‹건강검진코호트DB Meeting 내용 압축 정리›

- 2023.02.21. ~ 2023.03.22.까지의 내용 압축 정리

- 작성자 : 이은경

*** 2023.02.20. 이후의 내용 중, 필요한 부분만 압축하여 정리한 Document입니다.**

‹주요 안건›

① 정의한 Covariate Missing Imputation 방안 고려한 후, Imputation 과정 진행

② ① 과정 진행한 후 Covariate Missing rate 다시 확인 — Final Version

③ 분석용 Data 형태 & Data를 구성하는 변수들의 요약 통계량(Table 1) 정리

④ "One-time BMI Exposure"(Cohort entry date 당시의 BMI 측정치) 기준으로 BMI 관측치 Grouping 한 후, "Moderate" group을 baseline으로 하여 Naive하게 Coxph model 적합 — 2가지 Version이 있음.

⑤ Study population에 포함되는 객체들의 Cohort entry date 이후 BMI, WAIST Trajectory와 측정 횟수 분포 파악

---

1) Covariate Missing value Imputation & Missing Rate re-check

: Missing value를 대체할 때 "Cohort entry date 이전 기록 중 "결측이 아닌" 가장 최근의 기록 가져오는 방안"을 선택하였다.

↳ "SODA" 섭취 빈도 공변량, "Snack 섭취 빈도 공변량", "과일 & 채소 섭취 빈도 공변량"은 결측률 자체가 높고, 해당 변수를 확인할 수 있는 문진항목이 별로 없어 세 공변량은 Coxph model 적합 때 아예 고려하지 않기로 결정했으므로 결측 대체 과정 진행하지 않음.

1)-①. "Demographic" 관련 공변량 : Missing인 관측치는 없는 것으로 확인됨.

1)-②. "가족력" 공변량

: 가족력의 경우, 위에서 언급한 대체 방안으로도 결측을 대체할 수 없는 관측치는 해당 가족력이 없다고 판단하고, 변수 값을 모두 "0"으로 대체함. -- Missing인 관측치는 없음.

1)-③. "Biomarker" 공변량

: 결측 대체 과정 거친 후 재파악한 변수 별 결측률은 아래 표와 같다.

Biomarker	Missing obs / N
TOT_CHOLE	168 / 168,339 (0.099%)
BLDS	2 / 168,339 (0.0011%)
OLIG_PROTE_CD	641 / 168,339 (0.38%)
SGOT_AST	30 / 168,339 (0.017%)
SGPT_ALT	29 / 168,339 (0.017%)
GAMMA_GTP	26 / 168,339 (0.015%)

My Guess) "TOT_CHOLE" 변수, 총콜레스테롤의 경우 혈액검사를 통해 해당 지표를 검사하는 연령대가 제한적이다. (20대인 경우, 혈액검사를 통해 총콜레스테롤 지표를 알 수 없음) -- 총콜레스테롤 측정대상이 아니어서 생긴 결측일 수 있지 않을까?

⇒ 나머지 변수의 경우, 결측 값을 특정 값으로 대체할 수 없는 방안이 딱히 없고, 결측 비율 또한 낮기에 결측인 관측치는 삭제해도 무방하다고 판단됨.

1)-④. "음주습관" 공변량
: 위에서 언급한 방안으로 결측이 대체되는 관측치는 없었으며, "음주습관"이 결측인 객체는 음주를 하지 않는다고 판단해 값을 모두 "0"으로 변경했다.

1)-⑤. "운동습관"(EXERCI_HABIT_COV) / "운동지속시간"(METS_minutes) 공변량
: 검진DB 자체에서 운동습관 변수를 모두 정의한 후, 이를 study population tbl과 joint 한 다음에 결측이 아닌 가장 최근의 기록 가져오는 방식으로 결측 대체 진행
→ 해당 logic 적용 후에도 여전히 4,007명의 관측치는 "운동습관" 변수가 결측임을 확인
⇒ 해당 관측치는 운동을 하지 않는다고 판단하고, 모두 "0"으로 변환
∴ 따라서, "운동습관"이 결측인 관측치는 0개이다.

