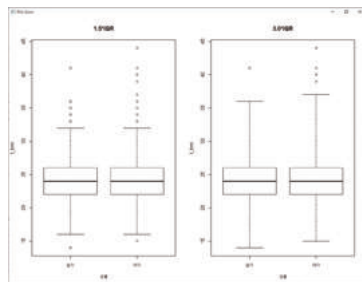
안울타리 안쪽에서 가장 가까운 값을 인접값이라고 하고, 안울타리를 벗어난 값을 이상치로 판단한다. 이때 이상치를 판단하는 기준으로는 보통 사분위범위(IQR)의 1.5배($1.5 \times \text{IQR}$)가 사용되지만, IQR의 3배($3 \times \text{IQR}$) 등 사용자가 원하는 값으로 변경할 수 있으며, $3 \times \text{IQR}$ 을 벗어날 경우 심한 이상치로 구분한다.

• 그룹별 상자 그림 적용

예제

성별(t_sex) 체질량지수(t_bmi)의 비교를 위한 상자 그림

`boxplot(t_bmi ~ t_sex, data = base_data_final, 기타 <argument>)`



구분	argument	내용
이상치 판단기준	range = 1.5	$1.5 \times \text{IQR}$
	range = 3	$3 \times \text{IQR}$
이상치 출력여부	outline = FALSE	이상치 미출력
제목, 축 정보	main = "제목"	상자그림 제목 지정
	xlab = "성별"	x축 이름 지정



정규성 검토

KoGES 수집자료의 경우 대용량 자료로 중심극한정리에 의해 자료의 분포가 정규분포를 따르는지에 대한 검정이 필요하지 않다. 하지만 자료의 규모가 작으면 연속형 변수가 정규분포를 따르지 않을 경우 적용하는 통계분석법이 달라질 수 있으므로, 이에 대한 검토가 필요하다. 앞서 다룬 체질량지수(t_bmi) 변수를 사용해 정규성 검토를 위한 방법에 대해 알아두도록 하자.

• Kolmogorov-Smirnov 검정

귀무가설 : 체질량지수의 분포는 정규분포를 따른다.

유의확률 < 0.05 이면, 귀무가설 기각

```
KS <- ks.test(base_data_final$t_bmi, pnorm)
```

KS

```

Console Terminal Jobs
D:/koges/ >
> KS <- ks.test(base_data_final$t_bmi, "pnorm" )
Warning message:
In ks.test(base_data_final$t_bmi, "pnorm") :
  ties should not be present for the Kolmogorov-Smirnov test
> KS

      One-sample Kolmogorov-Smirnov test

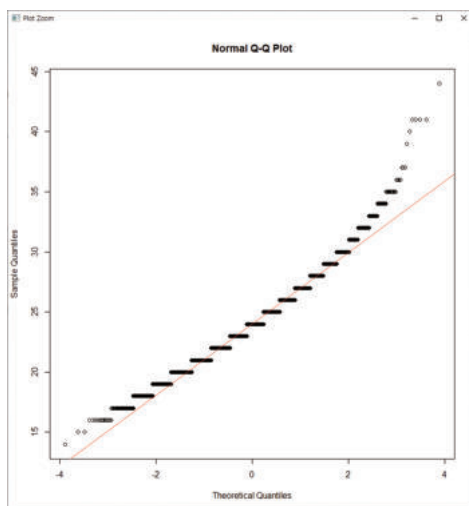
data:  base_data_final$t_bmi
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
  
```


• Q-Q plot

점들이 직선에 일치할수록 정규분포에 가까움

```
qqnorm(base_data_final$bmi)
```

```
qqline(base_data_final$bmi, col = 2)
```



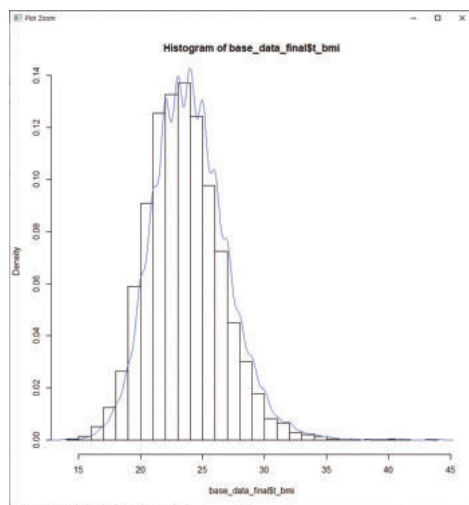
• Histogram

freq=FALSE : y축이 빈도가 아닌 확률 의미

```
hist(base_data_final$bmi, freq = FALSE, breaks = 30)
```

Kernel Density Plot

```
lines(density(base_data_final$bmi), col = 4)
```



5-5. 두 집단 평균 비교

독립적인 두 집단 평균이 같은지에 대한 검정은 t-검정(t-test)을 통해 이루어지며, R에서는 `t.test()` 함수를 사용한다. t-검정에 앞서 `var.test()` 함수를 이용해 두 집단 간 분산이 동질적인지를 먼저 판단하고, 등분산성 여부에 따라 그에 맞는 t-검정을 수행해야 한다. t-검정시 종속변수의 변수 유형은 연속형 변수이어야 하며, 독립성, 정규성 검토가 필요하다. 만약 정규성을 만족하지 않는다면, 비모수적인 방법을 사용해야 한다.

예제

성별에 따른 연령 평균 비교

I. 가설 설정

- 귀무가설 : 성별에 따라 연령의 평균은 차이가 없다.
- 대립가설 : 성별에 따라 연령의 평균은 차이가 있다.

II. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
t_sex	성별	1=남자, 2=여자	범주형
t_age	연령	만 ()세	연속형

II. R을 이용한 통계분석

```
# 성별에 따른 평균 연령 확인 ----
psych::describeBy(base_data_final$t_age, base_data_final$t_sex)

# 두 모집단의 등분산성 가정 검토 ----
var.test(t_age ~ t_sex, data = base_data_final)

# 두 모집단의 모평균 차이 검정 ----
t.test(t_age ~ t_sex, data = base_data_final, var.equal = FALSE) # var.equal = FALSE (이분산)
```

결과**① 성별에 따른 평균 연령 확인**

```
> psych::describeBy(base_data_final$t_age, base_data_final$t_sex)
```

```
Descriptive statistics by group
```

```
group: 남자
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
x1	1	3434	54.55	8.91	54	54.39	10.38	40	83	43	0.15	-0.87	0.15

```
group: 여자
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
x1	1	6510	53.25	8.59	52	52.88	8.9	40	86	46	0.37	-0.52	0.11

② 집단 간 등분산성 검토

```
> var.test(t_age ~ t_sex, data=base_data_final)
```

```
F test to compare two variances
```

```
data: t_age by t_sex
```

```
F = 1.0759, num df = 3433, denom df = 6509, p-value = 0.01367
```

```
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
```

```
1.015109 1.141045
```

```
sample estimates:
```

```
ratio of variances
```

```
1.075948
```

③ 집단 간 평균 차이 검정

```
> t.test(t_age ~ t_sex, data=base_data_final, var.equal=FALSE)
```

```
Welch Two Sample t-test
```

```
data: t_age by t_sex
```

```
t = 7.0046, df = 6766.5, p-value = 2.715e-12
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
0.9363759 1.6641635
```

```
sample estimates:
```

```
mean in group 남자 mean in group 여자
```

```
54.55096 53.25069
```

결과 해석

- 성별에 따른 연령 평균은 남자 54.55(세), 여자 53.25(세)로 남자의 평균 연령이 높다.
- var.test() 함수를 이용하여 두 집단 간의 분산이 같다는 귀무가설에 대한 F검정을 실시한 결과, 유의확률이 0.0137로 유의수준 0.05보다 작으므로 두 집단 간 분산이 같다고 할 수 없다.
- 등분산성 검정 결과에 따라 t.test() 함수를 이용하여 두 집단(남,여)의 연령 평균에 대한 차이를 검정한 결과, 유의확률은 유의수준 0.05보다 현저히 작아 '성별에 따라 연령의 평균은 차이가 없다'는 귀무가설을 기각할 수 있다.



집단 간 중심 차이에 대한 검정

여러 모집단의 중심 차이에 대한 추정과 검정에 대해서는 정규분포 가정을 만족하는 경우 모수적 검정을 시행하고, 정규분포 가정을 충족하지 못하거나 모집단 분포 형태를 모르는 경우에는 비모수적 검정을 시행한다.

• 모수와 비모수 검정법 비교

구분	모수적 검정		비모수적 검정	
	통계 분석	R에서의 함수	통계 분석	R에서의 함수
2 sample	2 sample t-test	t.test()	Wilcoxon rank sum test (Mann-Whitney U-test)	wilcox.test()
	paired 2 sample t-test	t.test(paired=TRUE)	Wilcoxon signed rank test	wilcox.test(paired=TRUE)
≥3 sample	one-way ANOVA	aov() 또는 oneway.test()	Kruskal-Wallis test	kruskal.test()

※ 위에서 소개한 방법은 큰 틀에서의 통계적인 방법을 소개한 것으로, 각 방법별 세부적인 가정 만족 여부에 따라 분석 방법이 달라질 수 있음.

5-6. 분산분석

분산분석(ANOVA, analysis of variance)은 셋 이상의 집단 간 평균을 비교하고자 할 때 사용한다. 요인의 수가 하나인 경우 1요인 분산분석, 둘인 경우 2요인 분산분석이라고 한다. 독립변수는 범주형이고 종속변수는 연속형이면서 정규성과 등분산성을 만족해야 하지만 정규성가정에 크게 제약을 받지 않으므로, 등분산성을 만족하는지만 확인하도록 하자. 등분산성을 만족하면 일반적인 ANOVA 분석을 시행하면 되고, 만약 만족하지 않는다면 이질적 분산에 대한 보정을 위해 Welch's ANOVA 분석을 수행해야 한다. 전자라면 R에서 제공되는 함수 중 `aov()` 함수를 이용하면 되고, 후자라면 `oneway.test()` 함수를 이용하면 된다.

분산분석 결과 귀무가설이 기각(최소한 어느 한 집단과 다른 집단 간 평균 차이가 있다) 되었다면, 어느 집단 간에 평균의 차이가 있는지에 대한 추가적인 확인이 필요하다. 추가적인 확인을 위해 사후검정 즉, 다중비교검정(multiple comparison tests)을 시행한다. 다중비교 분석 기법에는 '등분산성 여부'와 비교 집단의 '표본크기 동일 여부'에 따라 사용할 수 있는 방법이 달라지며, 어떠한 사후 검정 방법을 사용하느냐에 따라 조금씩 다른 결과가 나올 수 있음을 유념하고 사용해야 한다.

예제

비만도(저체중/정상체중/과체중/비만)에 따른 혈당 수준의 평균 차이 분석하기

I. 가설 설정

귀무가설 : 비만도에 따라 혈당 수준의 평균 차이가 없다.

대립가설 : 비만도에 따라 혈당 수준의 평균 차이가 있다.

II. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
t_glu0	혈당	() mg/dL	연속형

II. R을 이용한 통계분석

```
# 비만도 집단별 혈당 평균 ----
psych::describeBy(base_data_final$t_glu0, base_data_final$bmi_gr)

# 등분산성 검토 ----
car::leveneTest(t_glu0 ~ bmi_gr, data = base_data_final)

# Welch's ANOVA ----
welch <- oneway.test(t_glu0 ~ bmi_gr, data = base_data_final, var.equal = FALSE)
welch

# 사후검정(games-howell) ----
install.packages("userfriendlyscience")
library(userfriendlyscience)
base_data_final_post <- base_data_final %>% dplyr::filter(! (is.na(t_glu0)))
userfriendlyscience::posthocTGH(y = base_data_final_post$t_glu0,
                                x = base_data_final_post$bmi_gr,
                                method = "games-howell")
```

결과**① 비만도 집단별 혈당 평균 확인**

```
> psych::describeBy(base_data_final$t_glu0, base_data_final$bmi_gr)
```

Descriptive statistics by group

```
group: 저체중
  vars  n  mean  sd median trimmed  mad min max range skew kurtosis  se
x1     1 187  94.72  84.02     88   89.02 10.38  62 356   294  5.3   33.47  2.49
-----
group: 정상체중
  vars  n  mean  sd median trimmed  mad min max range skew kurtosis  se
x1     1 2940  91.89 18.01     89   89.54  8.9  49 296   247  4.66   34.42  0.33
-----
group: 과체중
  vars  n  mean  sd median trimmed  mad min max range skew kurtosis  se
x1     1 2611  94.42 19.12     91   91.67 10.38  53 296   243  3.76   23.87  0.37
-----
group: 비만
  vars  n  mean  sd median trimmed  mad min max range skew kurtosis  se
x1     1 3982  98.24 22.99     93   94.54 11.86  58 394   336  3.98   26.53  0.36
```

② 집단 간 등분산성 검토

```
> car::leveneTest(t_glu0 ~ bmi_gr, data=base_data_final)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group   3 19.095 2.423e-12 ***
9716
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

③ 집단 간 평균 차이 검정 (Welch' ANOVA)

```
> #### welch's ANOVA ####
> welch <- oneway.test(t_glu0 ~ bmi_gr, data=base_data_final, var.equal=FALSE)
> welch
```

One-way analysis of means (not assuming equal variances)

```
data: t_glu0 and bmi_gr
F = 55.343, num df = 3.0, denom df = 870.9, p-value < 2.2e-16
```

④ 집단 간 차이가 유의한 경우에 한하여, 어느 집단 간에 차이가 있는지 사후 검정
이질적 분산 & 표본크기가 상이한 경우, 대표적으로 사용하는 'games-howell' 검정법을 실시한 결과

	diff	ci.lo	ci.hi	t	df	p
정상체중-저체중	-2.83	-9.3	3.7	1.13	193	.67
과체중-저체중	-0.31	-6.8	6.2	0.12	195	1
비만-저체중	3.52	-3.0	10.0	1.40	194	.5
과체중-정상체중	2.52	1.2	3.8	5.04	5379	<.01
비만-정상체중	6.35	5.1	7.6	12.87	6896	<.01
비만-과체중	3.82	2.5	5.2	7.32	6232	<.01

참고

① 만약 등분산성을 만족한다면, 집단 간 평균 차이 검정 (ANOVA)

```
> bmi.aov <- aov(t_glu0 ~ bmi_gr, data=base_data_final)
> summary(bmi.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
bmi_gr	3	70552	23517	54.06	<2e-16 ***
Residuals	9716	4226480	435		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
224 observations deleted due to missingness

② 만약 등분산성 만족 & 표본크기가 상이할 경우, 사후검정(Scheffe) 실시

```
> # 등분산성을 만족하고 표본크기가 상이할 경우, 사후 검정(Scheffe) ----
> options(digits = 2 )
> library(agricolae)
> agricolae::scheffe.test(bmi.aov, "bmi_gr", group=F, console=T)
```

Study: bmi.aov ~ "bmi_gr"

Scheffe Test for t_glu0

Mean Square Error : 435

bmi_gr, means

	t_glu0	std	r	Min	Max
과체중	94	19	2611	53	296
비만	98	23	3982	58	394
저체중	95	34	187	62	356
정상체중	92	18	2940	49	296

Alpha: 0.05 ; DF Error: 9716
Critical Value of F: 2.6

Comparison between treatments means

	Difference	pvalue	sig	LCL	UCL
과체중 - 비만	-3.82	0.0000	***	-5.29	-2.4
과체중 - 저체중	-0.31	0.9981		-4.72	4.1
과체중 - 정상체중	2.52	0.0002	***	0.96	4.1
비만 - 저체중	3.52	0.1662		-0.85	7.9
비만 - 정상체중	6.35	0.0000	***	4.93	7.8
저체중 - 정상체중	2.83	0.3571		-1.57	7.2

결과 해석

- 비만도 수준별 혈당 수준의 평균값을 살펴보면 저체중 94.72(kg/m²), 정상체중 91.89(kg/m²), 과체중 94.42(kg/m²), 비만이 98.24(kg/m²) 이었다.
- leveneTest 함수를 이용하여 네 집단 간의 분산이 같다는 귀무가설을 검정한 결과, 유의확률이 유의수준 0.05보다 현저히 작으므로 네 집단 간 분산이 같다는 귀무가설을 기각할 수 있다.
- 집단 간 등분산성을 만족하지 못하므로 집단별 혈당 수준의 평균값의 차이를 검정하는 방법으로 Welch's ANOVA를 실시한 결과, 검정통계량 F값이 55.34이고, 유의확률은 유의수준 0.05보다 현저히 작으므로 4개의 집단(저체중, 정상체중, 과체중, 비만) 중 차이가 나는 집단이 적어도 하나는 존재한다고 할 수 있다.
- 어느 집단 간에 차이가 있는지 사후검정하기 위해 'games-howell'방법에 의한 다중비교를 수행한 결과, 과체중-비만, 과체중-정상체중, 비만-정상체중 집단 간에 평균 혈당 수준값이 통계적으로 유의미한 차이가 있음을 알 수 있다.



