<건강검진코호트DB 결과 정리>

- 3월 9일 Version
- 작성자 : 이은경

<What TO DO>

: 결측이 존재하는 공변량 중, 대체가 가능한 공변량의 경우 Imputation을 진행하고, Coxph model 재적합하기

: Never 흡연자, 현재 흡연자들 대상으로 Coxph model 적합 해보고 결과 확인하기

: Study population 대상으로 Cohort entry date 이후의 BMI, WAIST trajectory 가져오기.

<Share & Result>

1) Biomarker 관련 변수

: Cohort entry date 이전 기록 중 "결측이 아닌" 가장 최근의 기록 가져오는 것으로 결측 대체 진행

-- "BLDS" 변수 외의 해당 대체 방안으로 결측치 값이 채워진 변수 없음.
("BLDS"의 경우 1명 대체 가능했음.)

Question) Biomarker 관련 변수 중 "TOT_CHOLE" 변수가 존재, 총콜레스테롤을 의미하는데, 혈액검사를 통해 해당 지표를 검사하는 연령대가 제한적이다. (20대인 경우, 혈액검사를 통해 총콜레스테롤 지표를 알 수 없음.) -- 총콜레스테롤을 측정하는 대상이 아니어서 발생한 결측 일 수 있지 않을까?

---

2) 음주습관(DRINK_FREQ_COV)

: Cohort entry date 이전 기록 중 "결측이 아닌" 가장 최근의 기록 가져오는 것으로 결측 대체 진행

-- 해당 과정으로 결측이 대체되는 관측치는 없었음.

: 위 과정으로도 대체되지 않는 값은 모두 "0"으로 처리

---

3) 운동습관(EXERCI_HABIT_COV) / 운동지속시간(METS_minutes)

: 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 값은 0으로 변환

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

⇒ 해당 과정을 통해 총 312개의 관측치가 특정 값으로 대체되었다.
/ 그러나 여전히 15,818명의 Missing value 존재

My Guess) Missing value가 존재하는 이유 중 하나는 2008년도 생애 전환기 검진대상이 아니어서 파악할 수 없었기 때문이 아닐까? (∵ 2009년 이전에 "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 값 0으로 대체함.

---

4) 흡연상태 / 빈도 관련 변수
: 흡연상태가 결측인 경우, 모든 흡연 관련 변수 또한 결측 임을 확인 → 값 모두 "0"으로 대체
: 해당 변수의 경우 결측이 존재하는 원인 중 가장 큰 원인은 2002년~2008년 사이 과거 흡연자들 대상, "PAST_DSQTY", "PAST_PACK_YEAR" 변수가 모두 결측인 상황이다. (단, 과거 흡연자들 중 흡연 기간이 결측인 경우는 없음을 확인)

⇒ 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.
(Study population 수는 총 168,339명) 따라서, 해당 tbl에 포함되지 않은 객체들의 흡연 관련 변수는 모두 Never 흡연자와 동일한 방식으로 처리함.

---

5) SODA, SNACK, 과일 & 채소 섭취 빈도 공변량
: 해당 변수들은 결측 률이 너무 높기 때문에 Coxph model 적합 때 사용하지 않음.

6) Comorbidities, 약물 처방 이력, 개인 과거력
: 해당 변수들의 결측치는 모두 "0"으로 대체

cf) Study population 총 수는 168,339명이나, Cohort entry date에서 Death 사건이 발생하기까지 걸린 시간이 "음수"인 객체가 한 명 있어서 Coxph model fitting 할 때 해당 관측치는 아예 제거함.
-- 분석용 데이터 N수는 "168,338명"

[Coxph model fitting 결과 정리]

1. 결측치 대체한 후 다시 한 번 Naive하게 Coxph model fitting

1-1) 정의한 모든 공변량을 설명변수로 사용

1-1)-①. Categorical BMI

BMI Category	Parameter Estimate	P-value	HR
Moderate	-0.56006	< .0001	0.571
Overweight	-0.71997	< .0001	0.487
Obesity	-0.55826	< .0001	0.572

1-1)-②. Numerical BMI

Parameter Estimate	P-value	HR
-0.04122	<.0001	0.960

1-2) "AGE", "AGE^2", "SEX", "Cohort_entry_Duration" 공변량만 이용해 Coxph model 적합

1-2)-①. Numerical BMI

Parameter Estimate	P-value	HR
-0.05091	<.0001	0.950

1)-2-②. Categorical BMI

BMI Category	Parameter Estimate	P-value	HR
Moderate	-0.65645	< .0001	0.519
Overweight	-0.85977	< .0001	0.423
Obesity	-0.64702	< .0001	0.524

→ 결측치 대체 전의 결과와 크게 다르지 않다.

