<산업안전보건연구원 - Data EDA Plan>

- 작성자 : 이은경 shared with 윤서
- 6월 17일 Version

Goal)

: 공변량을 정의하기 위해 공통적으로 전처리가 필요한 과정 계획과 코드 공유
/ 겹치는 공변량 정의 역할 분배

파악한 부분 & Plan)

0. 각자 정의한 공변량 tbl joint 위해 KEEP해 두어야 하는 변수

: "NO"(연번), "INDI_ID"(개인 일련 번호) -- 공통 변수

1. Covariate 목록

(윤서)

(취득일, 상실일)	최초 고용보험 등록 일자	근속 연수	업종정보 (대분류, 중분류, 소분류)	직종 코드	직종 차수	상시 근로자 수	사업장 주소 정보

↳ "근속 연수"(DURATION) 변수는 새로 생성 필요 — 윤서 comment

(은경)

생년월일	최초 고용보험 일자 + entry (최초 등록 연도 범주화 변수)	성별	AGE (입사 당시 나이)	근속 연수	(취득일, 상실일)	암 이력 + 진단 일자	사망 여부 + 사망 일자	Outcome (폐암 / 백혈병 진단 여부 + 시점)

↳ "entry" 변수는 Dropbox-meeting-0614 회의 자료에 나타나 있던 변수이며, 추가 생성이 필요함을 교수님께 확인받음 — 은경 comment

↳ 만약, "OUT_DT"(상실일)이 빈 칸이면 '2018/12/31'로 모두 변경하는 것으로 확인받음 — 은경 comment

1-1) 겹치는 공변량

변수명	담당자
(취득일, 상실일)	은경, 윤서 둘 다 가지고 가는 방향
최초 고용보험 일자 + entry(최초 등록 연도 범주화 변수)	은경
근속 연수(DURATION)	Data 정제 때 생성하는 방향

↳ (취득일, 상실일) pair의 경우 time-varying covariate Cox-ph model 적합 때 반드시 필요한 변수이므로 ((start, stop) 형태) 각자 가지고 가는게 낫다고 판단.

↳ "NEW_Y=Y"가 최초 고용보험 등록을 나타내는 것이 아님 + 최초 고용보험 일자 정의 방식 confirm 받은 은경이 정의하는 것이 더 효율적이라 판단.

↳ DURATION 정의가 (상실일-취득일)/365.25와 맞지 않으므로 기본적인 Data 정제 때 새로 생성하는 방향으로 가는 것이 효율적이라 판단. (DURATION 이용해 데이터 제외 과정 필요)

2. Raw Data 전처리

① 데이터 수치형 혹은 날짜형 변환

: “BYEAR”, “BMONTH”, “BDAY”, “ECNY_DT”, “OUT_DT”, “AGE”, “DURATION”, “BIZ_NO”(상시근로자수)

↳ 연도 이용한 데이터 추출, 분리, 근속연수, 입사 시 연령 계산이 용이

② INDI_ID가 ‘1000000000001’인 객체(4명)에 대한 관측치 전부 제외

: INDI_ID가 ‘1000000000001’은 주민 번호 누락을 의미하므로 제외 필요

③ 최초 고용보험 등록 이전에 암 발생 혹은 사망한 객체에 대한 관측치 전부 제외

: 노동으로 인한 암 발병 risk 추정에 편향을 최대한 줄이기 위함.

Question) 최초 고용보험 등록 이전에 사망한 객체가 존재 가능한가?