다중 비교

세 집단 이상의 평균 비교에서 어느 집단(수준)간에 차이가 나는지 알고 싶을 때 사용하는 사후검정법이 다중 비교(multiple comparison)이다. 세 집단 이상의 평균 비교를 't-검정'처럼 두 집단씩 짝을 지어 비교하게 되면, 애초에 정한 유의수준보다 제 1종 오류(참을 거짓으로 판정하는 오류)가 훨씬 커지게 된다. 따라서 유의수준을 초기 설정한 크기로 유지하면서 모든 두 집단 짝의 평균을 동시에 비교할 수 있도록 고안된 것이 다중비교이며, 분석 기법에는 크게 모수적 기법과 비모수적 기법으로 나뉘는데, 본 가이드북에서는 모수적 기법만 소개하기로 한다. 모수적 기법에는 '등분산여부'와 비교 집단의 '표본크기' 동일 여부에 따라 사용할 수 있는 사후 검정 방법이 달라지게 되며, 어떠한 사후 검정 방법을 사용하느냐에 따라 조금씩 다른 결과가 나올 수 있음을 유념하고 사용해야 한다.

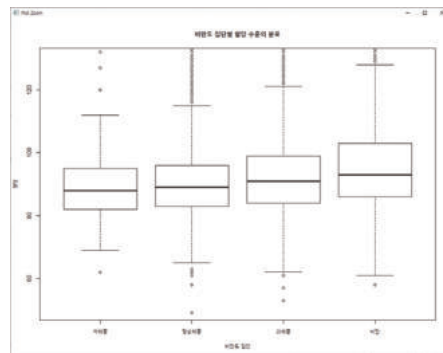
다중 비교	등분산여부	표본 크기	분석 기법
	등분산	동일	Tukey, Duncan, Dunnett 등
		상이	Bonferroni, Scheffe, Tukey, Dunnett 등
	이질적 분산	상이	Games Howell, Dunnett T3, Dunnett C 등

집단 평균 비교 - 시각화

Boxplot과 Histogram을 사용하여 시각적으로 집단별 평균을 비교할 수 있으며 그 방법은 다음과 같다.

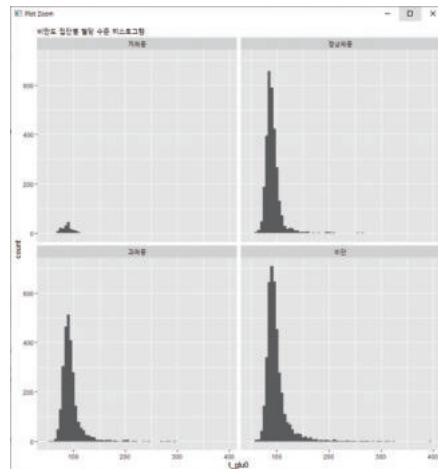
• Boxplot (상자그림)

```
boxplot(t_glu0 ~ bmi_gr,
        data = base_data_final,
        main = "비만도 집단별 혈당 수준의 분포",
        xlab = "비만도 집단",
        ylab = "혈당",
        ylim = c(50, 130))
```



• Histogram (히스토그램)

```
install.packages("ggplot2")
library(ggplot2)
ggplot(base_data_final,
        aes(x = t_glu0)) +
  facet_wrap(bmi_gr~.) +
  geom_histogram(binwidth = 5) +
  ggtitle("비만도 집단별 혈당 수준 히스토그램")
```



5-7. 선형 회귀분석

회귀분석은 독립변수와 종속변수 사이의 관계를 모형화하는 통계적 분석 방법이다. 선형 회귀분석은 독립변수와 종속변수의 관계가 직선적이라는 가정하에, 선형함수 관계를 기울기와 절편으로 표현한다. 즉, 독립변수의 값이 1 단위 증가함에 따라 종속변수가 어느 정도 증가하는지를 분석할 수 있다. 독립변수가 한 개일 경우 단순 선형 회귀분석, 독립변수가 두 개 이상일 경우 다중 선형 회귀분석으로 검정한다. 선형 회귀분석은 종속변수와 독립변수 간 선형성을 만족하고 오차항의 정규성/독립성/등분산성 만족을 기본적으로 가정한다.

예제

체질량 지수(BMI)가 혈당에 미치는 영향 분석하기 - 단순 선형 회귀분석

I. 모형 설정

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

여기서 ε_i 은 오차로, 오차들은 서로 독립이며 동일한 분포 $N(0, \sigma^2)$ 를 따른다.

II. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
t_bmi	체질량지수(BMI)	() kg/m ²	연속형
t_glu0	혈액_Glucose	() mg/dL	연속형

III. R을 이용한 통계분석

```
# 단순 선형 회귀 ----
reg.simple <- lm(t_glu0 ~ t_bmi, data = base_data_final)
summary(reg.simple)
# 산점도 및 회귀식 ----
plot(t_glu0 ~ t_bmi, data = base_data_final)
abline(coef(reg.simple))
```

결과

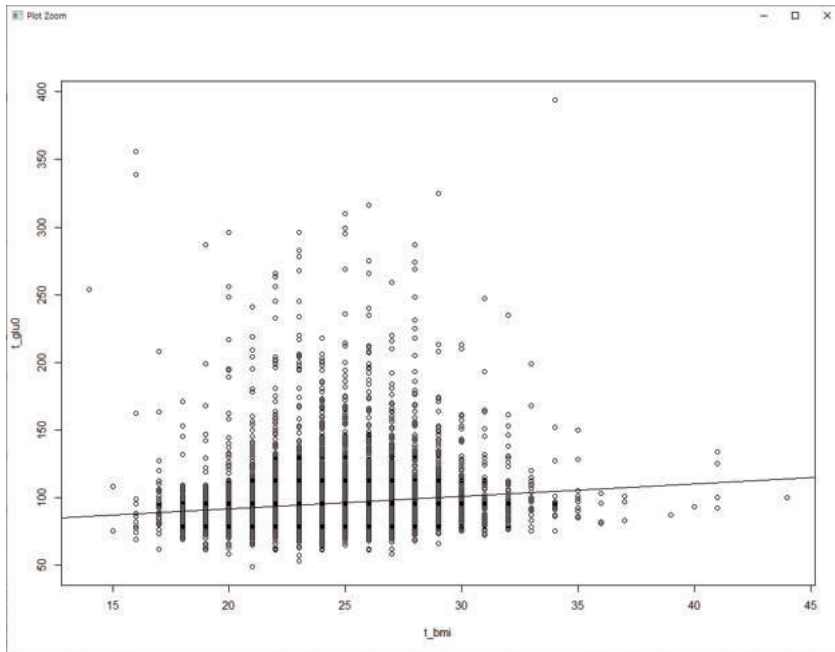
```
> summary(reg.simple)

Call:
lm(formula = t_glu0 ~ t_bmi, data = base_data_final)

Residuals:
    Min       1Q   Median       3Q      Max
-43.404 -10.338  -4.077   3.940  289.446

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.77639    1.71576   42.42  <2e-16 ***
t_bmi         0.93463    0.07089   13.18  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.84 on 9718 degrees of freedom
(224 observations deleted due to missingness)
Multiple R-squared:  0.01757,    Adjusted R-squared:  0.01747
F-statistic: 173.8 on 1 and 9718 DF,  p-value: < 2.2e-16
```



결과 해석

- 회귀모형의 적합성을 살펴보기 위해 F-검정을 실시한 결과, 유의확률이 유의수준 0.05보다 현저히 작으므로 회귀식이 상당한 의미가 있다고 할 수 있으며, 'R-Squared' 값을 통해 회귀모형의 설명력이 1.76%로 혈당(t_glu0)의 변동은 BMI(t_bmi)에 의해 1.76% 설명됨을 알 수 있다.
- 추정된 회귀계수의 유의확률은 유의수준 0.05보다 현저히 작아 귀무가설(회귀계수=0)을 기각해 통계적으로 유의하므로 다음과 같이 회귀식을 추정할 수 있다.
- 추정 회귀식 : $t_glu0 = 72.78 + 0.93 * t_bmi$, 추정 회귀식으로부터 BMI(t_bmi)가 1단위(kg/m^2) 증가할때 혈당(t_glu0)이 0.93mg/dL만큼 증가하고 이는 통계적으로 유의하다'라고 해석한다.

예제

연령과 성별을 보정한 체질량 지수(BMI)가 혈당에 미치는 영향 분석하기
- 다중 선형 회귀분석

I. 모형 설정

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i, \quad i = 1, 2, \dots, n$$

여기서 ε_i 은 오차로, 오차들은 서로 독립이며 동일한 분포 $N(0, \sigma^2)$ 를 따른다.

II. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
t_age	연령	만 () 세	연속형
t_sex	성별	1=남자, 2=여자	범주형
t_bmi	체질량지수(BMI)	() kg/m ²	연속형
t_glu0	혈액_Glucose	() mg/dL	연속형

III. R을 이용한 통계분석

회귀분석에서는 독립변수 중 연속형이 아닌 범주형 변수가 섞여 있다면 그 변수를 더미변수로 변환하여 처리를 해주어야 한다. 여기서 더미변수란 0과 1로 구성된 이항변수로 0으로 표시되는 범주를 기준범주라 하며, 이는 범주간 비교의 기준이 된다. 만일 범주형 변수의 범주가 n개일 경우, 생성해야할 더미변수 n-1개 이다. R에서는 더미변수로 변환하지 않아도 자동으로 기준 범주를 설정하여 번거로운 작업을 피할 수 있다. 하지만 R이 자동 설정한 기준 범주가 연구자의 의도와 다를 경우, 연구자가 원하는 특정 범주를 기준으로 설정하는 방법을 익혀두어야 한다.

다음의 예제를 통해서 범주형 변수의 기준 범주가 자동으로 처리된 경우와 연구자가 임의로 지정하여 분석하는 방법을 익히고, 그 결과를 비교해보도록 하자. 참고로 결과 비교가 용이하기 위해 기준이 되는 범주를 동일하게 설정하였다.

```
# 다중 선형 회귀 ----
# 더미변수 처리하지 않은 경우 ----
reg.fit <- lm(t_glu0 ~ t_bmi + t_age + t_sex, data = base_data_final)
summary(reg.fit)

# 더미변수 처리한 경우 ----
base_data_final_dummy <- transform(base_data_final,
                                   sex_dummy = as.factor(ifelse(t_sex == "여자", 1, 0)))
reg.dummy <- lm(t_glu0 ~ t_bmi + t_age + sex_dummy, data = base_data_final_dummy)
summary(reg.dummy)

# 다중공선성 진단 ----
install.packages("car")
options(digits = 3)
library(car)
car::vif(reg.fit) # VIF>10이면, 다중공선성 의심
```

결과

〈자동 : 더미변수 처리하지 않은 경우〉

```
> summary(reg.fit)

Call:
lm(formula = t_glu0 ~ t_bmi + t_age + t_sex, data = base_data_final)

Residuals:
    Min       1Q   Median       3Q      Max
-48.202 -10.244  -3.619   4.013  294.371

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.83631    2.09021   30.062 <2e-16 ***
t_bmi        0.79052    0.07051   11.212 <2e-16 ***
t_age        0.30125    0.02408   12.510 <2e-16 ***
t_sex여자   -4.24362    0.44098  -9.623 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.57 on 9716 degrees of freedom
(224 observations deleted due to missingness)
Multiple R-squared:  0.04368, Adjusted R-squared:  0.04338
F-statistic: 147.9 on 3 and 9716 DF, p-value: < 2.2e-16
```

〈임의 : 더미변수 처리한 경우〉

```
> summary(reg.dummy)

Call:
lm(formula = t_glu0 ~ t_bmi + t_age + sex_dummy, data = base_data_final_dummy)

Residuals:
    Min       1Q   Median       3Q      Max
-48.202 -10.244  -3.619   4.013  294.371

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.83631    2.09021   30.062 <2e-16 ***
t_bmi        0.79052    0.07051   11.212 <2e-16 ***
t_age        0.30125    0.02408   12.510 <2e-16 ***
sex_dummy1   -4.24362    0.44098  -9.623 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.57 on 9716 degrees of freedom
(224 observations deleted due to missingness)
Multiple R-squared:  0.04368, Adjusted R-squared:  0.04338
F-statistic: 147.9 on 3 and 9716 DF, p-value: < 2.2e-16
```

```
> car::vif(reg.fit) # VIF>10이면, 다중공선성 의심
t_bmi t_age t_sex
1.02 1.01 1.01
```

결과 해석

- 독립변수 중 범주형 변수인 성별에 대해 더미변수 처리하지 않은 경우와 처리한 경우의 회귀분석 결과를 비교해본 결과, 동일한 결과가 산출되므로 연구자가 편한 방법을 권장한다. 다만 R이 자동으로 설정하는 기준범주가 아닌 다른 범주를 기준으로 설정하고자 할 경우, 후자의 방법을 이용하면 된다.
- 회귀모형의 적합도 검정 결과(F-검정) 유의확률이 유의수준 0.05보다 현저히 작으므로 회귀식이 상당한 의미가 있다고 할 수 있으며, 추정된 모든 회귀계수의 유의확률이 유의수준 0.05보다 현저히 작아 회귀계수가 통계적으로 유의하므로 다음과 같이 회귀식을 추정할 수 있다.
남자의 회귀식 : $t_glu0 = 62.84 + 0.79 * t_bmi + 0.3 * t_age$
여자의 회귀식 : $t_glu0 = 62.84 + 0.79 * t_bmi + 0.3 * t_age - 4.24$
- 추정된 회귀식 해석
 - 성별과 연령을 통제(보정)했을 경우 'BMI가 1단위(kg/m²) 증가할 때 혈당이 0.79mg/dL만큼 증가'하고 이는 통계적으로 유의하다.
 - 성별과 체질량지수(BMI)를 통제(보정)했을 경우 '나이가 1세 증가할 때 혈당이 0.3mg/dL만큼 증가'하고 이는 통계적으로 유의하다.
 - 나이와 체질량지수(BMI)를 통제(보정)했을 경우 '남자(기준 범주)에 비해 여자의 혈당이 4.24mg/dL만큼 낮고, 이는 통계적으로 유의하다.
- 'Adjusted R-Squared' 값을 통해 회귀모형의 설명력이 4.34%로 혈당의 변동은 3개의 설명변수(BMI, 연령, 성별)에 의해서 4.34% 설명된다.
- 독립변수에서 분산팽창요인(Variance Inflation)이 10이하로 나타나 다중공선성의 위험은 없는 것으로 보인다.



회귀분석 - 기본가정 검토

회귀분석은 기본적으로 오차항의 정규성, 독립성, 등분산성을 가정하므로 이를 만족하는지 살펴보아야 한다.

• 오차항의 독립성

Durbin-Watson 값은 0~4의 분포를 가지며, 2에 가까울수록 독립이라고 볼 수 있으며, 만족하지 않을 경우 누락된 독립변수를 모형에 포함하거나 종속변수와 독립변수를 모두 변수변환 한다.

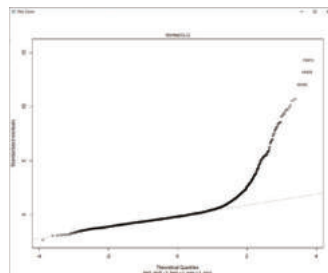
```
options(digits = 4)
car::durbinWatsonTest(reg.fit)
```

```
> car::durbinWatsonTest(reg.fit) # 독립성
lag Autocorrelation D-W Statistic p-value
1 0.00109 1.998 0.88
Alternative hypothesis: rho != 0
```

• 오차항의 정규성

잔차에 대한 정규확률그림(Normal Q-Q)에서 잔차의 분포가 일직선이면 정규분포와 비슷하다고 판단하며, 만족하지 않을 경우 특이값(outlier)을 제거하거나 변수변환 한다.

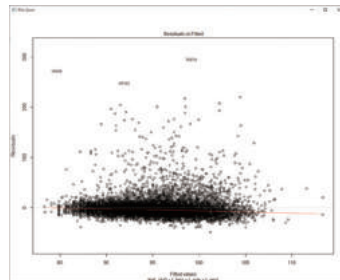
```
plot(reg.fit, which = 2)
```



• 오차항의 등분산성

적합값에 대한 잔차 그림(Residuals vs Fitted)에서 특정 패턴이 존재하지 않는다면 등분산성 만족한다고 판단하고, 만약 만족하지 않을 경우 대수변환/역변환 등의 변수변환 한다.

```
plot(reg.fit, which = 1)
```



회귀분석 - 다중공선성

다중 선형 회귀분석에서 독립 변수들끼리의 영향은 존재할 수밖에 없으나, 그 영향이 클 경우 다중공선성이 발생한다. 다중공선성이 존재할 경우 그 징후는 무엇이며, 이를 진단하고 해결 방안을 살펴보도록 하자.