: "운동 지속시간"을 정의할 때 이전에는 "WLK30_WEK_FREQ_ID"(30분 이상 걷기) 변수를 고려하지 않았었음.
(∵ 18~19년도 검진DB에는 해당 변수가 없으므로 변수 생성 과정 통일성 위해)
그러나, Cohort entry date 이전 기록 중 18~19년 기록은 없으므로 METS_minutes 계산할 때 "WLK30_WEK_FREQ_ID" 변수도 추가로 고려함.

⇒ 해당 과정을 통해 총 312개의 관측치가 특정 값으로 대체되었다.
∴ Imputation 방안 통해 결측치 대체 결과, Missing인 obs는 15,016명이다. (8.92%)

My Guess) "운동 지속시간" 변수에 Missing value가 존재하는 이유 중 하나는 2008년도 생애 전환기 검진대상이 아니어서 파악할 수 없었기 때문이 아닐까?
(∵ 2009년 이전 기록에 대해 "운동 지속시간"을 정의할 때는 2008년도 생애 전환기 검진DB만 이용하였기 때문)
∴ 해당 대체 과정 진행 후, "운동습관" = 0이면서 "운동 지속시간" 변수가 결측이면 "운동 지속시간" 변수 값을 0으로 대체함.

---

1)-⑥. "흡연상태", "흡연 기간", "흡연량" 변수
: "흡연상태" 변수가 결측이면, (과거) / (현재) 흡연 기간, (과거) / (현재) 흡연량 변수 전부가 결측임을 모든 검진DB에서 확인함. 따라서, 해당 객체들은 모두 "Never Smoker"로 간주해도 무방하다고 판단

: "흡연 기간", "흡연량" 변수에 결측이 존재하는 원인 중 가장 큰 원인은 2002년~2008년 사이 과거 흡연자들 대상, "PAST_DSQTY", "PAST_PACK_YEAR" 변수가 모두 결측인 상황이다.
∵ 2002~2008년도 검진DB에는 (과거) / (현재) 흡연량을 조사하는 변수가 없음. (흡연 기간 변수만 존재함)

⇒ 흡연 관련 변수 중 하나라도 결측이 있는 관측치 개수가 총 12,929개인데 이 중 12,928개가 위에서 제시한 "과거 흡연자" case 때문이다. / 나머지 1개의 관측치는 현재 흡연자이나 현재 흡연 관련 변수가 모두 결측임.

※ 현재 가지고 있는 검진DB로 흡연 관련 정보를 가지고 올 수 있는 객체는 총 164,036명이다.
: 흡연 관련 변수 정의할 때 "흡연상태" 변수 값이 '1', '2', '3' 인 경우로 데이터 나누어 공변량 정의하였음.
⇔ "흡연상태"가 결측인 경우 고려하지 못해서 발생한 것이라 추측됨.
/ 따라서, 해당 tbl에 포함되지 않은 객체들의 흡연 관련 변수는 모두 Never 흡연자와 동일한 방식으로 처리함.

: Imputation 과정 거친 후, **현재 흡연자 대상**, “흡연 기간”, “흡연량” 관련 공변량의 Missing rate는 아래 표와 같다.

Covariate	Missing obs / N
흡연 기간	1 / 168,339
흡연량	1 / 168,339
PACK YEAR	1 / 168,339

: Imputation 과정 거친 후, **과거 흡연자 대상**, “흡연 기간”, “흡연량” 관련 공변량의 Missing rate 또한 아래 표와 같다.

Covariate	Missing obs / N
흡연 기간	0 / 168,339
흡연량	12,928 / 168,339
PACK YEAR	12,928 / 168,339

---

1)-⑦. “개인 과거 병력” 공변량

: 해당 공변량이 결측인 객체는 과거 병력이 없다고 판단하고 모두 값을 “0”으로 처리하였다.
따라서, Missing인 obs는 0명이다.

cf) 개인 과거 병력이 존재하는 객체 비율

개인 과거 병력	객체 수
Myocardial infraction	0 / 168,339
Heart failure	0 / 168,339
TIA or Stroke	0 / 168,339
Hypertension	55,970 / 168,339
Cancer	0 / 168,339

※ 개인 과거 병력이 존재하는 객체가 없는 이유 추측
: “MI”, “Heart failure”, “TIA or Stroke”, “Cancer”의 경우 Study population 정의할 때 제외조건으로 적용했던 병명으로, Cohort entry date 이전에 해당 병명이 있는 객체들은 모집단에서 제외했으므로 과거 병력이 존재하는 객체가 없는 것은 당연해 보인다. 이때, “Hypertension”의 경우는 제외조건에 포함했던 병명이 아니었기 때문에 과거 병력이 잡히는 것으로 파악됨.