2. “Never” 흡연자들 대상으로 Coxph model Naive하게 fitting (결측 대체 과정 거친 Data 이용)

2-1) 정의한 모든 공변량을 설명변수로 사용

2-1)-①. Categorical BMI

BMI Category	Parameter Estimate	P-value	HR
Moderate	-0.576	< .0001	0.562
Overweight	-0.729	< .0001	0.482
Obesity	-0.541	< .0001	0.582

2-1)-②. Numerical BMI

Parameter Estimate	P-value	HR
-0.03237	<.0001	0.968

2-2) “AGE”, “AGE^2”, “SEX”, “Cohort_entry_Duration” 공변량만 이용해 Coxph model 적합

2-2)-①. Numerical BMI

Parameter Estimate	P-value	HR
-0.02867	<.0001	0.972

2-2)-②. Categorical BMI

BMI Category	Parameter Estimate	P-value	HR
Moderate	-0.56327	< .0001	0.569
Overweight	-0.70582	< .0001	0.494
Obesity	-0.46973	< .0001	0.625

[BMI, WAIST Trajectory 파악]

: Study population 대상으로 Cohort entry date 이후 BMI, WAIST Trajectory 파악

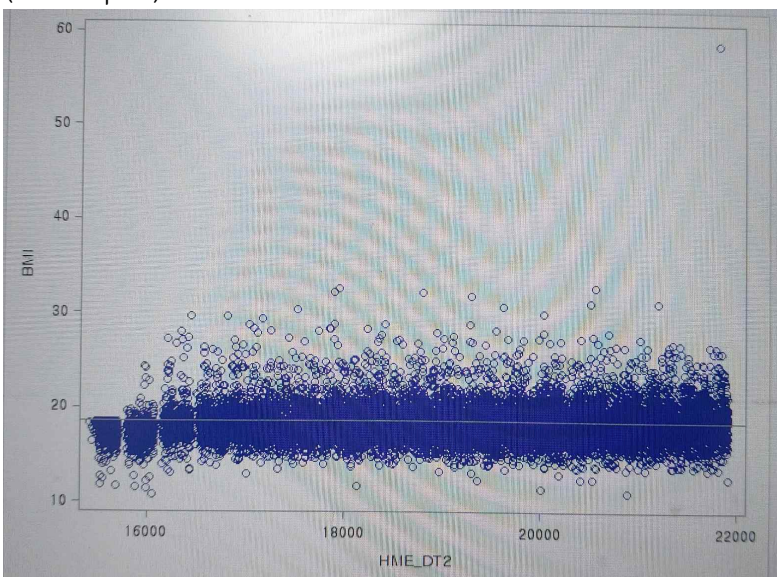
1. BMI

1-①. "Underweighted" group (이전에 정의한 "Category" 변수 기준)

(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)



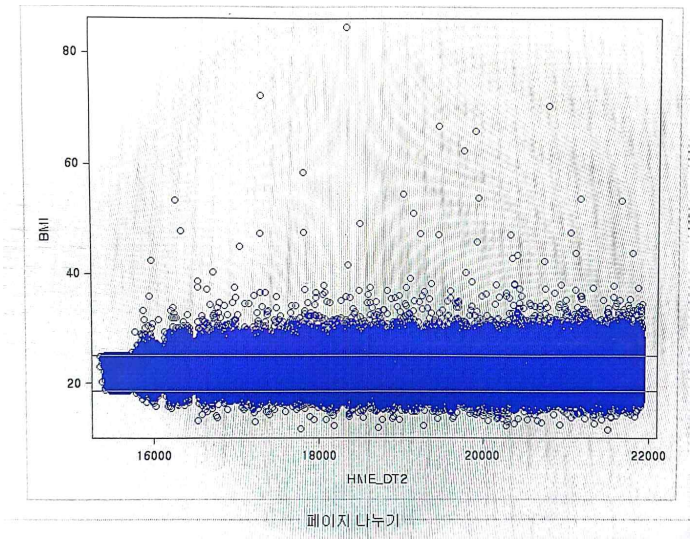
: 기준선은 BMI=18.5를 의미함.