• 다중공선성 징후 및 진단과 해결방안

징후	독립변수들이 높은 상관관계를 보임 이론적으로 종속변수와 상관관계가 높을 것으로 예견되는 독립변수의 회귀계수가 유의하지 않음 독립변수를 추가 또는 제거했을 때, 회귀계수의 변화가 큰 경우
진단	분산팽창요인(VIF)가 10이상이면 다중공선성 의심
해결 방안	상관성이 높은 독립변수를 제거(변수 선택법 활용) 주성분회귀 등 수행

회귀분석 - 변수 선택

앞서 살펴본 다중공선성을 해결하고 최적의 모형 적합을 위해서는 여러 개의 독립변수 중 적절한 변수를 선택(제거)하는 과정이 필요한데, 이러한 과정을 변수 선택이라고 부른다. 변수 선택에는 3가지(전진, 후진, 단계적) 방법이 있다.

• 3가지 방법

전진 선택법	독립변수 0개부터 시작하여 가장 유의한 변수부터 하나씩 추가하는 방법
후진 제거법	모든 독립변수를 모형에 넣고, 필요 없는 변수부터 하나씩 제거하는 방법
단계적 방법	독립변수 0개부터 추가했다, 제거했다를 반복하는 방법

• 변수 선택 (예제)

```
# step() 함수에서 변수 선택법은 단계적인 방법 사용
reg.fit.step <- step(reg.fit, direction="both") # 전진(forward), 후진(backward), 단계적(both)
```

• 변수 선택 후 회귀모형 비교

설정한 회귀모형에서 만약 변수를 제거할 경우, 제거된 변수의 회귀모형과 기존 회귀모형을 비교하여 제거된 변수의 회귀모형에 대한 기여도를 평가할 수 있다.

```
# anova() 함수 사용
anova(reg.fit, reg.fit.step)
```

5-8. 로지스틱 회귀분석

로지스틱 회귀분석은 연속형이나 범주형 독립변수와 이분화된 종속변수의 관계를 알아보고자 할 때 이용된다. 주로 특정 질병의 위험인자가 무엇인지 추정하고자 할 때 사용하며, 기준 그룹 대비 특정 그룹의 오즈비(Odds ratio, OR)를 구할 수 있다. 독립변수가 하나인 경우 단순 로지스틱 회귀분석으로, 독립변수가 두 개 이상일 경우 다중 로지스틱 회귀분석으로 검정한다.

예제

비만도에 따라 당뇨병 유병 위험에 차이가 있는지 분석하기 - 단순 로지스틱 회귀분석

I. 모형 설정

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i}, \quad p_i = p(y=1)$$

II. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
dm	당뇨병 유병 여부	0=정상, 1=당뇨병	범주형

III. R을 이용한 통계분석

```
# 기준범주 변경 : releval() 함수 이용 ----
base_data_final_level <- base_data_final
base_data_final_level$bmi_gr <- releval(base_data_final_level$bmi_gr, ref = "정상체중")

# 단순 로지스틱 회귀 ----
logit.simple <- glm(dm ~ bmi_gr, data = base_data_final_level, family = "binomial")
summary(logit.simple)

options(digits = 1)
exp(coef(logit.simple))      # 오즈비
exp(confint.default(logit.simple)) # 95% 신뢰구간
```


결과

```
> summary(logit.simple)

Call:
glm(formula = dm ~ bmi_gr, family = "binomial", data = base_data_final_level)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5042  -0.5042  -0.4310  -0.3333   2.4160

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.86291    0.08078  -35.442  < 2e-16 ***
bmi_gr저체중  0.40047    0.28084   1.426    0.154
bmi_gr과체중  0.53338    0.10553   5.054 4.32e-07 ***
bmi_gr비만   0.86460    0.09414   9.185  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6044.8  on 9943  degrees of freedom
Residual deviance: 5950.5  on 9940  degrees of freedom
AIC: 5958.5

Number of Fisher Scoring iterations: 5

> exp(coef(logit.simple)) # 오즈비
      (Intercept) bmi_gr저체중 bmi_gr과체중 bmi_gr비만
           0.06           1.49           1.70           2.37
>
> exp(confint.default(logit.simple)) # 95% 신뢰구간
            2.5 % 97.5 %
(Intercept)  0.05   0.07
bmi_gr저체중  0.86   2.59
bmi_gr과체중  1.39   2.10
bmi_gr비만   1.97   2.86
> |
```

결과 해석

- 비만도가 정상체중인 그룹 대비 과체중 그룹의 당뇨병 유병의 오즈비(OR) 및 95% CI는 1.70(1.39-2.10), 비만인 그룹은 2.37(1.97-2.86)로 나타났다. 이를 통해 과체중 그룹은 정상체중인 그룹에 비하여 당뇨병 유병의 오즈가 1.7배 더 높고, 비만인 그룹은 정상체중인 그룹에 비하여 당뇨병 유병의 오즈가 2.37배 높았으며, 이는 통계적으로 유의하였다.

예제

성별과 연령을 보정했을 때 비만도에 따라 당뇨병 유병 위험에 차이가 있는지 분석하기
- 다중 로지스틱 회귀분석

I. 모형 설정

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}, \quad p_i = p(y=1)$$

II. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
t_age	연령	만 ()세	연속형
t_sex	성별	1=남자, 2=여자	범주형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
dm	당뇨병 유병 여부	0=정상, 1=당뇨병	범주형

III. R을 이용한 통계분석

```
# 다중 로지스틱 회귀 ----
logit.multi <- glm(dm ~ bmi_gr + t_age + t_sex, data = base_data_final_level, family = "binomial")
summary(logit.multi)
exp(coef(logit.multi))          # 오즈비
exp(confint.default(logit.multi)) # 95% 신뢰구간
```

결과

```
> summary(logit.multi)

Call:
glm(formula = dm ~ bmi_gr + t_age + t_sex, family = "binomial",
    data = base_data_final_level)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.016  -0.481  -0.371  -0.280   2.797

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9292     0.2529  -23.44 < 2e-16 ***
bmi_gr저체중   0.2827     0.2877   0.98   0.33
bmi_gr과체중   0.4564     0.1070   4.27 2.0e-05 ***
bmi_gr비만     0.7301     0.0955   7.65 2.1e-14 ***
t_age          0.0608     0.0041  14.83 < 2e-16 ***
t_sex여자     -0.3956     0.0719  -5.51 3.7e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6044.8 on 9943 degrees of freedom
Residual deviance: 5688.6 on 9938 degrees of freedom
AIC: 5701

Number of Fisher Scoring iterations: 5
```

```

> exp(coef(logit.multi)) # 오즈비
(Intercept) bmi_gr저체중 bmi_gr과체중 bmi_gr비만 t_age t_sex여자
0.003 1.327 1.578 2.075 1.063 0.673
> exp(confint.default(logit.multi)) # 95% 신뢰구간
2.5 % 97.5 %
(Intercept) 0.002 0.004
bmi_gr저체중 0.755 2.332
bmi_gr과체중 1.280 1.946
bmi_gr비만 1.721 2.502
t_age 1.054 1.071
t_sex여자 0.585 0.775

```

결과 해석

- 성별과 연령을 통제(보정)하였을 때, 체질량지수(BMI) 기준 비만도가 정상체중인 그룹에 비하여 저체중인 그룹의 오즈비(OR) 및 95% CI는 1.33(0.76-2.33), 과체중인 그룹은 1.58(1.28-1.95), 비만인 그룹은 2.08(1.72-2.50)로 나타났다. 이를 통해 과체중 그룹은 정상체중인 그룹에 비하여 당뇨병 유병의 오즈가 1.58배 더 높고, 비만인 그룹은 정상체중인 그룹에 비하여 당뇨병 유병의 오즈가 2.08배 높았으며, 이는 통계적으로 유의하였다.



오즈비(Odds ratio, OR)

오즈(Odds)와 오즈비(Odds ratio)의 개념을 다음의 예제를 통해 알아보도록 하자.

예제

구분	폐암	
	유	무
흡연자	30	70
비흡연자	1	99

예제를 통해 흡연자와 비흡연자가 폐암에 걸릴 확률을 먼저 구해보면 다음과 같다.

- 흡연자가 폐암에 걸릴 확률(R1)은 $30/100 = 0.3$
- 비흡연자가 폐암에 걸릴 확률(R2)은 $1/100 = 0.01$

반면에 오즈는 위의 확률과는 조금 다른 개념으로, 흡연자나 비흡연자로 구분했을 때, 각각이 폐암에 걸릴 확률과 폐암에 걸리지 않을 확률의 '비율'이다.

- 흡연자가 폐암에 걸릴 오즈(odds1)는 $30/70 \approx 0.4285$
- 비흡연자가 폐암에 걸릴 오즈(odds2)는 $1/99 \approx 0.0101$

위의 두 오즈의 비가 오즈비(OR)이며, 계산해보면 다음과 같다.

- $OR = odds1/odds2 = \text{흡연자가 폐암에 걸릴 오즈} / \text{비흡연자가 폐암에 걸릴 오즈} \approx 42.43$

이를 해석하면, "흡연자가 폐암에 걸릴 오즈는 비흡연자가 폐암에 걸릴 오즈보다 42.43배 높다."라고 할 수 있으나, 그 의미를 직관적으로 파악하는데 어려움이 있다.

상대위험도(Relative risk, RR)

상대위험도(RR)란 위험인자가 있는 경우 어떤 사건이 발생할 확률과 위험인자가 없는 경우 어떤 사건이 발생할 확률의 비이며, 상대위험도가 클수록 위험인자와 사건간의 연관성이 크다는 것을 의미한다.

위의 예제를 적용해 상대위험도(RR)를 산출하면 다음과 같다.

- $RR = \text{흡연자가 폐암에 걸릴 확률} / \text{비흡연자가 폐암에 걸릴 확률} = 30$

이를 해석하면, "흡연이라는 위험인자에 노출된 사람은 그렇지 않은 사람에 비해 폐암(사건)에 걸릴 확률이 30배 높다."라고 할 수 있다. 이처럼 상대위험도는 직관적인 해석이 가능해 논문에서도 잘 쓰이며, 특히 코호트 연구에서 잘 활용된다.

한국인유전체역학조사사업 (KoGES)
Korean Genome and Epidemiology Study

KoGES 데이터 분석 가이드북

R편

1장 소개

2장 역학자료 소개

3장 기반조사 자료 분석하기

4장 추적조사 자료 분석하기

책
파

KoGES 데이터 분석 가이드북
[R편]

Korean Genome and Epidemiology Study

4장.

한국인유전체역학조사사업(KoGES) 추적조사 자료 분석하기

1. 추적조사 교육용 데이터 이해하기
2. 자료 불러오기
3. 자료 결합하기
4. 자료 분석 준비하기
5. 자료 분석하기

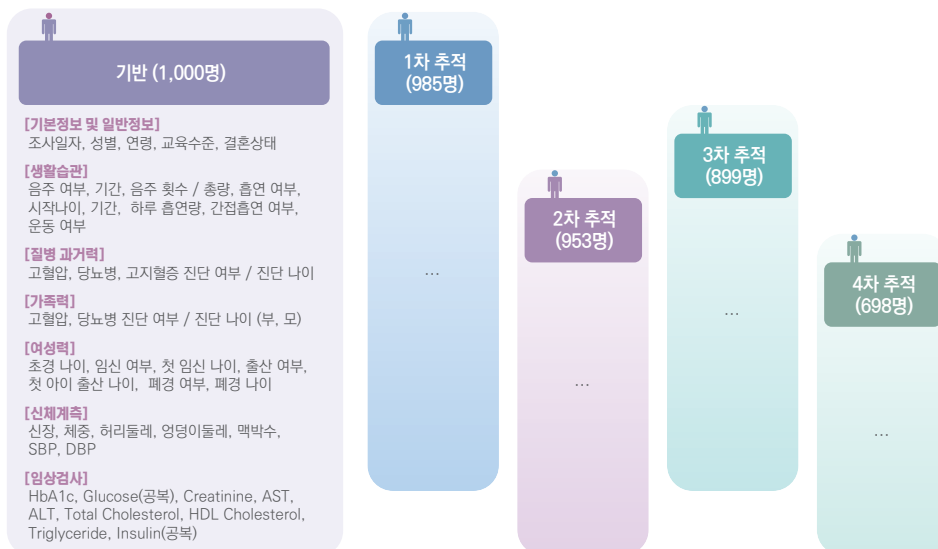
4장.

한국인유전체역학조사사업(KoGES) 추적조사 자료 분석하기

1. 추적조사 교육용 데이터 이해하기

1-1. 자료 구성

KoGES 추적조사 교육용 데이터는 CSV 데이터로 제공 된다. 기반조사 ~ 4차 추적조사까지 참여한 대상자 1,000명으로 이루어져 있으며, 변수는 기반조사 63개, 1~4차 추적조사 54개로 구성되어 있다.



| 그림 18 | KoGES 추적조사 교육용 데이터 구성

1-2. 코드북

KoGES 추적조사 교육용 데이터의 코드북은 공개 자료 코드북과 동일하게 테이블명, 변수명, 변수 설명, 변수값 (코드) 설명, 변수타입, 통합 설문지로 구성되어있다.

KoGES 추적조사 교육용데이터(기반)									
변수내용							통합 설문지		
일련 번호	데이터명(영문)	데이터명	변수명	변수설명		변수 유형			
0	Full_Survey_Base	일반정보	000_1st	개인식별자		(06666=>조사대상, 77777=>제당검진, 99999=>직장/무종교/비종교)	변수 유형	ID	■ 개인식별자
1	Full_Survey_Base	일반정보	000_001a_1st00	단위 대위번호		F01, F02, F03, F04, F05, F06, F07, F08, F09, F10, F11, F12, F13, F14, F15, F16, F17, F18, F19	일자리	날짜	■ 단위 대위번호
2	Full_Survey_Base	일반정보	001_001a	KoGES 추적조사 교육용 데이터 코드북(4차 추적)					
변수내용							통합 설문지		
일련 번호	데이터명(영문)	데이터명	변수명	변수설명		변수 유형			
3	Full_Survey_Base	일반정보	000_001a	개인식별자		(06666=>조사대상, 77777=>제당검진, 99999=>직장/무종교/비종교)	변수 유형	ID	■ 개인식별자
4	Full_Survey_Base	일반정보	001_001a	단위 대위번호		F01, F02, F03, F04, F05, F06, F07, F08, F09, F10, F11, F12, F13, F14, F15, F16, F17, F18, F19	일자리	날짜	■ 단위 대위번호
5	Full_Survey_Base	일반정보	001_001a	조사일		999999	날짜	■ 조사일	
				성별		1=남자, 2=여자	문자열	■ 성별	
				연령		1=10대, 2=20대, 3=30대, 4=40대, 5=50대, 6=60대, 7=70대, 8=80대, 9=90대, 10=100대	문자열	■ 연령	
6	Full_Survey_Base	일반정보	001_001a	종교		1=불교, 2=기독교, 3=이슬람교, 4=유교, 5=기타, 6=없음, 7=기타	일자리	■ 종교	
7	Full_Survey_Base	일반정보	001_001a	직업		1=관리직, 2=사무직, 3=판매직, 4=서비스직, 5=농림어업, 6=제조업, 7=건설업, 8=서비스업, 9=기타, 10=없음	일자리	■ 직업	

그림 19 | KoGES 추적조사 교육용 데이터 코드북

2. 자료 불러오기

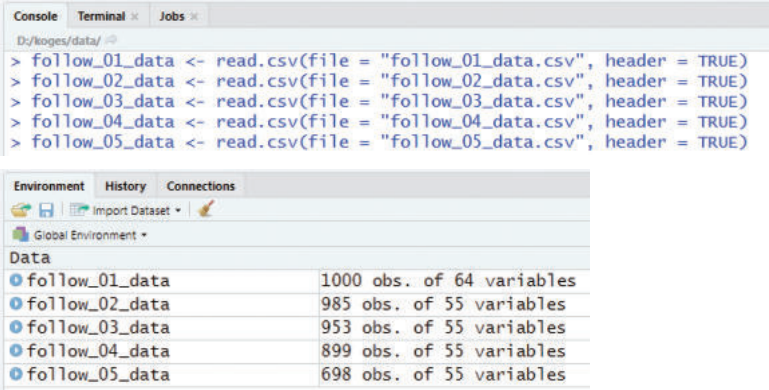
2-1. 자료 불러오기(CSV 파일)

파일을 불러오기 전, RStudio에서 생성한 데이터 및 분석 결과를 저장하거나 분석할 데이터 파일을 불러올 때 사용하는 폴더를 원하는 경로에 생성하고 이를 워킹 디렉토리로 설정해준다. 현재 설정된 워킹 디렉토리를 알고 싶다면 `getwd()` 함수를 이용하면 되고, 새롭게 설정하고 싶다면 `setwd()` 함수를 이용하면 된다. 본 가이드북에서 사용하는 KoGES 교육용 데이터는 [koges]라는 폴더 하위에 있는 [data] 폴더에 저장되어 있다. 예를 들어 [koges] 폴더 전체를 D 드라이브에 저장했다면, 데이터가 들어있는 파일의 경로는 "D:/koges/data" 가 되며, `setwd()` 함수를 이용해 다음의 코드 `setwd("D:/koges/data")`를 실행하면 워킹 디렉토리가 지정된다. 이처럼 워킹 디렉토리가 지정되었다면, 그 다음으로 `read.csv()` 함수를 이용해 바로 CSV파일을 R로 불러올 수 있다.

```
# 워킹 디렉토리 ----
getwd()           # 워킹 디렉토리 확인
setwd("D:/koges/data") # 워킹 디렉토리 설정

# CSV 파일 불러오기 ----
follow_01_data <- read.csv(file = "follow_01_data.csv", header = TRUE)
follow_02_data <- read.csv(file = "follow_02_data.csv", header = TRUE)
follow_03_data <- read.csv(file = "follow_03_data.csv", header = TRUE)
follow_04_data <- read.csv(file = "follow_04_data.csv", header = TRUE)
follow_05_data <- read.csv(file = "follow_05_data.csv", header = TRUE)
```

결과



The screenshot shows the RStudio interface. The Console pane displays the execution of R commands to load five CSV files. The Environment pane shows the resulting data objects in the Global Environment.