1)-⑧. Comorbidities / 약물 처방 이력 공변량

: Comorbidities, 약물 처방에 관련된 변수들이 결측인 객체는 해당 이력이 없다고 판단하고 모두 값을 “0”으로 처리함.
따라서, 결측인 객체는 없다.

2) 분석용 Data 재생성

: 기존에는 가장 빨리 발생한 Outcome 종류 & 해당 Outcome 발생 때까지 걸린 시간만 저장하였으나,
재생성한 Data에는 고려하는 모든 Outcome에 대해 Outcome 발생 여부 & Cohort entry date에서 Outcome 발생 때까지 걸린 시간 변수 모두 추가 — 만약, Outcome이 발생하지 않았다면, Last follow Up(2019.12.31.)까지의 기간 계산.

: 추가로, Cohort entry date와 Disease onset date 간 Duration 변수도 추가하였다.
∵ 해당 변수도 Coxph model 적합 때 공변량으로 사용

[재생성한 Data 형태]

PERSON_ID	Cohort_entry_Duration	Death	FU_Duration_death	MI	MI_Duration	Cancer	Cancer_Duration	Heart_failure	HF_Duration	TIA_Stroke
10000355	1.7002053388	1	5.3196440794	0	16.227241615	0	16.227241615	0	16.227241615	0
10000821	0	1	3.9288158795	0	16.109514031	0	16.109514031	0	16.109514031	0
10002153	0	1	5.8535249629	0	16.646132786	0	16.646132786	0	16.646132786	0
10003229	1.8206707734	1	14.020533881	0	16.169746749	0	16.169746749	0	16.169746749	0
10003681	6.1464750171	1	7.7754962355	0	10.064339493	0	10.064339493	0	10.064339493	0
10005231	0.0164271047	1	8.9911019849	0	16.271047228	0	16.271047228	0	16.271047228	0
10006251	0.9472963723	1	6.5270362765	0	16.465434634	0	16.465434634	0	16.465434634	0
10006326	0	1	6.3436002738	0	16.580424367	0	16.580424367	0	16.580424367	0
10006642	1.3579739904	1	9.697467488	0	16.574948665	0	16.574948665	0	16.574948665	0
10007049	0.9993155373	1	15.26899384	0	16.574948665	0	16.574948665	0	16.574948665	0
10007259	1.4099931554	1	14.069815195	0	16.481861739	0	16.481861739	0	16.481861739	0
10007371	0.7091033539	0	14.67761807	0	14.67761807	0	14.67761807	0	14.67761807	0
10007707	0	1	7.5099247091	0	16.057494867	0	16.057494867	0	16.057494867	0
10008531	0.1861738535	1	5.6399726215	0	16.542094456	0	16.542094456	0	16.542094456	0
10008871	1.4182067077	1	5.5550992471	0	16.396988364	0	16.396988364	0	16.396988364	0