1-②. "Moderate" group (이전에 정의한 "Category" 변수 기준)

(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)



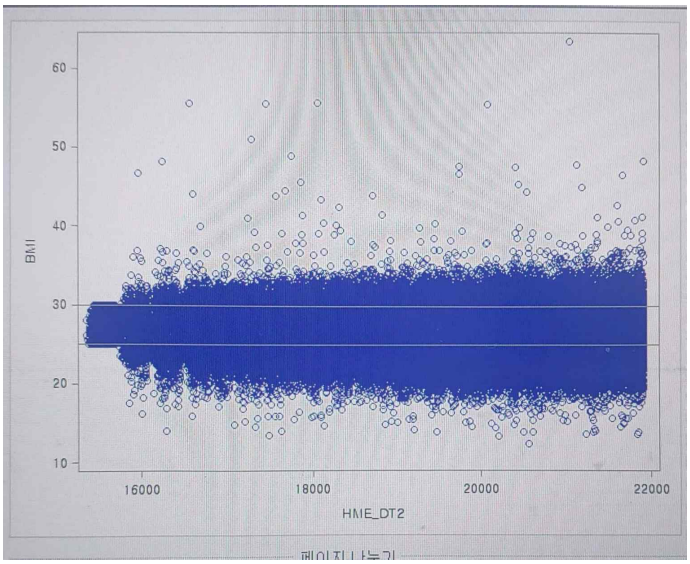
: 기준선은 각각 BMI=18.5, BMI=25를 의미함.

1-③. “Overweighted” group (이전에 정의한 “Category” 변수 기준)

(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)



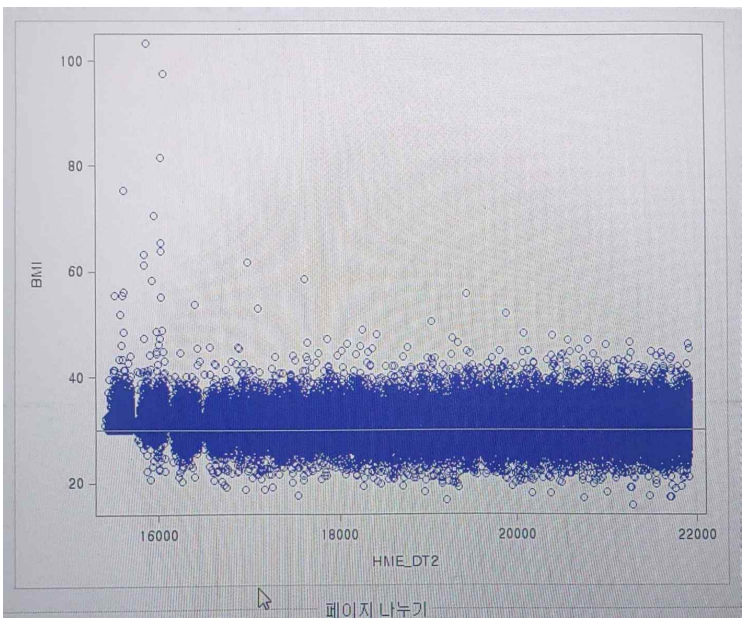
: 기준선은 각각 BMI=25, BMI=30를 의미함.

1-④. “Obesity” group (이전에 정의한 “Category” 변수 기준)