Object	Size / Description
follow_01_data	1000 obs. of 64 variables
follow_02_data	985 obs. of 55 variables
follow_03_data	953 obs. of 55 variables
follow_04_data	899 obs. of 55 variables
follow_05_data	698 obs. of 55 variables

2-2. 불러온 자료 확인하기

데이터를 불러와서 가장 먼저 해야할 일은 불러온 데이터의 구조와 관측값 수, 변수에 대한 정보(이름, 유형, 값) 등을 확인하는 일이다. R에서 제공하는 다양한 함수들을 이용해 불러온 자료의 정보를 확인할 수 있다.

```
# 불러온 자료 확인하기 ----
options(digits = 3)
str(follow_01_data) # 데이터셋 구조
View(follow_01_data) # 데이터셋 전체 새로운 창으로 보기
```

결과

```
> str(follow_01_data)
'data.frame': 1000 obs. of 64 variables:
 $ t00_id      : Factor w/ 1000 levels "K_FOLLOW_0001",...
 $ t00_data_class: Factor w/ 14 levels "F01","F03","F04",...
 $ t01_edate   : int  200412 200401 200309 200504 200402 2
 $ t00_sex     : int  1 1 1 2 1 1 2 2 1 2 ...
 $ t01_age     : int  56 40 52 60 49 50 48 48 42 62 ...
 $ t01_edu     : int  1 3 2 2 3 3 3 3 5 1 ...
 $ t01_marry   : int  2 2 2 2 2 2 2 2 3 ...
 $ t01_drink   : int  3 3 2 1 3 3 3 1 3 1 ...
 $ t01_drdrq   : int  4 4 3 77777 4 4 1 77777 4 77777 ...
 $ t01_takfq   : int  0 0 0 0 2 0 0 1 0 ...
 $ t01_takam   : num  77777 77777 77777 77777 77777 ...
 $ t01_ricefq  : int  0 0 0 0 0 0 0 0 0 ...
 $ t01_riceam  : num  77777 77777 77777 77777 77777 ...
 $ t01_winefq  : int  0 0 0 0 0 0 0 0 0 ...
 $ t01_wineam  : num  77777 77777 77777 77777 77777 ...
 $ t01_sojufq  : int  4 5 0 0 3 4 0 0 4 0 ...
 $ t01_sojuam  : num  3 24 77777 77777 7 ...
 $ t01_beerfq  : int  0 1 0 0 0 1 1 0 2 0 ...
 $ t01_beeram  : num  77777 10 77777 77777 77777 ...
 $ t01_hliqfq  : int  0 0 0 0 0 0 0 0 0 ...
 $ t01_hliqam  : num  77777 77777 77777 77777 77777 ...
 $ t01_hliqam  : num  77777 77777 77777 77777 77777 ...
 $ t01_smoke   : int  3 2 2 1 3 2 1 1 2 1 ...
 $ t01_smag    : int  20 20 21 77777 23 23 77777 77777 19
 $ t01_smdu    : num  40 20 15 77777 20 ...
 $ t01_snam    : num  20 40 20 77777 20 ...
 $ t01_psm     : int  2 1 1 99999 1 2 2 2 2 1 ...
```

t00_id	t00_data_class	t01_edate	t00_sex	t01_age	t01_edu	t01_marry
1	K_FOLLOW_0001	F05	200412	1	56	1
2	K_FOLLOW_0002	F19	200401	1	40	3
3	K_FOLLOW_0003	F05	200309	1	52	2
4	K_FOLLOW_0004	F05	200504	2	80	2
5	K_FOLLOW_0005	F19	200402	1	48	3
6	K_FOLLOW_0006	F19	200401	1	50	3
7	K_FOLLOW_0007	F05	200410	2	48	3
8	K_FOLLOW_0008	F19	200308	2	48	3
9	K_FOLLOW_0009	F19	200301	1	42	5
10	K_FOLLOW_0010	F01	200312	2	62	1
11	K_FOLLOW_0011	F10	200502	2	48	2
12	K_FOLLOW_0012	F17	200312	2	52	1
13	K_FOLLOW_0013	F10	200502	2	88	1
14	K_FOLLOW_0014	F05	200512	1	47	1
15	K_FOLLOW_0015	F05	200408	1	61	2
16	K_FOLLOW_0016	F19	200508	2	42	3
17	K_FOLLOW_0017	F19	200411	2	48	3

3. 자료 결합하기

KoGES 추적조사 교육용 데이터는 기반조사 자료와 1~4차 추적조사 자료가 각각의 데이터셋으로 구분되어 있다. 따라서 자료 분석을 위해서는 연구 목적에 따라 분석에 필요 변수를 선택하여 하나의 데이터 셋으로 결합(merge)하는 작업이 요구된다.

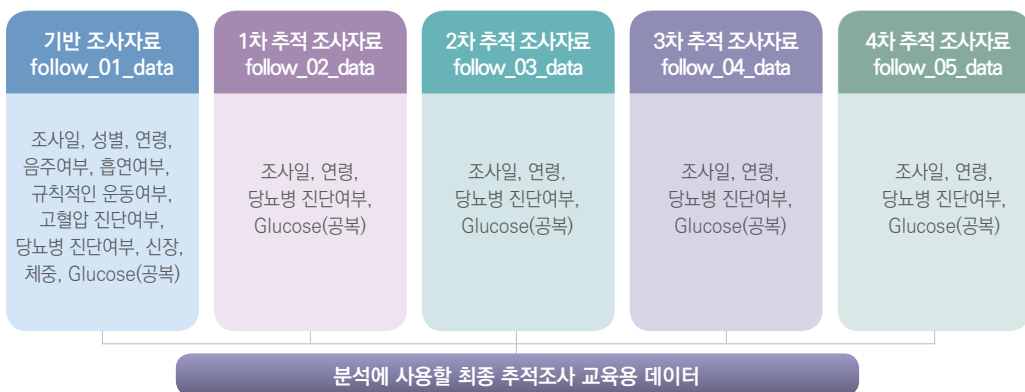
각각의 개별 테이블에는 공통적으로 참여자의 개인 식별번호인 'ID' 변수가 포함되어 있으며, 이를 기준으로 하나의 자료로 결합할 수 있다. 보통의 통계패키지에서는 결합 전 반드시 'ID' 변수로 정렬하는 작업이 필요하나, R에서는 사전 정렬 없이 바로 테이블 결합이 가능하다.

기반			1차 추적								기반			1차 추적		
ID	A1_HTN	A1_HTNAG	+	ID	A2_HTN	A2_HTNAG	=	ID	A1_HTN	A1_HTNAG	A2_HTN	A2_HTNAG				
NO_1	1	77777		NO_1	2	55		NO_1	1	77777	2	55				
NO_2	2	60		-	-	-		NO_2	2	60	-	-				
NO_3	2	58		NO_3	1	77777		NO_3	2	58	1	77777				
NO_4	1	77777		NO_4	2	63		NO_4	1	77777	2	63				
NO_5	1	77777		-	-	-		NO_5	1	77777	-	-				

| 그림 20 | KoGES 추적조사 데이터 조인키

예제

추적조사 교육용 데이터를 구성하는 5개 데이터 셋(기반~4차 추적)에 포함된 여러 변수 중 필요한 변수를 선택하여 하나의 데이터 셋으로 결합하기



| 그림 21 | 분석에 사용할 최종 추적조사 교육용 데이터

각 데이터셋에서 필요한 변수만 가져오기 ----

```
library(dplyr)
```

```
fw_select1 <- follow_01_data %>%
```

```
dplyr::select(t00_id, t00_sex, t01_age, t01_edate, t01_htn, t01_dm, t01_drink, t01_smoke, t01_exer,
  t01_height, t01_weight, t01_glu0)
```

```
fw_select2 <- follow_02_data %>% dplyr::select(t00_id, t02_age, t02_edate, t02_dm, t02_glu0)
```

```
fw_select3 <- follow_03_data %>% dplyr::select(t00_id, t03_age, t03_edate, t03_dm, t03_glu0)
```

```
fw_select4 <- follow_04_data %>% dplyr::select(t00_id, t04_age, t04_edate, t04_dm, t04_glu0)
```

```
fw_select5 <- follow_05_data %>% dplyr::select(t00_id, t05_age, t05_edate, t05_dm, t05_glu0)
```

자료 결합 ----

```
fw_merge1 <- left_join(fw_select1, fw_select2)
```

```
fw_merge2 <- left_join(fw_merge1, fw_select3)
```

```
fw_merge3 <- left_join(fw_merge2, fw_select4)
```

```
fw_merge4 <- left_join(fw_merge3, fw_select5)
```

```
follow_data <- fw_merge4
```

결합된 데이터 확인하기 ----

```
str(follow_data)      # 총 1000명의 대상자, 28개 변수 유형 확인
```

```
View(follow_data)     # 데이터셋 확인
```

결과

```
Console Terminal Jobs
RStudio
> # 결합된 데이터 확인하기 ----
> str(follow_data) # 총 1000명의 대상자, 28개 변수와 변수유형 확인
'data.frame': 1000 obs. of 28 variables:
 $ t00_id : chr "K_FOLLOW_0001" "K_FOLLOW_0002" "K_FOLLOW_0003" ...
 $ t00_sex : int 1 1 1 2 1 1 2 2 1 2 ...
 $ t01_age : int 56 40 52 60 49 50 48 48 42 62 ...
 $ t01_edate : int 200412 200401 200309 200504 200402 200401 2 ...
 $ t01_htn : int 1 1 2 1 1 1 1 1 2 1 ...
 $ t01_dm : int 1 1 1 1 1 1 1 1 2 1 ...
 $ t01_drink : int 3 3 2 1 3 3 3 3 1 1 ...
 $ t01_smoke : int 3 2 2 1 3 2 1 1 2 1 ...
 $ t01_exer : int 1 1 1 1 1 2 2 2 2 2 ...
 $ t01_height : int 159 169 165 165 166 174 158 157 168 148 ...
 $ t01_weight : int 50 94 63 70 69 77 61 62 74 63 ...
 $ t01_glu0 : int 82 130 83 89 95 93 95 270 96 81 ...
 $ t02_age : int 58 41 54 62 51 52 50 49 44 64 ...
 $ t02_edate : int 200609 200510 200505 200704 200601 200606 2 ...
 $ t02_dm : int 1 2 1 1 1 1 1 1 1 1 ...
 $ t02_glu0 : int 90 99999 90 106 124 81 94 141 104 90 ...
 $ t03_age : int 61 44 57 65 54 56 53 54 47 66 ...
```

	t00_id	t00_sex	t01_age	t01_edate	t01_htn	t01_dm	t01_drink	t01_smoke
1	K_FOLLOW_0001	1	56	200412	1	1	3	3
2	K_FOLLOW_0002	1	40	200401	1	1	3	2
3	K_FOLLOW_0003	1	52	200309	2	1	2	2
4	K_FOLLOW_0004	2	60	200504	1	1	1	1
5	K_FOLLOW_0005	1	49	200402	1	1	3	3
6	K_FOLLOW_0006	1	50	200401	1	1	3	2
7	K_FOLLOW_0007	2	48	200410	1	1	3	1
8	K_FOLLOW_0008	2	48	200308	1	2	1	1
9	K_FOLLOW_0009	1	42	200301	2	1	3	2
10	K_FOLLOW_0010	2	62	200312	1	1	1	1
11	K_FOLLOW_0011	2	46	200302	1	1	3	1
12	K_FOLLOW_0012	2	52	200312	1	1	3	1
13	K_FOLLOW_0013	2	69	200302	2	1	1	1
14	K_FOLLOW_0014	1	47	200312	1	1	3	3
15	K_FOLLOW_0015	1	61	200408	1	1	2	2

4. 자료 분석 준비하기

4-1. 기본코드 결측치 처리하기

KoGES 역학자료는 설문 문항의 '미상/무응답', '설문 문항 간의 상·하위 관계(해당없음)', '해당변수의 조사유무(조사안함)', '반복추적조사 통합자료의 경우, 추적조사 참여유무(추적조사 미참여)' 등을 구분하기 위하여, 기본코드가 적용되어 있으며, 구체적인 기본 코드는 다음과 같다.

표 9 | KoGES 기본코드의 종류와 정의

구분	코드명	코드	코드 정의		
결측	미상/무응답	99999	Null값(missing value) 또는 조사항목 상의 미상/무응답 값		
	해당없음	77777	조사항목에 대해 응답의 대상이 아닌 경우 예)		
			변수명	변수설명	변수값(코드)
			HTN	고혈압 과거력 - 진단여부 (1=아니오, 2=예)	1
	HTNAG	고혈압 과거력 - 처음 진단 나이	77777		
	조사안함	66666	특정 조사단위에 조사되지 않은 항목의 경우		
추적조사 미참여		55555	반복추적조사 통합자료에서 해당 차수의 조사에 참여하지 않은 경우		

기본코드를 처리하지 않고 분석할 경우 해당 값('99999', '77777', '66666', '55555')이 하나의 변수 값(관찰 값)으로 포함된 결과가 출력되므로, 분석 전 기본코드를 결측치로 처리해주는 작업이 선행되어야 한다.

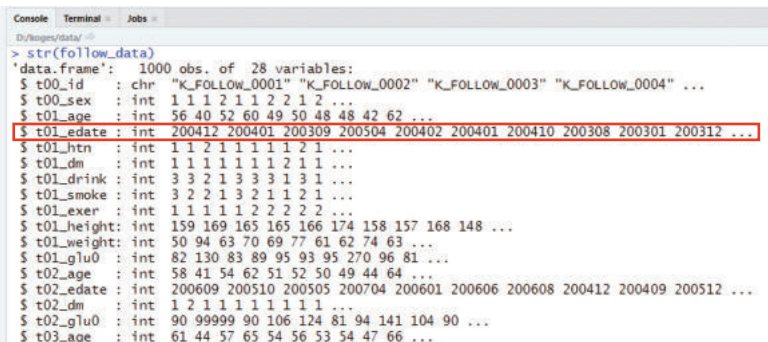
예제

추적조사 교육용 데이터에 포함된 모든 변수의 기본코드 결측치 처리하기

```
# 기본코드 결측치 처리 ----
follow_data_null <- follow_data
follow_data_null[follow_data_null == 99999 | follow_data_null == 77777 | follow_data_null == 66666 |
follow_data_null == 55555] <- NA
```

4-2. 변수 유형 변환하기

자료 분석에 앞서 올바른 분석과 결과해석을 위해 R로 불러온 데이터의 변수 유형을 파악하고, 필요 시 변수 유형을 변환해 주는 작업이 필요하다. 아래의 <그림 22>는 제공된 KoGES 추적조사 교육용 데이터를 R로 불러와 결합한 데이터 셋을 가지고 str()함수를 이용해 데이터 구조 등을 확인한 결과이다. 각각의 변수는 R이 인식하는 변수 유형으로 불러와져 있다.



```

> str(follow_data)
'data.frame':   1000 obs. of  28 variables:
 $ t00_id      : chr  "K_FOLLOW_0001" "K_FOLLOW_0002" "K_FOLLOW_0003" "K_FOLLOW_0004" ...
 $ t00_sex     : int   1 1 1 2 1 1 2 2 1 2 ...
 $ t01_age     : int   56 40 52 60 49 50 48 48 42 62 ...
 $ t01_edate   : int  200412 200401 200309 200504 200402 200401 200410 200308 200301 200312 ...
 $ t01_htn     : int   1 1 2 1 1 1 1 1 2 1 ...
 $ t01_dm      : int   1 1 1 1 1 1 1 2 1 1 ...
 $ t01_drink   : int   3 3 2 1 3 3 3 1 3 1 ...
 $ t01_smoke   : int   3 2 2 1 3 2 1 1 2 1 ...
 $ t01_exer    : int   1 1 1 1 1 2 2 2 2 2 ...
 $ t01_height  : int  159 169 165 165 166 174 158 157 168 148 ...
 $ t01_weight  : int   50 94 63 70 69 77 61 62 74 63 ...
 $ t01_glu0    : int   82 130 83 89 95 93 95 270 96 81 ...
 $ t01_age     : int   58 41 54 62 51 52 50 49 44 64 ...
 $ t02_edate   : int  200609 200510 200505 200704 200601 200606 200608 200412 200409 200512 ...
 $ t02_dm      : int   1 2 1 1 1 1 1 1 1 1 ...
 $ t02_glu0    : int   90 99999 90 106 124 81 94 141 104 90 ...
 $ t03_age     : int   61 44 57 65 54 56 53 54 47 66 ...
  