TIA_Stroke	TIA_Duration	Outcome_Duration	Outcome_Category	BMI	Category	SEX	AGE	CTRB_PT_TYPE_CD	BLDS	TOT_CHOLE	
0	16.227241615	5.3196440794	Death	24.54	Moderate	1	80	8	76	156	1
0	16.109514031	3.9288158795	Death	19.84	Moderate	2	80	3	137	182	1
0	16.646132786	5.8535249629	Death	18.86	Moderate	1	80	1	149	180	1
0	16.169746749	14.020533881	Death	22.81	Moderate	2	80	1	168	187	1
0	10.064339493	7.7754962355	Death	18.67	Moderate	2	86	10	86	211	1
0	16.271047228	8.9911019849	Death	21.60	Moderate	1	80	10	86	219	1
0	16.465434634	6.5270362765	Death	24.44	Moderate	2	80	9	90	255	1
0	16.580424367	6.3436002738	Death	19.30	Moderate	2	80	1	139	190	1
0	16.574948665	9.697467488	Death	25.85	Overweight	2	80	2	280	244	1
0	16.574948665	15.26899384	Death	22.07	Moderate	2	80	6	120	193	1
0	16.481861739	14.069815195	Death	20.70	Moderate	2	80	3	72	224	1
0	14.67761807	14.67761807	Last_follow	16.89	Underweighted	2	82	8	78	204	1
0	16.057494867	7.5099247091	Death	20.93	Moderate	2	80	1	131	174	1
0	16.542094456	5.6399726215	Death	16.49	Underweighted	1	80	10	89	264	1
0	16.396988364	5.5550992471	Death	18.07	Underweighted	2	80	4	271	279	1
0	16.424366872	16.424366872	Last_follow	20.54	Moderate	2	80	9	214	212	1
0	12.128678987	12.128678987	Last_follow	22.15	Moderate	1	84	10	91	173	1
0	16.358658453	7.9663107461	Death	22.83	Moderate	1	80	9	109	135	1
0	12.607802875	7.40862423	Death	24.52	Moderate	1	84	6	142	154	1
0	16.687200548	3.6577666516	Death	28.13	Overweight	2	80	10	91	246	1
0	16.361396304	2.5160848734	Death	27.31	Overweight	1	80	2	67	206	1
0	16.002737851	12.969199179	Death	25.33	Overweight	2	80	9	93	197	1
0	16.164271047	16.164271047	Last_follow	19.23	Moderate	1	80	1	127	188	1

---

3) Data를 구성하는 변수들의 요약 통계량 파악

: "Table1_Version0.xlsx" file에 자세하게 기록함.

〈주목할만한 점〉

1) Study population에 포함되는 객체 중 대부분은 Exposure(One time BMI) Category가 "Moderate", "Overweighted"로 분류된다.

2) 고려하는 모든 종류의 Secondary Outcome에 대해 BMI 수치가 증가할수록 Cohort entry date에서 Outcome이 발생하기까지의 기간 또한 길어진다. (Exposure Category 간 유의미한 차이를 보이는지는 알 수 없음 / Outcome이 발생하지 않은 사람들의 Duration 값도 포함되어 있기 때문이다. -- Last follow Up date는 2019.12.31.이다.)

3) "AGE", "BLDS", 과거 / 현재 흡연자 대상 흡연 기간("SMK_TERM"), "EXERCI_HABIT"(운동습관) 수치 경향이 BMI Category와 반비례하다.
: BMI 수치가 증가할수록 각 변수의 중심값이 줄어드는 경향을 보인다. (각 Category 별 차이가 유의미한지는

파악하지 않음.)

4) 과거 흡연자 대상, 흡연량("DSQTY", "PACK_YEAR") 수치 경향 또한 BMI Category와 반비례하다.
: BMI 수치가 증가할수록 각 변수의 중심값이 줄어드는 경향이 눈에 띈다. (이 또한, 각 Category 별 차이가 유의미한지는 파악하지 않음)

5) Numerical variable과 다르게 Categorical variable의 경우 모든 변수, 범주에 대해 "Moderate" - "Overweighted" - "Obesity" - "Underweighted" 순서와 N수(%) 순위와 동일하다. (변수별 특별한 특징이 보인다기보다는 해당 Category에 속하는 사람들이 많을수록 병력 / 약물 처방 이력이 잡히는 듯 하다.)

---

4) One-time BMI grouping 기준으로 Naive하게 Coxph model 적합

: Reference paper 중 하나인 "bmj.i2195"에 "Trajectory-mortality association(for all-cause mortality)이 never-smoker group보다 ever-smoker group에서 더 약하게 나타남", "normal-weight 대비, Overweight group의 all-cause mortality HR이 더 낮음"이라고 서술되어 있음.

↳ 해당 내용을 토대로 Categorical BMI Exposure 이용해 Coxph model 적합할 때 BMI Category Baseline은 "Moderate" group으로 지정하였으며, "Total Study population", "Never-Smoker", "Current-Smoker" 세 group을 대상으로 모형을 적합해 보았다.

