20230620_Data_EDA_결과(공통부분)

담당자	은경 이
상태	완료
마감일	@2023년 6월 20일

What TO DO)

: 공통적으로 처리가 필요한 데이터 정제 과정 시행 / 제외조건, filtering 적용함에 따라 줄어드는 N수 파악

Share & Result

1. 사용하지 않는 변수 제거 + 생년월일 값 filtering

```
/* Raw Data EDA */

LIBNAME dir 'E:\koshri-23\data';

/* 출생연도가 1940 ~ 1999년인 사람들의 데이터 필터링 */
/* DROP the unneeded variable */
DATA cohort;
SET dir.Db3 (DROP = AGE DURATION DREGDATE REAL_AUTH_CODE2 DTH_POS MRR_STATUS DTH_EDU
                    DTH_AGE NATIONALITY_CLASS_F NATIONALITY_F CODE_103 CODE_56
                    DTH_JOB NEW_DTH_CODE1 NEW_DTH_CODE2);
WHERE (input(BYEAR, 4.) BETWEEN 1940 AND 1999) and (input(BMONTH,2.) between 1 and 12)
      and (input(BDAY, 2.) between 1 and 31);
run;

/* Check N */
proc sql;
create table check_N as select distinct INDI_ID from dir.Db3;
quit;

proc sql;
create table change_N as select distinct INDI_ID from cohort;
quit;
```

: 기본 Raw data의 distinct INDI_ID는 총 31,361,258명이었다.

→ 생년월일 값 제한(생년이 1940년 ~ 1999년 사이 + 월은 1월 ~ 12월 + 일은 1일 ~ 31일)한 후

파악한 distinct INDI_ID는 **총 30,603,505명**

2. ECNY_DT(취득일) = OUT_DT(상실일) 혹은 ECNY_DT(취득일)이 2019년 이후인 관측치 제거

```
/* ECNY_DT filtering */
DATA cohort2;
SET cohort;
IF (ECNY_DT = OUT_DT) or (ECNY_DT>'20181231') THEN delete;
RUN;
```

: 암 추적 기간이 2018년까지이므로 취득일이 2019년 이후인 관측치 제거 필요

: ECNY_DT = OUT_DT인 것은 기록 작성 상의 오류이므로 제거 필요

→ 관측치가 110330685에서 101177476으로 감소함. 즉, **총 9,153,209개의 관측치가 제거**됨.

3. ENCY_DT, OUT_DT 변수 값 quality check + OUT_DT가 결측인 관측치 값 대체

3-1) “ECNY_DT” 변수와 “OUT_DT” 변수는 기본적으로 YYYYMMDD 형식이어야하므로, 변수가 결측이 아니면서 길이가 8이 아닌 관측치는 제거 필요

```
/* 상실일, 취득일의 입력 방식이 옳지 않은 데이터 check */
proc sql;
create table temp as select * from cohort2
where length(ECNY_DT) <> 8 or (not missing(OUT_DT) and  length(OUT_DT) <> 8);
quit;
```

: 파악 결과, 총 44개의 관측치가 잘못 입력되어 있음을 확인 (이때, 모든 관측치에 대해 취득일 변수는 결측이 아니므로 ECNY_DT 변수의 결측 유무는 조건에 추가하지 않음.)

3-2) “OUT_DT” (상실일) 변수가 결측인 것은 아직까지 근무 중임을 의미하는 것이고, 마지막 추적 기간을 2018년으로 설정하였으므로 **해당 결측값은 ‘20181231’로 대체**

```
/* 상실일, 취득일의 입력 방식이 옳지 않은 데이터 delete + OUT_DT가 결측인 관측치 대체 */
DATA cohort3;
SET cohort2;
IF length(ECNY_DT) ^= 8 or (not missing(OUT_DT) and length(OUT_DT) ^= 8) then delete;
IF OUT_DT = '' then OUT_DT = '20181231';
RUN;
```

4. AGE(연령), DURATION(근속연수) 변수 재생성 + DURATION 변수 Filtering

: “AGE”(연령), “DURATION” (근속연수) 변수 재생성 필요성을 인지 / 두 변수 재생성

→ 생년월일과 취득일(ECNY_DT), 상실일(OUT_DT) 변수 이용

: 근속연수의 경우, 음수가 발생하면 안되기 때문에 DURATION 변수가 음수인 관측치는 추가 제거

```
/* Define "AGE", "DURATION" variable + IF DURATION < 0 delete */
DATA cohort4;
SET cohort3;
AGE = round((input(ECNY_DT, yymmdd8.) - input(cats(BYEAR,BMONTH,BDAY), yymmdd8.))/365.25,
            .01);
DURATION = round((input(OUT_DT, yymmdd8.) - input(ECNY_DT, yymmdd8.))/365.25, .01);
IF DURATION < 0  then delete;
RUN;
```

: Filtering 진행 결과, 관측치 수가 101177432개에서 101177399개로 감소함. 즉, 총 33개의 관측치가 제거됨을 확인.