```

그림 22 | R로 불러와 결합된 KoGES 교육용 데이터 변수 유형

예제

조사일 변수의 변수 유형을 날짜형으로 변환하기
: 기반 조사일(t01_edate) ~ 4차 추적 조사일(t05_edate)

KoGES 추적조사 교육용 데이터에는 각 조사 차수별 조사일이 YYYYMM로 입력되어 있는데, 위의 <그림 22>에서 살펴본 바와 같이 R에서는 날짜가 아닌 6자리의 숫자로 인식하고 있다. 또한 조사일 변수에는 년과 월까지의 정보만 있어, YYYYMMDD 형태를 갖추기 위해서는 ‘일’에 대한 값을 연구자가 임의적으로 1일, 15일, 30일 등으로 지정해주어야 한다. 본 가이드북에서는 ‘일’에 대한 값을 임의적으로 1일로 설정하기로 하고, 년과 월의 정보를 문자형으로 인식시킨 후 paste() 함수를 이용해 일("01") 정보를 결합하여 YYYYMMDD 형태의 자료를 우선 생성한다. 이렇게 생성된 조사일자 변수를 as.Date() 함수를 이용해 R에서 활용 가능한 날짜 형태(YYYY-MM-DD)로 변환한 후, as.integer() 함수를 사용해 해당 조사일이 R의 기준 날짜(1970년 1월 1일)로부터 몇 일이 경과하였는지 ‘일’을 계산해준다(1970년 1월 1일 = 0, 1970년 1월 2일 = 1, ...).

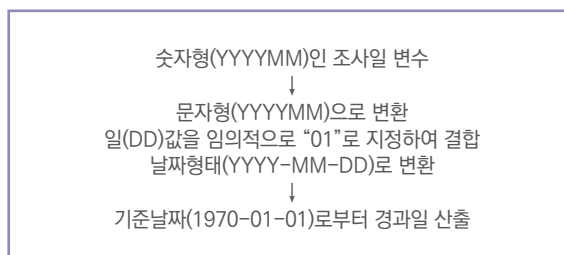


그림 23 | 조사일 변수를 숫자형에서 날짜형으로 변환하는 방법

```

# 변수 유형 변환 ----
follow_data_type <- follow_data_null
str(follow_data_type)

# 조사일자 변수 날짜형으로 바꾸기 ----
follow_data_type$edate1 <- as.Date(paste(as.character(follow_data_type$t01_edate), "01", sep=""), "%Y%m%d")
follow_data_type$edate2 <- as.Date(paste(as.character(follow_data_type$t02_edate), "01", sep=""), "%Y%m%d")
follow_data_type$edate3 <- as.Date(paste(as.character(follow_data_type$t03_edate), "01", sep=""), "%Y%m%d")
follow_data_type$edate4 <- as.Date(paste(as.character(follow_data_type$t04_edate), "01", sep=""), "%Y%m%d")
follow_data_type$edate5 <- as.Date(paste(as.character(follow_data_type$t05_edate), "01", sep=""), "%Y%m%d")

# 날짜 계산(1970-01-01 기준 계산) ----
follow_data_type$edate1_n <- as.integer(follow_data_type$edate1)
follow_data_type$edate2_n <- as.integer(follow_data_type$edate2)
follow_data_type$edate3_n <- as.integer(follow_data_type$edate3)
follow_data_type$edate4_n <- as.integer(follow_data_type$edate4)
follow_data_type$edate5_n <- as.integer(follow_data_type$edate5)

# 데이터 일부 확인 ----
follow_data_type[1:10, ] %>% dplyr::select(t01_edate, edate1, edate1_n)

```

결과

	t01_edate	edate1	edate1_n
1	200412	2004-12-01	12753
2	200401	2004-01-01	12418
3	200309	2003-09-01	12296
4	200504	2005-04-01	12874
5	200402	2004-02-01	12449
6	200401	2004-01-01	12418
7	200410	2004-10-01	12692
8	200308	2003-08-01	12265
9	200301	2003-01-01	12053
10	200312	2003-12-01	12387

5. 자료 분석하기

5-1. 분석 대상자 선정

자료 분석 전 분석에 포함할 연구 대상자를 선정하여야 한다. 일반적으로 의학 및 보건학 연구 논문에는 연구 대상자 흐름도(전체 대상자 중 특정 기준에 따라 본 분석에 포함된 대상자의 수를 나타낸 그림)를 확인 할 수 있다.

연구 주제 ‘비만한 사람에서 당뇨병 발생 위험이 높아질까?’와 관련한 분석을 위해서는 당뇨병과 비만을 정의하는 변수가 필수적이다. 또한 추적기간 동안의 당뇨병 발생을 확인하고자 하므로 기술통계 당시 당뇨병 유병자는 제외한 분석 데이터 셋을 생성하여 보자.

예제

당뇨병 또는 비만 관련 변수가 결측인 대상자와 기술통계 당시 당뇨병 유병자 제외하기

I. 변수 설명

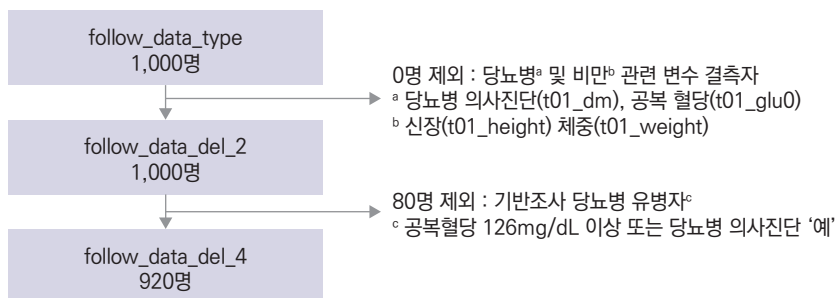
변수명	변수 설명	변수값 설명	변수 유형
t01_dm	기반, 당뇨병 의사진단	1=아니오, 2=예	범주형
t01_glu0	기반, 공복 혈당	() mg/dL	연속형
t01_height	기반, 신장	() cm	연속형
t01_weight	기반, 체중	() kg	연속형

II. 분석 대상자 선정

```
# 분석 대상자 선정 ----
# 당뇨병 및 비만 관련 변수 결측자 제외 #
follow_data_del_1 <- follow_data_type %>% dplyr::filter(! (is.na(t01_dm) & is.na(t01_glu0)))
follow_data_del_2 <- follow_data_del_1 %>% dplyr::filter(! (is.na(t01_height) | is.na(t01_weight)))

# 기술통계 당뇨병 유병자 제외 ----
follow_data_del_3 <- follow_data_del_2 %>%
  dplyr::mutate(t01_dm_d = ifelse((t01_glu0 >= 126 | t01_dm == 2), 1, 0))
follow_data_del_3$t01_dm_d[is.na(follow_data_del_3$t01_dm_d)] <- 0 # NA->0 처리
follow_data_del_4 <- follow_data_del_3 %>% dplyr::filter(t01_dm_d != 1)
```

결과



5-2. 변수 생성

원 자료에 포함된 여러 변수들을 조합하여 하나의 변수를 만들거나, 연속형 변수를 특정 값을 기준으로 나누어 범주형 변수로 생성하는 등 새로운 변수 생성이 필요한 경우가 있다. 아래 기준에 따라 기반 조사자료의 신장(t01_height)과 체중(t01_weight) 변수를 이용하여 체질량지수 및 비만도 변수를 생성하고, 1차 ~ 4차 추적 조사자료의 당뇨병 의사진단(t02_dm ~ t05_dm)과 공복 혈당(t02_glu0 ~ t05_glu0) 변수를 이용하여 각 추적 차수별 당뇨병 여부 변수와 추적 기간 동안의 당뇨병 발생 여부 변수(1차 ~ 4차 추적 조사 기간 동안 한 번 이상 당뇨병으로 분류된 경우로 정의)를 생성해 보자.

표 10 | 체질량지수 및 비만도 변수와 당뇨병 여부 변수 생성 기준

체질량지수 및 비만도 변수 생성 기준 ^a	당뇨병 여부 변수 생성 기준 ^b
· 체질량지수(BMI) : 체중(kg)/신장(m ²)	
· 저 체 중 : 체질량지수(BMI) < 18.5 kg/m ²	· 당 뇨 병 : 공복혈당 ≥ 126 mg/dL 이상 또는 과거 의사로부터 당뇨병을 진단 받은 적이 있는 경우 (당뇨병 의사진단 '예')
· 정 상 체 중 : 체질량지수(BMI) 18.5 ~ 23 kg/m ²	· 정 상 : 당뇨병에 해당하지 않은 모든 경우
· 과 체 중 : 체질량지수(BMI) 23 ~ 25 kg/m ²	
· 비 만 : 체질량지수(BMI) ≥ 25 kg/m ²	

^a 세계보건기구(WHO) 아시아-태평양 지역 기준

^b Report or a WHO/IDF consultation (2006) 기준

예제

체질량지수 및 비만도, 당뇨병 여부 변수 생성하기

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
t01_height	기반, 신장	() cm	연속형
t01_weight	기반, 체중	() kg	연속형
t02_glu0 ~ t05_glu0	1차 추적, 공복 혈당 ~ 4차 추적, 공복 혈당	() mg/dL	연속형
t02_dm ~ t05_dm	1차 추적, 당뇨병 의사진단 ~ 4차 추적, 당뇨병 의사진단	1=아니오, 2=예	범주형

II. 새로운 변수 생성

```
# 체질량지수 및 비만도 변수 생성 ----
follow_data_bmi <- follow_data_del_4 %>%
  dplyr::mutate(bmi = t01_weight / ((t01_height / 100) ^ 2))
follow_data_bmi_gr <- follow_data_bmi %>%
  dplyr::mutate(bmi_gr = ifelse(bmi < 18.5, 1,
                                ifelse(bmi >= 18.5 & bmi < 23, 2,
                                      ifelse(bmi >= 23 & bmi < 25, 3,
                                            ifelse(bmi >= 25, 4, NA))))))

# 각 추적자수별 당뇨병 유병 여부 변수 생성 ----
follow_data_bmi_dm <- follow_data_bmi_gr %>%
  dplyr::mutate(t02_dm_d = ifelse((t02_glu0 >= 126 | t02_dm == 2), 1, 0),
                t03_dm_d = ifelse((t03_glu0 >= 126 | t03_dm == 2), 1, 0),
                t04_dm_d = ifelse((t04_glu0 >= 126 | t04_dm == 2), 1, 0),
                t05_dm_d = ifelse((t05_glu0 >= 126 | t05_dm == 2), 1, 0))

follow_data_bmi_dm$t02_dm_d[is.na(follow_data_bmi_dm$t02_dm_d)] <- 0 # NA->0 처리
follow_data_bmi_dm$t03_dm_d[is.na(follow_data_bmi_dm$t03_dm_d)] <- 0
follow_data_bmi_dm$t04_dm_d[is.na(follow_data_bmi_dm$t04_dm_d)] <- 0
follow_data_bmi_dm$t05_dm_d[is.na(follow_data_bmi_dm$t05_dm_d)] <- 0

# 총 추적 기간 동안 당뇨병 발생 여부 변수 생성 ----
follow_data_bmi_dm_tot <- follow_data_bmi_dm %>%
  dplyr::mutate(dm_d = ifelse((t02_dm_d == 1 | t03_dm_d == 1 | t04_dm_d == 1 | t05_dm_d == 1), 1, 0))

# 저체중 대상자 향후 분석에서 제외(12명, 1.3%) ----
follow_data_bmi_dm_tot_del <- follow_data_bmi_dm_tot %>%
  dplyr::filter(bmi_gr != 1)
```

예제

관찰기간 변수 생성하기

추적자료를 이용하여 생존분석을 수행할 때에는 사건(예, 질병 발생 또는 사망 등)의 발생 유무와 함께 사건 발생까지의 시간 변수가 통계 분석에 포함되므로 해당 변수 생성이 필요하다. 추적조사 교육용 데이터를 이용하여 당뇨병 발생에 대한 관찰기간 변수를 생성해 보자. 관찰기간은 연구 시작 시점과 연구 종료 시점의 차이로 산출할 수 있다. 연구 시작 시점은 처음으로 연구에 참여한 기반 조사일이 되겠고, 연구 종료 시점은 당뇨병 발생 유무에 따라 달라진다. 당뇨병 발생군의 연구 종료 시점은 당뇨병이 발생한(확인된) 추적 조사일이 되겠고, 당뇨병 미 발생군의 경우 마지막 추적 조사일이 되겠다. 참고로 본 가이드북에서는 다루고 있지 않지만, 실제 연구에서는 관심 질병(event) 발생 없이 추적관찰 중 사망하는 대상자가 관찰될 수 있다. 이러한 경우, 해당 대상자는 중도절단(censoring, 사건발생='아니오') 처리해 주어야 하며, 관찰 기간은 기반 조사일 부터 사망일까지로 정의할 수 있다.

| 표 11 | 추적자료의 당뇨병 발생에 대한 관찰기간 산정 방법>

ID	기반	1차 추적	2차 추적	3차 추적	4차 추적	관찰기간
NO_1		O	O	O	O	기반 ~ 1차 추적 조사일
NO_2		X	O	O	O	기반 ~ 2차 추적 조사일
NO_3		X	X	O	O	기반 ~ 3차 추적 조사일
NO_4		X	X	X	X	기반 ~ 4차 추적 조사일
NO_5		X	X	-	-	기반 ~ 2차 추적 조사일

(O : 당뇨병 '예', X : 당뇨병 '아니오', - : 추적조사 미참여)

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
edate1_n ~ edate5_n	기반 조사일 - 기준일 ~ 4차 추적 조사일 - 기준일	1970년1월1일을 기준으로 해당 조사일의 경과일을 계산한 값	연속형
t02_dm ~ t05_dm	1차 추적, 당뇨병 여부 ~ 4차 추적, 당뇨병 여부	0=아니오, 1=예	범주형
dm_d	당뇨병 발생 여부	0=정상, 1=당뇨병	범주형

II. 새로운 변수 생성

```
# 당뇨 미발생군 연구 종료 시점(마지막 조사 참여시점) ----
follow_data_time <- follow_data_bmi_dm_tot_del
follow_data_time$max_date <- apply(follow_data_time[, c("edate2_n", "edate3_n", "edate4_n",
"edate5_n")], 1, max, na.rm = TRUE)

# 당뇨 발생군 연구 종료 시점(첫 발생 시점) ----
follow_data_time_end <- follow_data_time %>%
  dplyr::mutate(end_date = ifelse(t02_dm_d == 1, edate2_n,
                                ifelse(t02_dm_d == 0 & t03_dm_d == 1, edate3_n,
                                        ifelse(t02_dm_d == 0 & t03_dm_d == 0 & t04_dm_d == 1, edate4_n,
                                              ifelse(t02_dm_d == 0 & t03_dm_d == 0 & t04_dm_d == 0 &
t05_dm_d == 1, edate5_n, max_date))))))

# 연구 시작 시점, 연구 종료 시점 정의 ----
follow_data_time_end$start <- follow_data_time_end$edate1_n
follow_data_time_end$end <- follow_data_time_end$end_date

# 관찰기간(일) = (연구 종료 시점 - 연구 시작 시점) ----
follow_data_time_end$f_time_d <- follow_data_time_end$end - follow_data_time_end$start

# 관찰기간(년) ----
follow_data_time_end$f_time_y <- round(follow_data_time_end$f_time_d/365, 1)

# 생성 변수 확인 ----
follow_data_time_end %>% dplyr::select(t00_id, dm_d, start, end, f_time_d, f_time_y)
```

결과

	t00_id	dm_d	start	end	f_time_d	f_time_y
1	K_FOLLOW_0001	0	12753	16709	3956	10.8
2	K_FOLLOW_0003	0	12296	16222	3926	10.8
3	K_FOLLOW_0004	0	12874	15796	2922	8.0
4	K_FOLLOW_0005	1	12449	14245	1796	4.9
5	K_FOLLOW_0006	0	12418	15614	3196	8.8
6	K_FOLLOW_0007	0	12692	16679	3987	10.9
7	K_FOLLOW_0009	0	12053	15187	3134	8.6
8	K_FOLLOW_0010	0	12387	15584	3197	8.8
9	K_FOLLOW_0011	0	12084	15492	3408	9.3
10	K_FOLLOW_0012	0	12387	15614	3227	8.8
11	K_FOLLOW_0013	0	12815	16222	3407	9.3
12	K_FOLLOW_0014	0	13118	17136	4018	11.0
13	K_FOLLOW_0015	0	12631	16587	3956	10.8
14	K_FOLLOW_0016	0	12265	16467	4202	11.5
15	K_FOLLOW_0017	1	12723	16679	3956	10.8
16	K_FOLLOW_0018	0	12692	16648	3956	10.8
17	K_FOLLOW_0019	0	12418	15248	2830	7.8
18	K_FOLLOW_0020	0	12631	15887	3256	8.9
19	K_FOLLOW_0021	0	12784	16709	3925	10.8
20	K_FOLLOW_0022	0	13027	16953	3926	10.8
21	K_FOLLOW_0023	0	12631	16252	3621	9.9
22	K_FOLLOW_0024	1	12631	15522	2891	7.9
23	K_FOLLOW_0025	0	12965	16222	3257	8.9

R TIP

apply 계열 함수

본 가이드북에서는 apply 계열 함수 중 자주 사용되는 apply() 함수와 sapply() 함수, 그리고 tapply() 함수에 대해서 간략하게 소개하고자 한다. 또한 함수에 따라 적용 가능한 데이터 형태와 출력 형태가 다르다는 점을 인지하고 사용하길 바란다.