⇒ 자세한 결과는 "Coxph_result_Version0.xlsx", "Coxph_result_Version1.xlsx" file에 기록되어 있다.
※ Version 0은 "SODA", "SNACK", "FRUIT & VEGETABLES" 변수 제외, 모든 공변량을 이용해 적합한 Coxph 결과를 정리한 파일이며, Version 1의 경우에는 Version 0에서 "흡연량(DSQTY)", "PACK YEAR" 변수를 추가로 제외하였다.
(∵ 과거 흡연자의 경우 과거 흡연량 / PACK YEAR 변수를 정의할 수 없는 관측치가 존재함.)

cf) 참고논문을 살펴보면 흡연 관련 변수가 중요한 공변량인 것으로 보인다.

---

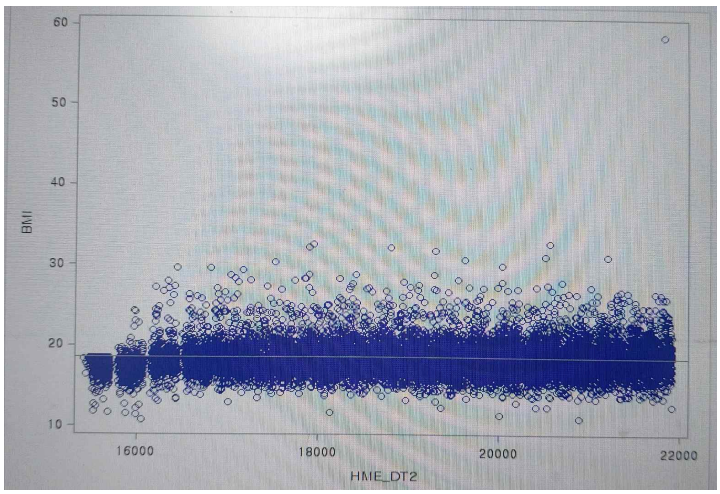
5) Study population에 포함되는 객체들의 Cohort entry date 이후 BMI, WAIST Trajectory와 측정 횟수 분포 파악

5)-1) BMI Trajectory

1)-①. "Under-weighted" group (One-time BMI Exposure 기준으로 BMI Category 나눔)
(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
13,831	18.355	1.990	10.8184	17.2383	18.0977	19.1484	58.8984

(Scatterplot)
: 기준선은 BMI=18.5를 의미함.



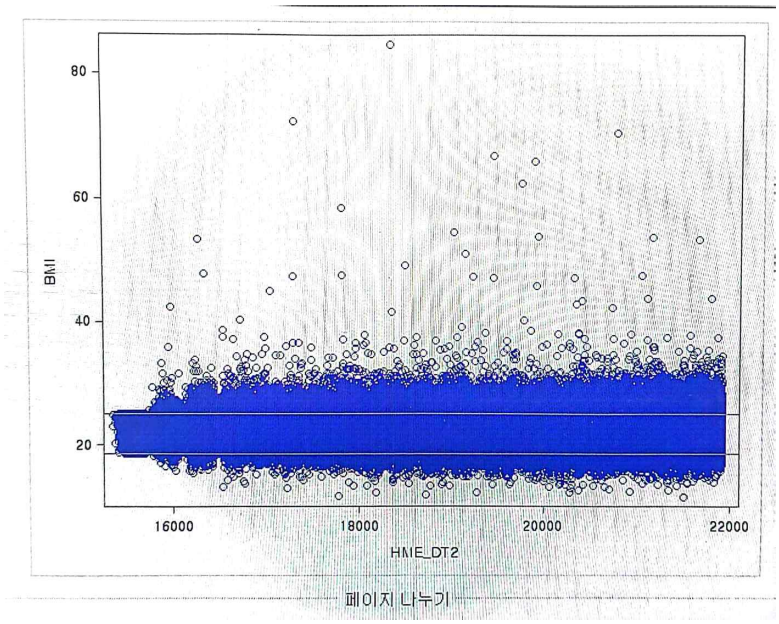
1)-②. “Moderate” group (One-time BMI Exposure 기준 BMI Category)

(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
620,473	22.89343	1.97712	11.3984	21.6172	23.0	24.218	84.6250

(Scatterplot)

: 기준선은 각각 BMI=18.5, BMI=25를 의미함.