④ 출생연도가 1940년 ~ 1999년 사이, 출생월이 1~12월 사이, 출생일이 1~31일 사이로 제한

: 월 / 일 중 가질 수 없는 값을 가지는 객체 존재 + 충분한 반복 관측치 확보 위해 출생연도를 1940~1999년 사이로 제한

⑤ AGE(입사 당시 연령)이 음수 또는 결측 또는 100이상인 객체에 대한 관측치 전부 제외

: 연령이 음수 혹은 결측인 관측치는 존재할 수 없음 + 입사 당시 연령이 100세 이상인 것은 상식에 부합함

⑥ DURATION(근속 연수) 변수가 음수인 값을 가지는 객체에 대한 관측치 전부 제외

: 근속 연수가 음수인 관측치는 존재할 수 없음

⑦ 취득일(ECNY_DT) > 상실일(OUT_DT) 인 경우나, 취득일(ECNY_DT)가 2019년 이후인 관측치는 삭제

: “취득일 > 상실일”은 상식에 부합함 + 충분한 추적 기간 확보하기 위해 취득일을 2019년 이전까지로 제한함

[구상한 code]

```
LIBNAME dir 'E:\koshri-23\data';
LIBNAME mine 'E:\koshri-23\eklee\data';

DATA dir.Num_Db3;
SET dir.Db3;
BYEAR2=input(BYEAR, 12.);
BMONTH2=input(BMONTH, 12.);
BDAY2=input(BDAY, 12.);
ECNY_DT_2=mdy(substr(ECNY_DT,5,2), substr(ECNY_DT,7,2), substr(ECNY_DT,1,4));
format ECNY_DT_2 yymmdd10.;
ECNY_DT2 = input(ECNY_DT_2, yymmdd8.);
OUT_DT_2=mdy(substr(OUT_DT,5,2), substr(OUT_DT,7,2), substr(OUT_DT,1,4));
format OUT_DT_2 yymmdd10.;
OUT_DT2 = input(OUT_DT_2, yymmdd8.);
AGE2=input(AGE, 12.);
/* DURATION 변수 새로 정의 */
DURATION2=(OUT_DT2 - ECNY_DT2)/365.25;
DROP BYEAR BMONTH BDAY ECNY_DT OUT_DT AGE DURATION BIRTH_1 ECNY_DT_2 OUT_DT_2;
RENAME BYEAR2=BYEAR BMONTH2=BMONTH BDAY2=BDAY ECNY_DT2=ECNY_DT OUT_DT2=OUT_DT AGE2=AGE
       DURATION2=DURATION;
RUN;
```

```
/* ②, ④, ⑤, ⑥ - Extract INDI_ID which need to be excluded */
PROC SQL;
create exclude_ID as select distinct INDI_ID from dir.Num_Db3
where (INDI_ID = '1000000000001') or (BYEAR not between 1940 and 1999) or
(BMONTH not between 1 and 12) or (BDAY not between 1 and 31) or
(AGE not between 1 and 99) or (DURATION < 0) or (DURATION < >' ');
quit;

/* ③ */
DATA temp;
SET dir.Num_Db3(KEEP = INDI_ID NO ECNY_DT fdx1-fdx6);
RUN;

PROC SORT DATA = temp;
BY INDI_ID ECNY_DT;
RUN;

DATA temp2;
SET temp;
BY INDI_ID ECNY_DT;
IF first.INDI_ID; /* 최초 고용보험 등록 일자 추출 */
IF ECNY_DT > fdx1 or ECNY_DR > fdx2 or ECNY_DT > fdx3 or ECNY_DT > fdx4 or ECNY_DT > fdx5
   ECNY_DT > fdx6;
KEEP INDI_ID;
RUN;

/* Exclude ID Merge */
DATA Exclude;
SET exclude_ID temp2;
RUN;

/* Extract Included ID */
PROC SQL;
create table include_ID as select distinct INDI_ID from dir.Num_Db3
except select * from Exclude;
quit;

/* Delete Obs which need to be excluded and ⑦ */
PROC SQL;
create table dir.Raw_temp as select a.* from dir.Num_Db3 as a join dir.include_ID as b
on a.INDI_ID = b.INDI_ID and a.ECNY_DT < > a.OUT_DT and a.ECNY_DT < a.OUT_DT and
a.ECNY_DT < 2019/01/01;
quit;
```

↳ 최종 Study population에 포함되는 distinct INDI_ID N수 count 필요

2-1) 전처리한 Data split

: 데이터는 기본적으로 “고용보험DB” + “암 DB” + “사망 DB”로 구성되어 있고 용량이 매우 크므로 각 DB로 split 한 뒤 공변량 정의하는 것이 메모리 상 효율적인 방안으로 판단됨. / 각 DB를 joint할 수 있는 “NO”, “INDI_ID”는 기본적으로 유지

DB	table명	구성 변수
고용보험 DB	“Company_raw” tbl	입사 시 연령(AGE) + 근속연수(Duration) + 취득일(ECNY_DT) + 상실일(OUT_DT) + 상실 사유(OUT_CZ) + 근속 년수(DURATION) + 신규 취득 여부(NEW_NY) + 10차 업종 코드(BIZ_INDUTY10) + 직종 코드(JSSFC_CD) + 직종 차수(JSSFC_NO) + 사업장명(BIZ_NM) + 상시 근로자수(BIZ_NO) + 사업장 주소(BIZ_ADDRESS) + 사업장 우편번호(BIZ_ZIP) + INDI_ID + NO
사망 DB	“Death_raw” tbl	사망연령(DTH_AGE) + 사망연월(DTH_DATE1~DTH_DATE3) + NO + INDI_ID
암 DB	“Cancer_raw” tbl	진단 일자(fdx1~fdx6) + 진단코드(icd10_1~icd10_6) + NO + INDI_ID
개인정보 DB	“Demographic_raw” tbl	INDI_ID + BIRTH_YMD + SEX