- apply(x, direction, function)

여기서 x는 2차원 데이터(배열, 행렬, 데이터프레임)를 의미하고, direction는 방향성을 말한다. 따라서 apply() 함수를 사용하면, 2차원 데이터에 방향성 기준으로 동일한 함수를 적용할 수 있게 된다.

예제 1

```
> a <- matrix(1:12, nrow = 3, ncol = 4)
> a
      [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
>
> apply(a, 1, max) # 행 기준
[1] 10 11 12
>
> apply(a, 2, max) # 열 기준
[1] 3 6 9 12
```

-- 1~12까지의 숫자로 이루어진 3행 4열의 행렬을 생성하고 이를 a라고 명명한다.

-- apply() 함수를 이용해 a의 행(열)기준 최댓값을 산출한 결과는 왼쪽과 같다.

예제 2

데이터셋(follow_data_time_end)의 추적 조사일자별 가장 큰 값을 산출하면 다음과 같다.

1 (행 기준)

```
> apply(follow_data_time_end[, c("edate2_n", "edate3_n", "edate4_n", "edate5_n")], 1, max, na.rm = TRUE)
[1] 16709 16222 15796 16405 15614 16679 15187 15584 15492 15614 16222 17136 16587 16467 16679 16648
[17] 15248 15887 16709 16953 16252 16556 16222 16587 15126 16953 16709 16648 16556 15887 16283 17014
[33] 16709 15553 15857 16709 16709 16252 15614 15979 17014 16375 16252 16617 15431 17198 15461 16679
      :
[881] 16922 16405 16648 16709 16892 16405 16832 16770 16587 16556 16648 16617 16770 16191 15279 16617
[897] 16556 15887 15522 16010 16556 15034 16405 16922 16010 16344 15979 16344
```

2 (열 기준)

```
> apply(follow_data_time_end[, c("edate2_n", "edate3_n", "edate4_n", "edate5_n")], 2, max, na.rm = TRUE)
edate2_n edate3_n edate4_n edate5_n
14123    15156    16375    17198
```

- sapply(x, function)

apply() 함수에서 방향성이 열 기준인 경우와 동일한 기능을 한다.

```
> sapply(follow_data_time_end[, c("edate2_n", "edate3_n", "edate4_n", "edate5_n")], max, na.rm = TRUE)
edate2_n edate3_n edate4_n edate5_n
14123    15156    16375    17198
```

- `tapply(x, group, function)`

`tapply()` 함수의 경우, 그룹별 정보를 요약할 때 주로 사용한다.

성별(`t00_sex`)에 따른 공복혈당(기반, `t01_glu0`)에 대한 기술통계량 구하고자 한다면, 다음과 같다.

```
> tapply(follow_data_time_end$t01_glu0, follow_data_time_end$t00_sex, summary, na.rm = TRUE)
$`1`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
59.00  80.00   86.00   87.75  95.00  125.00     1

$`2`
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
60.00  78.00   84.00   85.77  92.00  124.00
```

5-3. 총 관찰인년(Person-years) 산출

총 관찰인년이란 각 대상자의 관찰기간의 합이다. 대상자마다 관찰기간이 상이하므로 전체 대상자의 수와 함께 관찰인년을 고려해야한다. 예를 들어 10명의 대상자를 각 1년씩 관찰했을 경우 총 관찰인년은 10 인년이 되며, 10명 중 5명은 1년씩, 5명은 2년씩 관찰되었다면 총 관찰인년은 15 인년이 된다.

예제

비만도 그룹별 당뇨병 발생자 수 및 관찰인년 산출하기

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
dm_d	당뇨병 발생 여부	0=정상, 1=당뇨병	범주형
f_time_y	관찰기간	()년	연속형

II. R을 이용한 통계분석

```
# 비만도 그룹별 당뇨병 발생자수 구하기 ----
gmodels::CrossTable(follow_data_time_end$bmi_gr, follow_data_time_end$dm_d)

# 비만도 그룹별 관찰인년 산출하기 ----
options(digits = 5)
doBy::summaryBy(f_time_y ~ bmi_gr, data = follow_data_time_end, FUN = c(mean, sum, min, max))
```

결과

cell contents

Chi-square contribution	N
N / Row Total	
N / Col Total	
N / Table Total	

Total Observations in Table: 908

follow_data_time_end\$bmi_gr	follow_data_time_end\$dm_d		row total
	0	1	
2	265	11	276
	0.807	0.023	
	0.960	0.040	0.304
	0.321	0.133	
	0.292	0.012	
3	220	20	240
	0.017	0.271	
	0.917	0.083	0.264
	0.267	0.241	
	0.242	0.022	
4	340	52	392
	0.734	0.295	
	0.867	0.133	0.432
	0.412	0.627	
	0.374	0.057	
column total	825	83	908
	0.909	0.091	

```
> doBy::summaryBy(f_time_y ~ bmi_gr, data = follow_data_time_end,
  bmi_gr f_time_y.mean f_time_y.sum f_time_y.min f_time_y.max
1 2 9.8554 2720.1 2.0 11.8
2 3 9.8779 2370.7 1.6 11.6
3 4 9.4829 3717.3 1.4 11.7
```

결과 해석

· 비만도가 정상체중(bmi_gr=2)인 그룹은 총 276명의 대상자 중 당뇨병 발생자가 11명이고, 총 관찰인년은 2720.1이다. 과체중(bmi_gr=3)인 그룹은 총 240명의 대상자 중 당뇨병 발생자는 20명이고, 총 관찰인년은 2370.7이다. 마지막으로 비만(bmi_gr=4)인 그룹은 총 392명의 대상자 중 당뇨병 발생자는 52명이고, 총 관찰인년은 3717.3이다.

5-4. 생존함수 추정(생존곡선)

생존함수를 추정하는 방법에는 크게 생명표법(life table method)과 Kaplan-Meier법이 있다. 생명표법은 관찰 기간을 일정한 구간으로 구분한 후, 각 구간에서 관찰된 사망자로부터 구간 사망확률과 구간 생존확률을 구하고 이로부터 일정 기간까지의 구간 생존확률의 누적치인 누적 생존율을 산출한다. Kaplan-Meier법은 관찰 기간을 사망 또는 절단이 한 건 이상 나타나는 지점으로 구간을 구분한 뒤, 각 사망(또는 절단)이 발생한 시점에서 생존율을 산출해 나감으로써 누적 생존율을 산출하는 방법이다.

생존곡선은 사망/생존에 적용하기 위하여 제시된 방법이지만, 질병 발생 유무에도 적용이 가능하다.

우선적으로 생존분석의 특성상 time과 event를 먼저 정의하고 R에서 제시하는 survival 패키지의 Surv() 함수등을 이용해 이후 분석을 수행한다.

예제

비만도에 따른 당뇨병 발생 생존곡선 그리기 - Kaplan-Meier법 이용

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
dm_d	당뇨병 발생 여부	0=정상, 1=당뇨병	범주형
f_time_y	관찰기간	()년	연속형

II. R을 이용한 통계분석

```
# time과 event 정의 ----
install.packages("survival")
library(survival)
y <- survival::Surv(follow_data_time_end$f_time_y, follow_data_time_end$dm_d)

# 생존곡선(Kaplan Meier curve fitting) ----
follow_data_time_end$bmi_gr <- as.factor(follow_data_time_end$bmi_gr)
fit <- survival::survfit(y ~ bmi_gr, data = follow_data_time_end)
summary(fit)
plot(fit, main= "KM curve", ylab = "Survival", xlab = "year", col = 1:3, lty = 1:3)
legend("bottomleft", legend = c("정상체중", "과체중", "비만"), col = 1:3, lty = 1:3)

# log-rank test ----
log_rank <- survival::survdiff(y ~ bmi_gr, data = follow_data_time_end)
log_rank
```

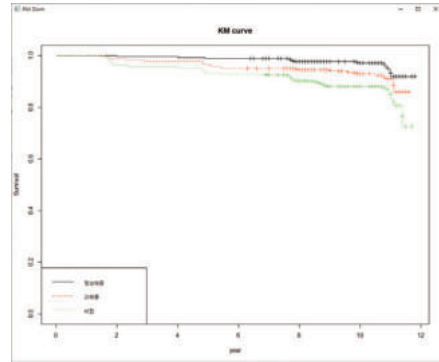
결과

```
> summary(fit)
Call: survfit(formula = y ~ bmi_gr, data = follow_data_time_end)

bmi_gr=2
time n.risk n.event survival std.err lower 95% CI upper 95% CI
2.0 276 1 0.996 0.00362 0.989 1.000
4.0 275 1 0.993 0.00511 0.983 1.000
4.9 274 1 0.989 0.00624 0.977 1.000
7.7 264 2 0.982 0.00814 0.966 0.998
7.8 259 1 0.978 0.00895 0.960 0.996
9.9 173 1 0.972 0.01053 0.952 0.993
10.8 153 1 0.959 0.01370 0.933 0.987
10.9 71 1 0.946 0.01904 0.909 0.984
11.0 35 1 0.919 0.03243 0.858 0.985

bmi_gr=3
time n.risk n.event survival std.err lower 95% CI upper 95% CI
1.6 240 1 0.996 0.00416 0.988 1.000
1.7 239 1 0.992 0.00587 0.980 1.000
1.8 238 1 0.988 0.00717 0.974 1.000
2.3 237 1 0.983 0.00826 0.967 1.000
2.9 236 1 0.979 0.00922 0.961 0.997
4.8 235 2 0.971 0.01086 0.950 0.992
4.9 233 1 0.967 0.01159 0.944 0.990
5.0 232 1 0.963 0.01226 0.939 0.987
5.1 231 1 0.958 0.01290 0.933 0.984
5.4 230 1 0.954 0.01350 0.928 0.981
5.5 229 1 0.950 0.01407 0.923 0.978
7.9 215 1 0.946 0.01468 0.917 0.975
9.0 180 1 0.940 0.01551 0.910 0.971
9.6 171 1 0.935 0.01637 0.903 0.967
9.9 168 1 0.929 0.01719 0.896 0.964
10.5 157 1 0.923 0.01807 0.889 0.959
10.8 146 2 0.911 0.01991 0.872 0.951
11.1 18 1 0.860 0.05264 0.763 0.970

bmi_gr=4
time n.risk n.event survival std.err lower 95% CI upper 95% CI
1.4 392 1 0.997 0.00255 0.992 1.000
1.6 391 1 0.995 0.00360 0.988 1.000
1.7 390 4 0.985 0.00620 0.973 0.997
1.8 386 7 0.967 0.00904 0.949 0.985
1.9 379 1 0.964 0.00937 0.946 0.983
2.3 378 1 0.962 0.00969 0.943 0.981
2.4 377 2 0.957 0.01029 0.937 0.977
4.0 375 2 0.952 0.01085 0.931 0.973
4.8 373 3 0.944 0.01162 0.921 0.967
4.9 370 5 0.931 0.01279 0.906 0.957
5.4 365 1 0.929 0.01301 0.903 0.954
6.9 364 1 0.926 0.01322 0.900 0.952
7.6 359 1 0.923 0.01343 0.897 0.950
7.7 353 3 0.910 0.01446 0.882 0.939
7.8 346 3 0.902 0.01504 0.873 0.932
8.3 310 1 0.900 0.01527 0.870 0.930
```



```
> log_rank
Call:
survdiff(formula = y ~ bmi_gr, data = follow_data_time_end)

N observed Expected (O-E)^2/E (O-E)^2/V
bmi_gr=2 276 11 25.6 8.350 12.145
bmi_gr=3 240 20 22.5 0.283 0.391
bmi_gr=4 392 52 34.8 8.445 14.631

Chisq= 17.2 on 2 degrees of freedom. p= 2e-04
```

결과 해석

- 로그-순위 검정(Log-Rank) 결과 유의수준 0.05하에서 유의확률이 0.0002로 '비만 그룹별(정상체중/과체중/비만) 당뇨병 발생에 대한 생존곡선은 같다'라는 귀무가설을 기각할 수 있다.

5-5. Cox 비례위험모형

Cox 비례위험모형은 생존분석 기법 중 가장 대표적인 방법으로, 생존기간과 여러 요인들 간의 관계를 알아보고자 할 때 이용된다. 주로 사망 또는 특정 질병 발생의 위험인자가 무엇인지 추정하고자 할 때 사용하며, 기준 그룹 대비 특정 그룹의 위험도(Hazard ratio, HR)를 구할 수 있다. 먼저 Cox 모형은 다음 식을 활용하여 표현할 수 있다.

$$h(t, x) = h_0(t) \exp(\beta_1 x_1 + \dots + \beta_k x_k)$$

여기서 $h_0(t)$ 는 기저 위험함수(Hazard function)라고 하며, 위험함수에 미치는 여러 설명변수들의 영향이 전혀 없을 경우를 가정한다. 또한 $h(t, x)$ 는 t시점에서 k개의 독립변수가 x_1, x_2, \dots, x_k 일 때의 위험함수를 말한다.

예를 들어 설명변수 x_1 에 대하여 $x_1 = 1$ 인 경우를 특정(비교) 그룹, $x_1 = 0$ 인 경우를 기준 그룹으로 가정했을 때,

특정 그룹의 hazard는 $h_0(t) \exp(\beta_1 + \beta_2 x_2 + \dots + \beta_k x_k)$,

기준 그룹의 hazard는 $h_0(t) \exp(\beta_2 x_2 + \dots + \beta_k x_k)$ 이 되며,

$$\text{Hazard ratio} = \frac{h_0(t) \exp(\beta_1 + \beta_2 x_2 + \dots + \beta_k x_k)}{h_0(t) \exp(\beta_2 x_2 + \dots + \beta_k x_k)} = \exp(\beta_1) \text{가 된다.}$$

설명변수 x_1 이외 다른 변수들의 값들이 일정하다고 할 때, $HR > 1$ 이면 설명변수 x_1 의 기준 그룹 대비 특정 그룹의 질병(사망) 발생 위험률이 더 크고, $HR = 1$ 이면 동일하며 $HR < 1$ 이면 작다고 해석한다. 만일 설명변수가 범주형이 아닌 연속형일 경우는 설명변수가 1 단위 증가함에 따른 질병(사망) 발생 위험률 비로 해석하면 된다.

해당 통계방법을 사용하기 위해서는 비례 위험 가정을 만족하는지에 대한 검토가 필요하다. 비례 위험 가정이란 설명변수의 효과가 시간과 관계없이 독립적으로 일정하다는 것으로 아래 예제를 통해 검토한 사항을 [R TIP]에 수록하였으니, 이를 참고하시기 바란다.

예제

비만도에 따라 당뇨병 발생 위험에 차이가 있는지 분석하기

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
dm_d	당뇨병 발생 여부	0=정상, 1=당뇨병	범주형
f_time_y	관찰기간	()년	연속형

II. R을 이용한 통계분석

```
# cox(1) ----
# y <- survival::Surv(follow_data_time_end$f_time_y, follow_data_time_end$dm_d)
cox1 <- survival::coxph(y ~ bmi_gr, data = follow_data_time_end)
options(digits = 2)
summary(cox1)
```

결과

```
> summary(cox1)
Call:
survival::coxph(formula = y ~ bmi_gr, data = follow_data_time_end)

n= 908, number of events= 83

      coef exp(coef) se(coef)      z Pr(>|z|)
bmi_gr3 0.727    2.069   0.376  1.94  0.05280 .
bmi_gr4 1.249    3.486   0.332  3.76  0.00017 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
bmi_gr3      2.07    0.483   0.991    4.32
bmi_gr4      3.49    0.287   1.819    6.68

Concordance= 0.635 (se = 0.026 )
Likelihood ratio test= 18.4 on 2 df,  p=1e-04
Wald test               = 15.6 on 2 df,  p=4e-04
Score (logrank) test = 17.2 on 2 df,  p=2e-04
```

결과 해석

- 비만도가 정상체중(bmi_gr=2)인 그룹과 비교하여 과체중(bmi_gr=3)인 그룹의 당뇨병 발생에 대한 HR 및 95% CI는 2.07 (0.99~4.32), 비만(bmi_gr=4)인 그룹은 3.49 (1.82~6.68)으로 유의수준 0.05 수준하에서 판단할 경우 비만인 그룹만 정상체중인 그룹 대비 당뇨병 발생 위험이 통계적으로 유의하게 높았다.