---

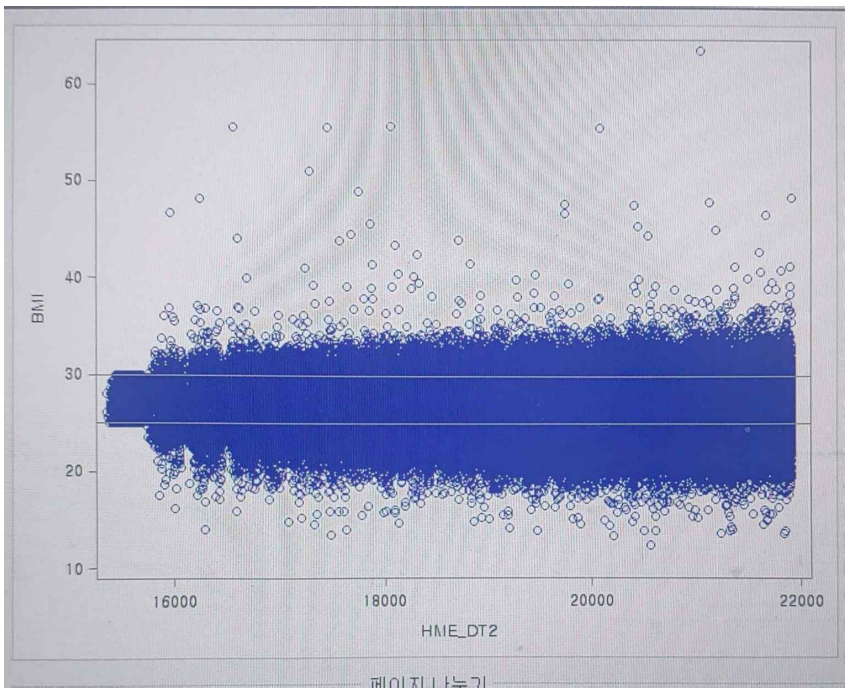
1)-③. “Overweighted” group (One-time BMI Exposure 기준 BMI Category)

(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
476,682	26.45350	1.97223	12.2988	25.2188	26.3672	27.6787	63.5

(Scatterplot)

: 기준선은 각각 BMI=25, BMI=30을 의미함.



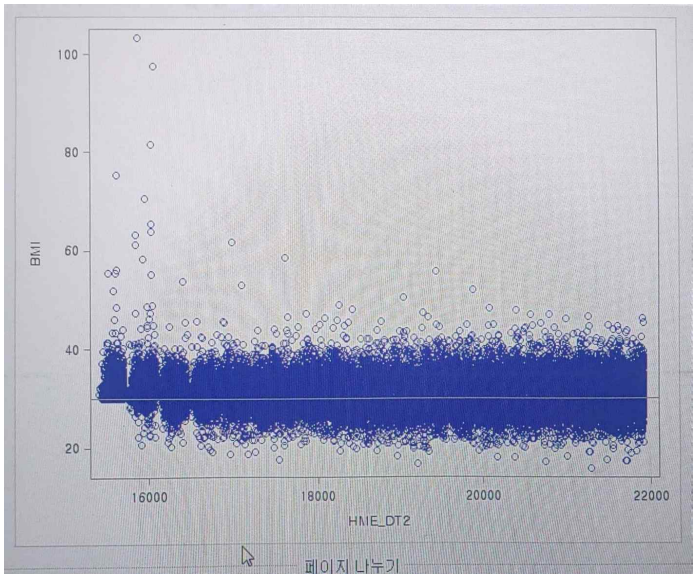
1)-④. “Obesity” group (One-time BMI Exposure 기준 BMI Category)

(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
54,460	30.928	2.825	15.699	29.3984	30.7969	32.3672	103.3594

(Scatterplot)

: 기준선은 BMI=30을 의미함.