5. Column 값이 모두 같은 행(중복) 삭제 + 최초 고용보험 등록 이전 암 발생 이력 존재하는 객체 제외

5-1) 모든 Column 값이 같은, 중복 행 제거

(20230619_Data_EDA_Plan(공통처리) paper 참고)

: 이전에 살펴본 바에 따르면, “NO”, “INDI_ID”, “ECNY_DT”, “OUT_DT”, “AGE”, “DURATION” 변수 값이 모두 같은 측정치가 존재함을 확인하였음.

→ 이를 바탕으로, 모든 Column 값이 같은, 중복 행 제거 과정 수행

```
/* Duplicate ROW delete */
proc sort data = cohort4 noduprecs;
BY INDI_ID;
run;
```

: 파악 결과, 중복인 행이 존재하지 않았음. (Raw data에서의 AGE, DURATION 변수 값 문제가 원인인 것으로 판단됨.)

5-2) 최초 고용보험 등록일 이전에 암 이력이 존재하는 객체 추출

: 데이터를 살펴보니, 암 이력이 존재하는 사람 대상 고용보험 일자와 상관없이 모든 반복 측정치에 대해 암 DB 변수 값이 동일함을 확인 + fdx1이 가장 과거 암 발병일자임을 나타냄

: 추가로, ECNY_DT, OUT_DT 변수처럼 fdx1 변수 값의 길이가 8이 아닌 경우도 존재함을 확인

EWTABLE: Work.Temp

연번	성별	취득일	상실일	상실사유	신
3118948	남	20040711	2005071	폐업,도산,공사중단	Y
3626498	여	20041022	2004125	계약만료, 공사종료	Y
1298299	남	19980101	2004021	개인사정으로 인한 자진퇴사	Y
2220119	여	20020415	2003121	계약만료, 공사종료	Y
589339	남	19950701	2003071		Y
1341744	남	20020101	2004101	기타개인사정	N
1259667	남	20031001	2003101	개인사정으로 인한 자진퇴사	N
1055350	남	19950701	2005021		Y
69041	남	20030401	2004030	그밖에회사사정에의한퇴직	N
2703949	여	19980301	2003121	그밖에회사사정에의한퇴직	Y
2282795	남	20031104	2004051	기타개인사정	N

→ 이를 바탕으로, first.ECNY_DT와 min(fdx1, … fdx6) 비교 통해 최초 고용보험 등록일 이전에 암 이력이 존재하는 객체만 추출

```
/* 최초고용보험등록일자 전에 암 발병한 사람 추출 */
proc sort data = cohort4;
by INDI_ID ECNY_DT;
run;

DATA cohort5 exclude;
SET cohort4;
BY INDI_ID ECNY_DT;
IF first.INDI_ID;
IF (not missing(fdx1) and length(fdx1) ^= 8) or  (not missing(fdx2) and length(fdx2) ^= 8)
or (not missing(fdx3) and length(fdx3) ^= 8)
or  (not missing(fdx4) and length(fdx4) ^= 8) or (not missing(fdx5) and length(fdx5) ^= 8)
or  (not missing(fdx6) and length(fdx6) ^= 8)  then delete; /* 21개 */

IF not missing(fdx1) and input(ECNY_DT, yymmdd8.) > min(input(fdx1, yymmdd8.),
input(fdx2, yymmdd8.), input(fdx3, yymmdd8.), input(fdx4, yymmdd8.)
, input(fdx5, yymmdd8.), input(fdx6, yymmdd8.)) then output exclude;
ELSE output cohort5;
run;
```

: 결과, **총 183,877개의 관측치가 제거**됨을 확인함.

6. 최종 Cohort에 포함되는 객체들의 자료만 가져와 “New_Raw” data 생성

```
proc sql;
create table dir.new_raw as select a.* from cohort4 as a join cohort5 as b
on a.INDI_ID = b.INDI_ID;
quit;
```

: 모든 제외조건 적용한 “New_Raw” data의 **총 관측치 수는 100681906개**임을 확인함.

→ 이 데이터를 기반으로 공변량 정의 예정