예제

성별과 연령을 보정했을 때 비만도에 따라 당뇨병 발생 위험에 차이가 있는지 분석하기

I. 변수 설명

변수명	변수 설명	변수값 설명	변수 유형
bmi_gr	비만도	1=저체중, 2=정상체중, 3=과체중, 4=비만	범주형
dm_d	당뇨병 발생 여부	0=정상, 1=당뇨병	범주형
f_time_y	관찰기간	()년	연속형
t_sex	성별	1=남자, 2=여자	범주형
t_age	연령	만 ()세	연속형

II. R을 이용한 통계분석

```
# 여성을 기준으로 재설정 ----
follow_data_time_end$t00_sex <- as.factor(follow_data_time_end$t00_sex)
follow_data_time_end$t00_sex <- relevel(follow_data_time_end$t00_sex, ref = "2")

# cox(2) 성별과 연령을 추가 보정 ----
cox2 <- survival::coxph(y ~ bmi_gr + t00_sex + t01_age, data = follow_data_time_end)
summary(cox2)
```

결과

```
> summary(cox2)
Call:
survival::coxph(formula = y ~ bmi_gr + t00_sex + t01_age, data = 
  n = 908, number of events = 83

              coef exp(coef) se(coef)      z Pr(>|z|)
bmi_gr3  0.7997    2.2249   0.3764  2.12  0.03360 *
bmi_gr4  1.2726    3.5700   0.3323  3.83  0.00013 ***
t00_sex1  0.2955    1.3438   0.2226  1.33  0.18440
t01_age   0.0460    1.0471   0.0126  3.65  0.00026 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
bmi_gr3          2.22      0.449    1.064    4.65
bmi_gr4          3.57      0.280    1.861    6.85
t00_sex1          1.34      0.744    0.869    2.08
t01_age           1.05      0.955    1.022    1.07

Concordance= 0.694 (se = 0.028 )
Likelihood ratio test= 32.4 on 4 df,  p=2e-06
Wald test               = 28.7 on 4 df,  p=9e-06
Score (logrank) test = 30.4 on 4 df,  p=4e-06
```

결과 해석

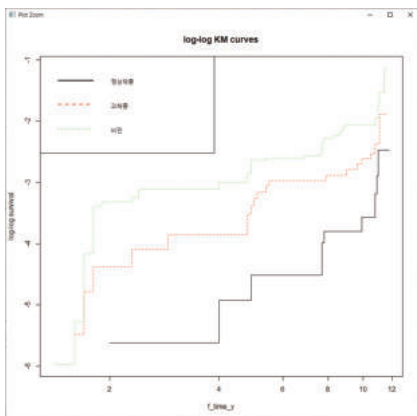
· 성별과 연령을 보정하였을 때, 비만도가 정상체중(bmi_gr=2)인 그룹과 비교하여 과체중(bmi_gr=3)인 그룹의 당뇨병 발생에 대한 HR 및 95% CI는 2.22 (1.06-4.65), 비만(bmi_gr=4)인 그룹은 3.57 (1.86-6.85)으로 과체중 및 비만 그룹은 정상 체중인 그룹에 비하여 모두 당뇨병 발생 위험이 통계적으로 유의하게 높았다.

비례성 가정 검토

Cox 모델을 사용하기 위해서는 비례 위험 가정을 만족하는지에 대한 검토가 필요하다. 비례 위험 가정이란 설명변수의 효과가 시간과 관계없이 독립적으로 일정하다는 것이며, Log-Log 생존그림을 그려 두 개 이상의 곡선이 평행(parallel)하거나 `cox.zph()` 함수 결과를 통해 통계적인 유의성을 판단할 수 있다. 만약 설명변수의 효과가 시간에 따라 변화하는 경우(비례 위험 가정을 만족하지 않을 경우), Stratified cox regression 또는 Time dependent variable에 의한 Extended cox regression을 이용하여 분석을 진행하여야 한다.

• Log-Log 생존그림 이용

```
# (1) log-log plot ----
# y <- survival::Surv(follow_data_time_end$f_time_y, follow_data_time_end$dm_d)
# fit<- survival::survfit(y ~ bmi_gr, data = follow_data_time_end)
plot(fit, fun = "cloglog", main = "log-log KM curves", ylab = "log-log survival",
     xlab = "f_time_y", col = 1:3, lty = 1:3)
```



Log-Log 생존그림을 확인한 결과 비만도(bmi_gr) 변수의 수준별 그래프가 서로 평행하게 변화하므로, 비례 위험 가정을 만족하는 것으로 판단된다.

- `cox.zph()` 함수 이용

```
# (2) cox.zph ----
cox1 <- survival::coxph(y ~ bmi_gr, data = follow_data_time_end)
survival::cox.zph(cox1) # 비교
# 더미변수 처리 후 적용
follow_data_time_end <- transform(follow_data_time_end,
                                   bmi_dum1 = as.factor(ifelse(bmi_gr == "3", 1, 0)),
                                   bmi_dum2 = as.factor(ifelse(bmi_gr == "4", 1, 0)))
cox1.dummy <- survival::coxph(y ~ bmi_dum1 + bmi_dum2, data = follow_data_time_end)
survival::cox.zph(cox1.dummy)
```

```
> survival::cox.zph(cox1)
      chisq df    p
bmi_gr   2.88  2 0.24
GLOBAL   2.88  2 0.24
>
> survival::cox.zph(cox1.dummy)
      chisq df    p
bmi_dum1 0.0677 1 0.79
bmi_dum2 1.8069 1 0.18
GLOBAL   2.8793 2 0.24
```

`cox.zph()` 함수를 적용하기 위해 비만도가 정상체중인 그룹을 기준으로 과체중과 비만일 경우를 더미변수 처리해 주었고, 그 결과 유의확률이 유의수준 0.05보다 모두 크므로 비례 위험 가정을 만족한다는 귀무가설을 기각할 수 없다.

KoGES 데이터 분석 가이드북
[R편]

Korean Genome and Epidemiology Study

부록

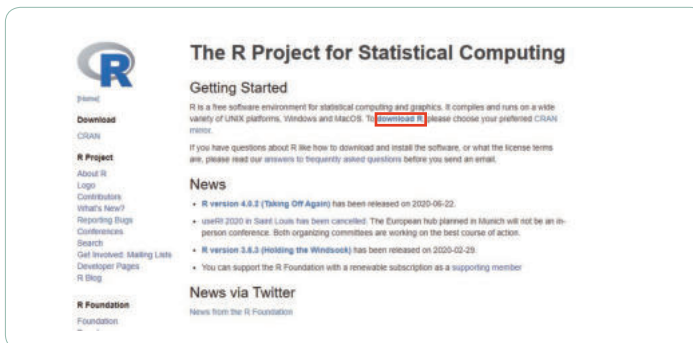
1. R과 RStudio 설치하기
2. RStudio 들어가기
3. 분석 목적에 따른 통계분석 방법 요약

부록

1. R과 RStudio 설치하기

1-1. R 설치하기 (Version 3.6.3)

R을 사용하려면 먼저 PC에 R을 설치해야 하며, RStudio를 추가로 설치한다. RStudio는 코딩, 파일관리 등 R 사용자의 편의를 위해 제공되는 통합개발환경(IDE) 소프트웨어이다.

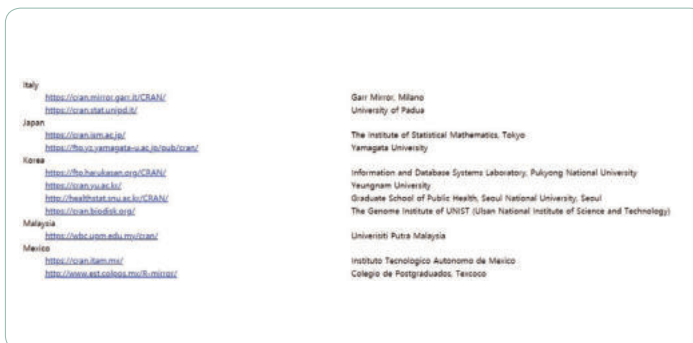


①

R 공식 웹 사이트

(<http://www.r-project.org/>)

에서 [download R]을 클릭



②

한국에 서버를 두고 있는

미러사이트 중 하나를 클릭

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages. **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for macOS](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2020-06-22, Taking Off Again) [R-4.0.2.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R.alpha](#) and [beta.releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions: About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

③
운영체제에 맞는 프로그램 클릭

R for Windows

Subdirectories:

- [base](#)
- [contrib](#)
- [old-contrib](#)
- [libdata](#)

Binaries for base distribution. This is what you want to **install R for the first time**.
Binaries of contributed CRAN packages (for R >= 2.13.0, managed by Uwe Ligges). There is also information on [this tool, software](#) available for CRAN Windows services and corresponding environment and make variables.
Binaries of contributed CRAN packages for outdated versions of R (for R < 2.13.0, managed by Uwe Ligges).
Tools to build R and R packages. This is what you want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

You may also want to read the [FAQ](#) and [R for Windows FAQ](#).

Note: CRAN does some checks on these binaries for viruses, but cannot give guarantees. Use the normal precautions with downloaded executables.

④
[base] 클릭

R-4.0.2 for Windows (32/64 bit)

[Download R 4.0.2 for Windows](#) (64 megabytes, 32/64 bit)
[Installation and other instructions](#)
[New features in this version](#)

If you want to double-check that the package you have downloaded matches the package distributed by CRAN, you can compare the [checksum](#) of the file to the [checksum](#) on the master server. You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- [Does R run under my version of Windows?](#)
- [How do I update packages to my previous version of R?](#)
- [Should I use 32-bit or 64-bit R?](#)

Please see the [FAQ](#) for general information about R and the [R-4.0.2 FAQ](#) for Windows-specific information.

Other builds

- Patches to this release are incorporated in the [patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [current snapshot build](#).
- [Previous releases](#)

Note to submitters: A stable link which will redirect to the current Windows binary release is: [CRAN_MIRROR/binary/windows/choose-release.html](#)

Last change: 2020-06-22

⑤
가장 최근의 버전(R 4.0.2)이 올라와 있으므로, 이전 버전을 다운을 위해
[Previous releases]를 클릭

Previous Releases of R for Windows

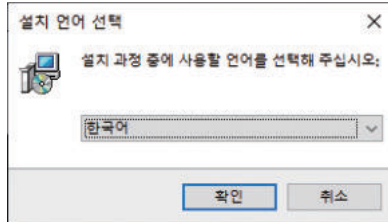
This directory contains previous binary releases of R for Windows.

The current release, and links to development snapshots, are available [here](#). Source code for these releases and others is available through [the main CRAN page](#).

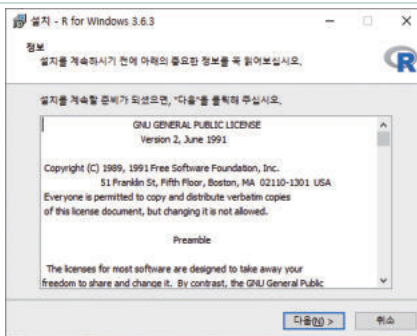
In this directory:

- R 4.0.2 (June, 2020)
- R 4.0.1 (June, 2020)
- R 4.0.0 (April, 2020)
- R 3.6.3 (February, 2020)**
- R 3.6.2 (December, 2019)
- R 3.6.1 (July, 2019)
- R 3.6.0 (April, 2019)
- R 3.5.2 (March, 2019)
- R 3.5.2 (December, 2018)
- R 3.5.1 (July, 2018)
- R 3.5.0 (April, 2018)
- R 3.4.4 (March, 2018)
- R 3.4.3 (November, 2017)
- R 3.4.2 (September, 2017)
- R 3.4.1 (June, 2017)
- R 3.4.0 (April, 2017)
- R 3.3.3 (March, 2017)
- R 3.3.2 (October, 2016)
- R 3.3.1 (June, 2016)
- R 3.3.0 (April, 2016)