---

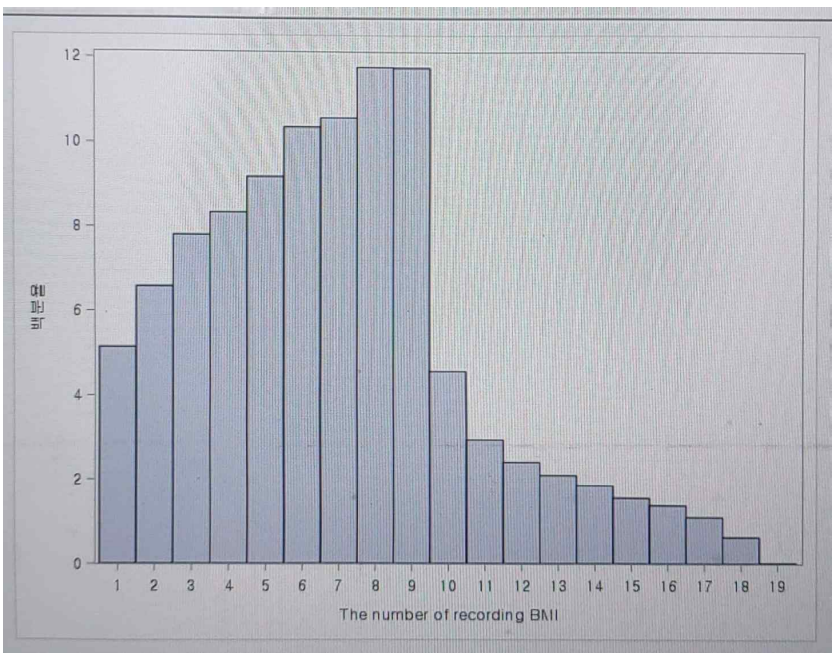
5)-2) BMI 측정횟수

(Summary 값)

mean	SD	Min	Q1	Q2	Q3	Max
6.923	3.6999	1	4	7	9	19

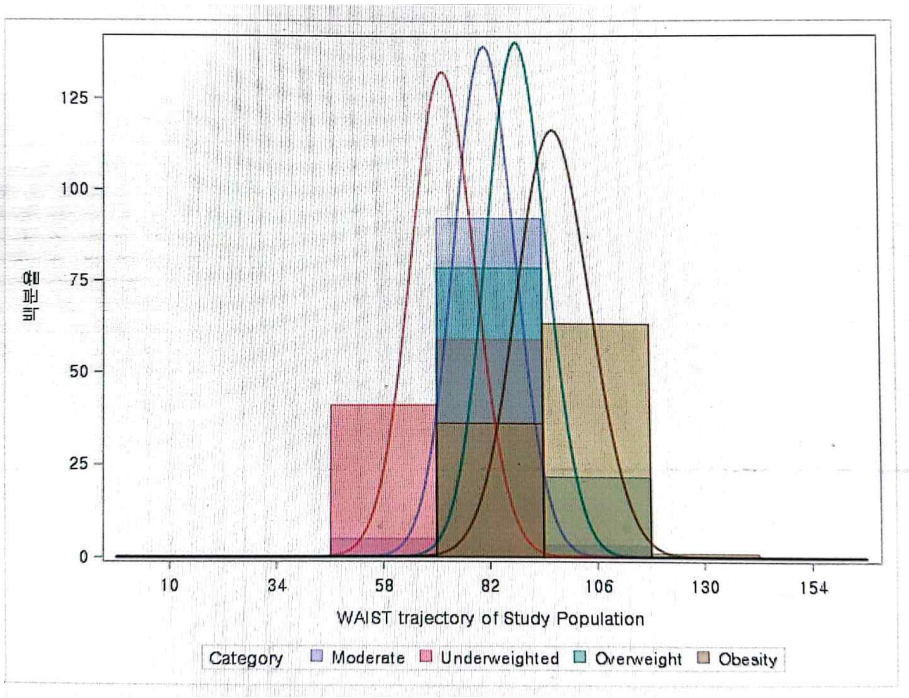
↳ 이때, BMI 측정 횟수가 1번인 객체는 총 8,658명 / 19번인 객체는 총 35명이다.

(Histogram)



5)-3) WAIST Trajectory

(WAIST 측정치 Histogram)



---

3)-①. “Underweighted” group (One-time BMI Exposure 기준 BMI Category)

(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
9,787	71.4487	7.231	50	66	71	76	114

3)-②. “Moderate” group

(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
447,677	81,0181	6.866	30	76	81	86	142

3)-③. “Overweight” group

(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
337,415	88,267	6.813	0	84	88	93	129

↳ 최솟값 “0”은 측정오류로 파악된다. 0 다음으로 작은 값이 38이므로 최솟값은 “38”로 보는 것이 더 적절해 보인다.