(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

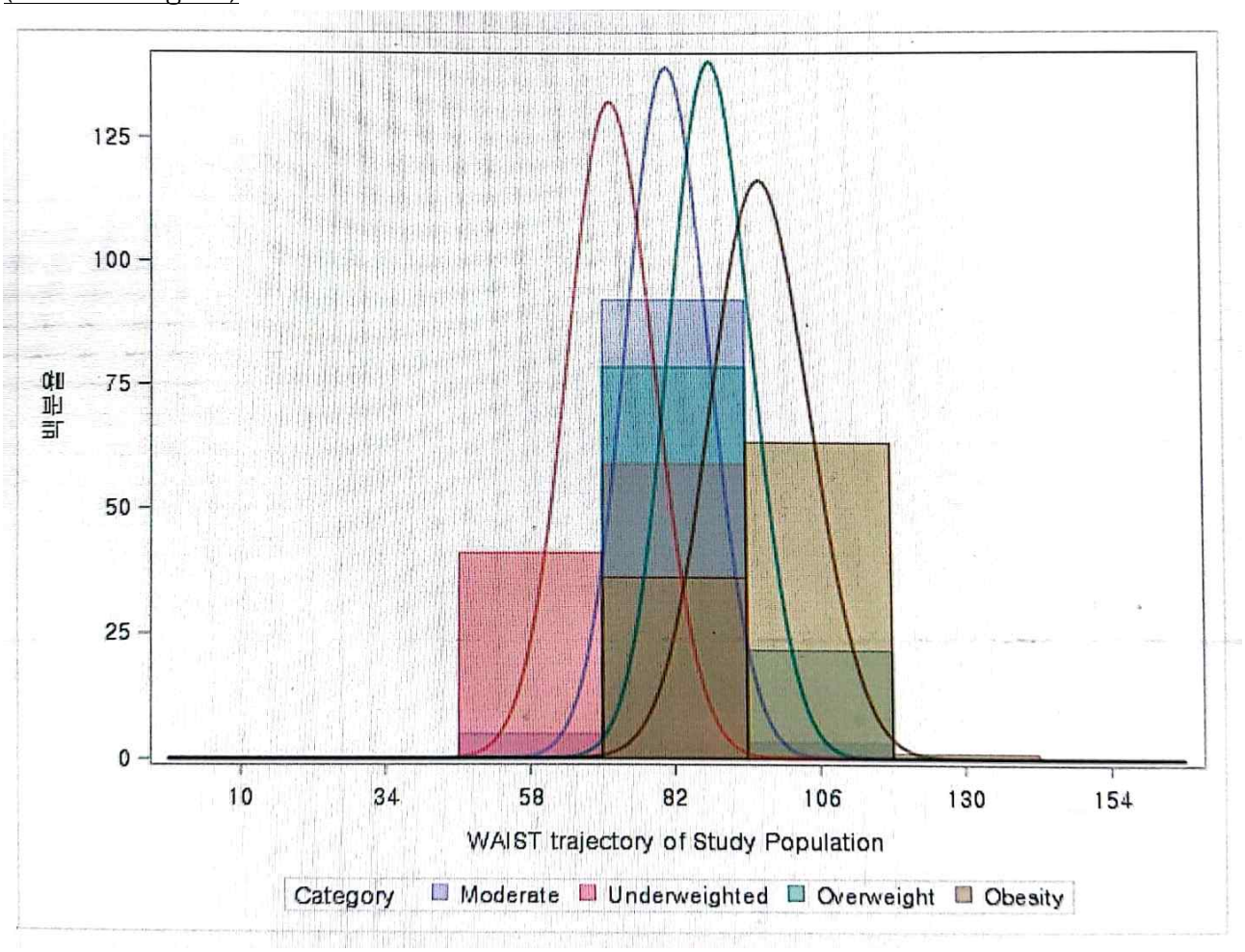
(Scatterplot)



: 기준선은 BMI=30를 의미함.

2. WAIST Trajectory

(WAIST Histogram)



2-①. “Underweighted” group

(Summary 값)

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

2-②. “Moderate” group

(Summary 값)

관측치 수	Mean	SD	Min	Q1	Q2	Q3	Max
447,677	81.0181	6.866	30	76	81	86	142

2-③. “Overweighted” group

(Summary 값)

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

2-④. “Obesity” group

(Summary 값)

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

<My Comment>

1. BMI

- 대부분 객체가 처음 분류된(One time BMI 값 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

- BMI처럼 Category = “Moderate”인 group을 집중해서 볼 필요가 있어 보인다. “Moderate” group의 WAIST 최댓값이 가장 크며, “Moderate” group의 WAIST 최솟값이 “Underweighted” group WAIST 최솟값보다 적다.

- Category = “Overweighted” group의 WAIST 최솟값이 “0”인데 이는 측정오류로 보인다. 0 다음으로 작은 값이 38이므로 “38”로 보는 것이 더 적절해보인다.