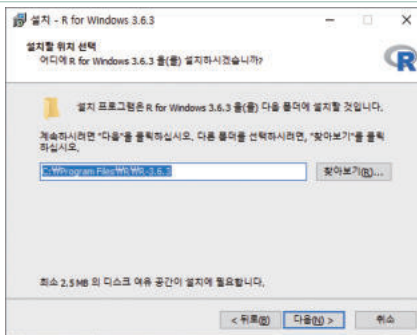
⑥
R 3.6.3 버전을 클릭한 후
R-3.6.3-win.exe를 실행



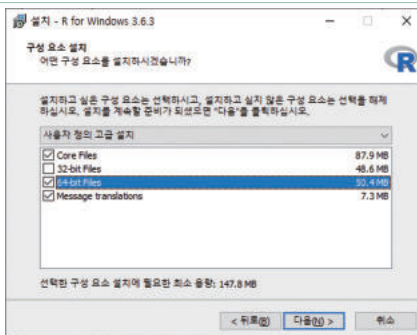
⑦
설치언어 [한국어] 선택



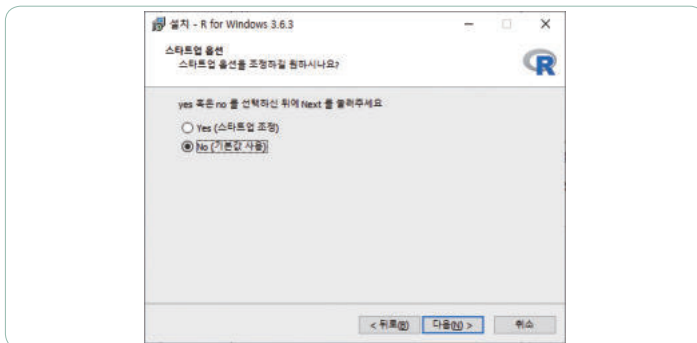
⑧
[다음] 선택



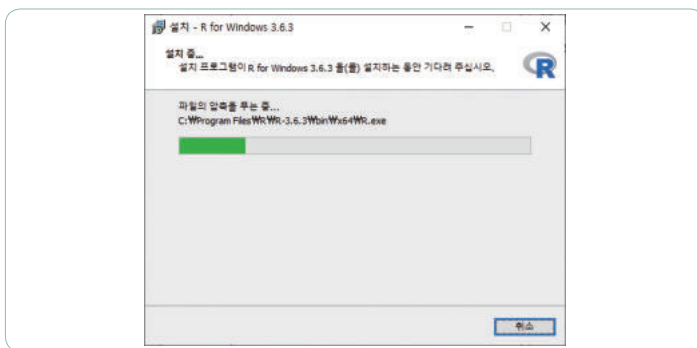
⑨
설치할 위치 지정



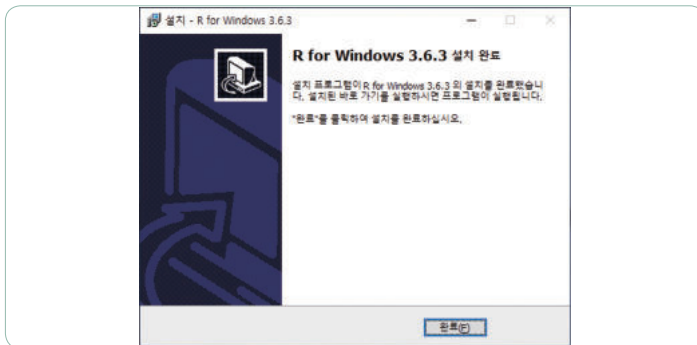
⑩
32bit 또는 64bit 중 선택



⑪
기본값 유지 후 [다음] 계속 클릭



⑫
설치 진행 중



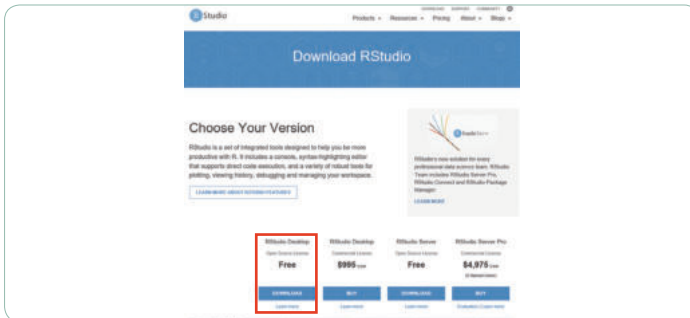
⑬
[완료] 버튼 클릭

1-2. RStudio 설치하기



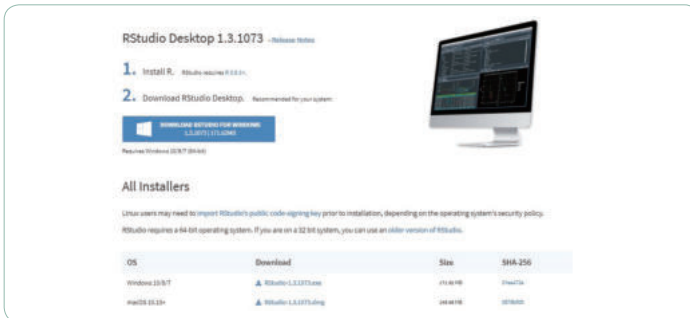
①

RStudio 웹 사이트
(<http://rstudio.com/>) 접속 후
상단의 [DOWNLOAD]을 클릭



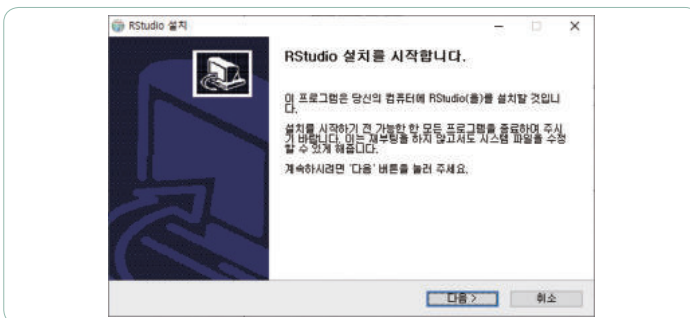
②

Open Source License로 제공
하는 RStudio Desktop의
[DOWNLOAD]을 클릭



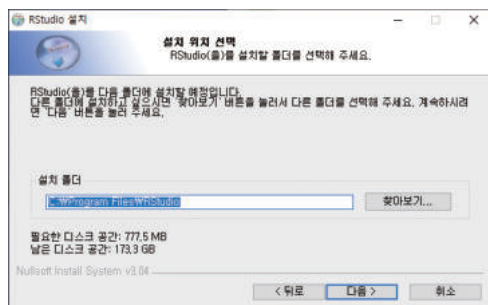
③

운영체제에 맞는 링크를 클릭해
설치 파일을 다운로드

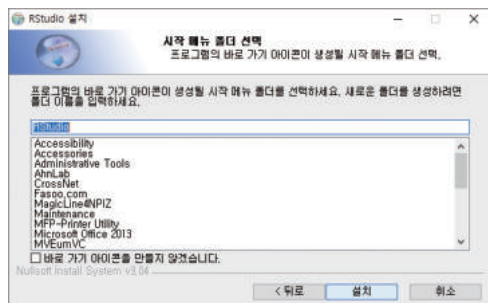


④

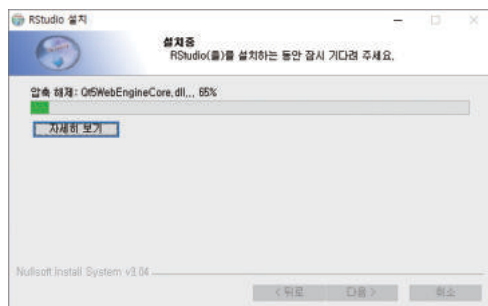
설치 시작 [다음] 버튼 클릭



- ⑤
설치 위치 선택 후 옵션 변경하지 않고 [다음] 버튼 클릭



- ⑥
시작 메뉴 폴더 선택에서 [설치] 버튼 클릭하면 설치 진행



- ⑦
설치중 화면



- ⑧
설치완료 되면 [마침] 버튼 클릭



설치 관련

• Window의 사용자 계정 확인

R과 RStudio를 정상적으로 설치했음에도 RStudio가 에러가 나거나 그래프가 잘 그려지지 않을 경우가 있다. 그럴 경우 가장 먼저 Window의 사용자 계정이 한글로 되어 있는지 확인해야 한다. 이는 RStudio가 계정이 한글일 경우 경로를 잘 인식하지 못하는 문제점을 가지고 있기 때문이다. 다음의 방법으로 사용자 계정을 영문 계정으로 변경한다.

- ① 시작
- ② 설정(Window 설정)
- ③ 계정
- ④ 기타 사용자
- ⑤ 이 PC에 다른 사용자 추가
- ⑥ 로컬 사용자 및 그룹 - 사용자(마우스 우클릭) - 새 사용자 - 새 사용자 만들기
- ⑦ 윈도우키 + X (종료 또는 로그아웃)
- ⑧ 새(영문) 계정으로 로그인

• RStudio 바탕화면에 아이콘 만들기

RStudio가 설치된 경로로 간 다음 RStudio > bin > rstudio.exe 마우스 우클릭 후 ⇒ 보내기 ⇒ 바탕 화면에 바로가기 만들기 ⇒ 이름 변경하기

• RStudio 권한 설정하기

RStudio 아이콘에서 마우스 우클릭 후 [속성] ⇒ [호환성] 클릭 ⇒ [관리자 권한으로 이 프로그램 실행] 체크 후 [확인] 버튼 클릭

2. RStudio 들어가기

2-1. RStudio 인터페이스

설치 후 RStudio를 처음 실행하면 콘솔(Console) 창, 환경(Environment) 창, 파일(Files) 창이 나타나며, 콘솔 창 오른쪽 위에 있는 아이콘을 클릭하면 콘솔 창이 작아지면서 스크립트(Script) 창이 나타난다. 이렇게 해서 RStudio는 크게 총 4개의 창으로 구분된다.

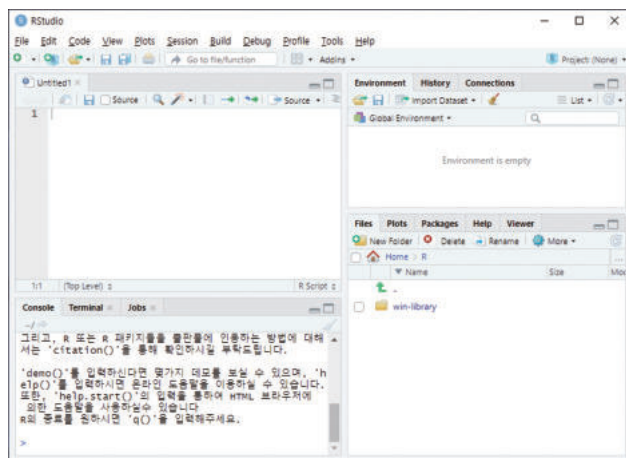


그림 24 | RStudio 인터페이스

① 스크립트(Script) 창

코드나 메모를 기록할 수 있는 일종의 문서 편집기 창이다.. 스크립트 창에 코드를 작성하고 **[Enter]**키를 누르면 커서가 다음 행으로 넘어가면서 행이 추가되며, 스크립트 창의 코드 실행 방법은 다음과 같다.

- ① 실행하려는 코드에 마우스 커서를 위치시킨 후 **[Ctrl]+[Enter]** 또는 Script창 상단 메뉴 'Run' 단추 클릭
- ② 마우스로 드래그 또는 블록 지정 후 **[Ctrl]+[Enter]**
- ③ 모든 코드를 실행하려면 **[Ctrl]+[Alt]+[R]**

위의 방법으로 코드를 실행하면, 그 코드가 콘솔 창으로 넘어가 실행되어 결과가 출력된다. 스크립트 창에는 보통 코드의 이해를 높이기 위해 일종의 메모를 달아줄 수 있는데 방법은 # 기호를 사용하면 되며, 이 메모는 주석문(comments)으로 취급되어 코드 실행시 결과에 영향을 주지 않는다.

② 콘솔(Console) 창

스크립트 창 또는 콘솔 창에서 작성한 프로그램 실행 결과를 볼 수 있으며, 그 밖에 패키지 설치, 연산 결과, 에러/오류 메시지 등의 로그를 볼 수 있는 창이다. 콘솔 창에서도 코드를 입력하여 **[Enter]**키를 눌러 1라인씩 실행되는 결과를 볼 수 있으며, 스크립트 창과는 다르게 사용 준비를 표시하는 프롬프트(">")가 나타난다.

③ 환경(Environment) 창

생성한 데이터 정보를 확인할 수 있는 환경 창과 스크립트 창 이력을 볼 수 있는 히스토리(History) 창 등이 있다.

④ 파일(Files) 창

윈도우 탐색기와 유사한 기능을 하는 파일(Files)창이 있어 폴더 생성, 파일 찾기, 작업경로(워킹 디렉토리) 설정 등을 할 수 있다. 그 밖에 그래프를 볼 수 있는 플롯(Plots) 창, 설치된 패키지 목록과 추가 설치 및 업데이트를 할 수 있는 패키지(Packages) 창, 도움말 검색 기능을 하는 도움말(Help) 창 등이 있다.

RStudio에서 자주 사용하는 단축키

• 스크립트 창

새로운 스크립트 창 열기 : Ctrl + Shift + n

코멘트 부호(#) 삽입 : Ctrl + Shift + c

할당연산자 (<-) : Alt + -

파이프 연산자(%)>%) 삽입 : Ctrl + Shift + m

스크립트 창 저장하기 : Ctrl + s

텍스트 찾고 바꾸기 : Ctrl + f

• 콘솔 창

콘솔 창 화면 모두 지우기 : Ctrl + l

• 기타 단축키 관련

Tools > Modify Keyboard Shortcuts 로 들어가면, 단축키 등록, 편집, 조회 가능

2-2. RStudio 유용한 환경설정

RStudio에는 RStudio 전반에 영향을 미치는 옵션인 글로벌 옵션(Global Options)과 프로젝트를 설정 후 해당 프로젝트가 열려있는 상태에서만 활성화되는 프로젝트 옵션(Project Options)으로 나뉘는데, 본 가이드북에서는 글로벌 옵션 중 주로 사용하는 옵션을 소개하고자 한다.

① 자동 줄바꿈 옵션 설정하기

스크립트 창에서 코드가 길어질 경우 자동으로 줄바꿈을 해주는 매우 유용한 옵션이다.

[Tools → Global Options]를 클릭한 후 [Code]탭에서 [Soft-wrap R source files] 항목을 체크해주면 된다.

② 한글 인코딩 오류 해결하기

스크립트 창의 한글이 깨질 경우가 있는데, 이는 인코딩 문제로 [Tools → Global Options]를 클릭한 후 [Code]탭에서 [Saving]탭으로 이동한 후 [Default text encoding] 항목의 [Change] 버튼을 클릭해 [UTF-8]이나 [Show all encoding]을 체크하고 1/3 지점의 [EUC-KR]을 체크해주면 된다.

③ 글꼴 및 테마 설정하기

[Tools → Global Options]를 클릭한 후 [Appearance]탭에서 사용자의 취향에 따라 글꼴 종류, 글자 크기, 배경 화면 등을 설정할 수 있다.

④ 창 위치 설정 변경하기

[Tools → Global Options]를 클릭한 후 [Pane Layout]탭에서 창의 위치를 변경할 수 있다.

⑤ RStudio가 실행될때마다 워킹 디렉토리를 자동으로 불러올 수 있도록 설정

[Tools → Global Options]를 클릭한 후 Options창이 열리면 [General]에서 [Default working directory] 항목의 [Browse] 버튼을 클릭하고 원하는 경로로 지정하고 최종적으로 [OK] 버튼을 클릭한다.

3. 분석 목적에 따른 통계분석 방법 요약

자료를 올바르게 활용하기 위해서는 연구 가설에 맞는 적합한 통계분석 방법을 찾고, 분석을 진행하는 것이 중요하다. 다음은 의학 및 보건학 연구에서 많이 사용되고 있는 분석 방법과 목적, 세부 내용이다.

3-1. 분석 목적에 따른 통계분석 방법

분석 방법	분석 목적	세부 내용
기술통계	자료의 전반적인 특성 파악	<ul style="list-style-type: none"> 백분율, 평균, 표준편차, 표준오차 등 (* 사례보고, 임상연구, 치료결과분석 등에 활용 가능)
분할표 분석	범주형 변수의 관련성 분석	<ul style="list-style-type: none"> Chi-square test(=Pearson's chi-square test) : 독립된 범주형 자료를 분할표로 나눌 수 있는 경우 McNemar's test(두 군) Stuart-Maxwell test(세 군) : 짝을 지은 범주형 자료를 분할표로 나눌 수 있는 경우 Fisher's Exact test(비모수적 방법) : 분할표의 셀이 하나라도 5보다 작은 경우)
t-tests	집단 간 평균 비교	<ul style="list-style-type: none"> One-sample t-test : 1개 집단의 평균값과 알려진 특정 값의 비교 Two-sample t-test : 독립적인 두 집단의 평균 비교 (* 비모수적 방법 : Mann-Whitney U-test) Matched-pair t-test : 짝을 지은 두 집단의 평균 비교 (* 비모수적 방법 : Wilcoxon signed rank test)
분산분석	세 집단 이상의 평균비교	<ul style="list-style-type: none"> ANOVA(analysis of variance) - 분산분석 : 셋 이상의 모집단의 모평균에 차이가 있는지 검정 ANCOVA(analysis of covariance) - 공분산분석 : 다변량 분석의 통합 모델로, 관심 독립변수 외의 변수는 통제하여 셋 이상의 모집단의 모평균의 차이 분석 (* 비모수적 방법 : Kruskal-Wallis test) MANOVA(multivariate analysis of variance) : 두 개 이상의 종속 변수가 서로 관련된 상황에 적용시키는 것으로 셋 이상의 집단 간 평균 차이 검증
상관분석	두 연속 변수 간 상관관계 분석	<ul style="list-style-type: none"> Pearson's correlation : 두 개의 연속형 변수 사이 선형관계에 대한 분석 방법 (* 비모수적 방법 : Spearman correlation test)
편상관분석	제3의 교란요인을 반영한 상관관계 분석	<ul style="list-style-type: none"> Partial correlation : 제3의 교란요인을 반영하여 분석하고자 하는 경우

분석 방법	분석 목적	세부 내용
회귀분석	연속변수 사이의 관련성 확인	<ul style="list-style-type: none"> • Simple linear regression(단순 선형 회귀분석) : $Y = \beta_0 + \beta_1 X$ - 독립변수 1개(연속형 변수) - 종속변수 1개(연속형 변수)
		<ul style="list-style-type: none"> • Multiple linear regression(다중 선형 회귀분석) : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$ - 독립변수 n개(연속형 변수) - 종속변수 1개(연속형 변수)
		<ul style="list-style-type: none"> • Logistic regression(로지스틱 회귀분석) : $P_x = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k}}$ - 독립변수 n개(연속 또는 범주형 변수) - 종속변수 1개(이분된 범주형 변수)
생존분석	어떤 사건a이 발생할 때까지의 시간b으로 자료가 주어진 경우 a 예: 사망, 질병 발생 b 예: 사망까지의 생존시간	<ul style="list-style-type: none"> • Life table method(생명표법) - 고정된 시간간격을 적용하는 방법으로, 관찰 대상자 수가 많을 때(각 군의 자료가 50개 이상) 적절한 방법임 - Event 발생과 상관없이 원하는 생존율의 기간을 정할 수 있음
생존분석	어떤 사건a이 발생할 때까지의 시간b으로 자료가 주어진 경우 a 예: 사망, 질병 발생 b 예: 사망까지의 생존시간	<ul style="list-style-type: none"> • Kaplan-Meier method / Log-rank test - 사건이 발생한 시점에 대해서 생존곡선이 추정됨 - 정확한 생존율을 구할 수 있으며, 사례가 소규모(각 군의 자료가 50개 이하)인 경우에 적절한 방법임
생존분석	어떤 사건a이 발생할 때까지의 시간b으로 자료가 주어진 경우 a 예: 사망, 질병 발생 b 예: 사망까지의 생존시간	<ul style="list-style-type: none"> • Cox 's proportional hazard model - 생존기간에 영향을 주는 인자에 대한 영향력 검정

3-2. 분석 목적에 따른 통계분석 방법 (R 사용)

변수유형	종속변수 (Y)	
	범주형	연속형
독립 변수 (X)	범주형	T-검정 t.test()
		예: 성별(남, 녀)에 따라 공복혈당 수준은 차이가 있을까?
	연속형	카이제곱 검정 chisq.test()
		예: 성별(남, 녀)과 당뇨병 유병여부(당뇨병 유병자, 당뇨병 비유병자)는 관련성이 있을까?
독립 변수 (X)	범주형	분산분석 aov() / oneway.test()
		예: 비만도(저체중, 정상체중, 과체중, 비만)에 따라 공복혈당 수준은 차이가 있을까?
	연속형	상관분석 cor.test()
독립 변수 (X)	연속형 & 범주형	예: BMI와 공복혈당은 상관성이 있을까?
		로지스틱 분석 glm()
독립 변수 (X)	연속형 & 범주형	회귀분석 lm()
		예: 연령이 증가할수록 공복혈당은 증가할까?

한국인유전체역학조사사업 (KoGES)
Korean Genome and Epidemiology Study

KoGES 데이터 분석 가이드북

R편

발 행 : 2020년 9월

편 집 : 김성수, 송대섭, 이두리, 이소현, 정은주, 조미진, 최선호

발행처 : 질병관리청 국립보건연구원 유전체역학과

| 주 소 | 충청북도 청주시 흥덕구 오송읍 오송생명2로 200
국립중앙인체자원은행 3층

| 전화번호 | 043-719-6710

| 팩 스 | 043-719-6759

KoGES 데이터 분석 가이드북



질병관리청 국립보건연구원 미래의료연구부 유전체역학과
충청북도 청주시 흥덕구 오송읍 오송생명2로 200

비매품/무료



9 788968 388491

ISBN 978-89-6838-849-1 (PDF)