3)-④. “Obesity” group

(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
37,829	96.4855	8.224	55	91	96	102	137

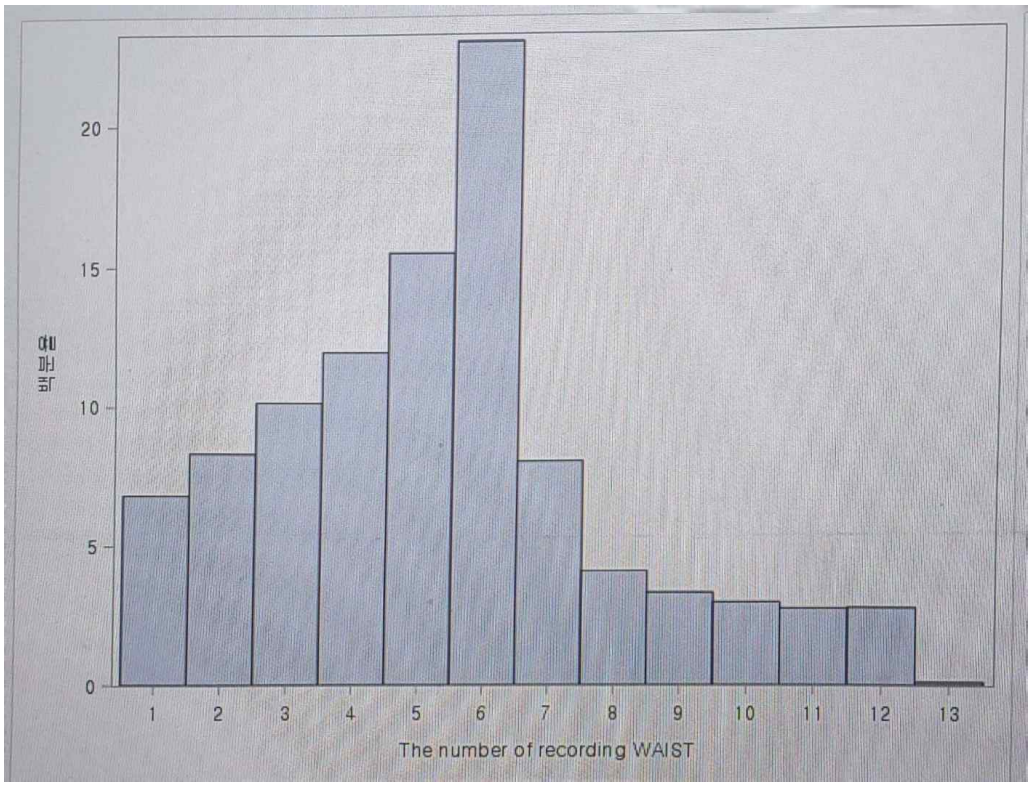
5)-4) WAIST 측정횟수

(Summary 값)

mean	SD	Min	Q1	Q2	Q3	Max
5.3133	2.6217	1	3	5	6	13

↳ 이때, WAIST 측정 횟수가 1번인 객체는 총 10,693명 / 13번인 객체는 총 191명이다.
/ Study population에 포함되는 객체 168,339명 중 156,742명에 대해서만 WAIST 지표가 측정되었다.

(측정횟수 Histogram)



(주목할만한 점)

: WAIST가 2008년부터 측정된 지표라고 알고 있으나, 2007년 생애전환기검진DB 항목에 "WAIST" 지표가 존재함. 따라서, WAIST가 측정된 검진 일자를 보면 2007년대도 존재하며, 2007년도에 측정된 WAIST 건수는 총 2,414건이다.

〈주목할만한 점〉

1. BMI Trajectory

- 대부분 객체가 처음 분류된(One time BMI Exposure 기준으로 분류) BMI Category에 머무르긴 하나, 체중 증가 혹은 감소한 객체가 분명히 존재하긴 함. (Min, Max 값 기준)

- Category가 "Moderate"에 분류된 사람들의 체중 변화가 가장 눈에 띈다. (∵ 최댓값과 최솟값의 차이가 제일 크다)
"Moderate" group의 BMI 최댓값이 "Overweighted" group의 BMI 최댓값보다 크다.

2. WAIST Trajectory

- BMI처럼 Category = "Moderate"인 group을 집중해서 볼 필요가 있어 보인다. "Moderate" group의 WAIST 최댓값이 가장 크며, "Moderate" group의 WAIST 최솟값이 "Under-weighted" group WAIST 최솟값보다 적다.