‹2023 산업보건연구원 Data Check›

- 3월 7일 ~ 9일 Version

‹What TO DO›

: Data Column 명과 변수명 Label Excel 표에 따로 정리

- 채우지 못한 부분은 파악할 수 없었던 부분(후에 기록 필요)

: Data 설명 텍스트 파일 읽고 궁금한 점 파악, 질문 정리

‹TRY TO DO›

1) SAS 상에서 N수가 1억 개인 데이터를 이용해 FAST option 추가한 Coxph model 적합 하려 하였음.

: Data의 Dimension이 커지면 Data 저장에서부터 메모리 부족 에러가 발생함.

: Data의 Column 개수를 10개 정도로 하면 Data 저장은 되나, "FAST" option을 추가한 Coxph도 메모리 문제로 인해 적합이 되지 않음.

‹확인사항›

Check1) 취득일과 상실일이 같은 관측치가 존재하는가?

: 총 15,602개의 관측치가 "취득일" = "상실일" 조건을 만족한다.

: 해당 조건을 만족하는 관측치에 대응하는 Distinct INDI_ID는 15,506명

(이는 이중으로 고용보험이 가입되거나, 직종이 여러 개인 경우를 뜻한다는 설명 보고 파악해봄.)

Check2) 취득일(ECNY_DT)가 2019.01.01. 이후인 관측치가 몇 개인가?

: 총 7,210,228개

(Data 설명 자료에서 2019년부터의 입사자는 버려야 한다는 내용 보고 파악해본 결과임.)

Check3) INDI_ID가 "1000000000001"인 관측치(주민번호 누락 의미)가 존재하는가, 존재한다면 몇 개?

: 4개의 관측치가 존재함.

Check4) INDI_ID가 결측인 관측치는 몇 개?

: "INDI_ID"가 결측인 관측치는 없음을 확인.

Check5) "DTH_POS"(사망 장소) 변수 값을 어떻게 표현하고 있는가?

	사망장소
1	1
2	10
3	2
4	3
5	4
6	5
7	6
8	7
9	8
10	9
11	99

: 해당 변수 값은 어떻게 해석해야 하는가?

Check6) “NATIONALITY_CLASS_F”(국적 구분) 변수 값 어떻게 표현하고 있는가?

	국적구분
1	1
2	2
3	N

: 해당 변수 값은 어떻게 해석해야 하는가?

〈질문사항〉

Question1) NO(연번)이 의미하는 것은 무엇인가?

--- “INDI_ID”가 같은 사람임에도 NO(연번)이 다른 관측치가 여럿 존재함을 확인.
한 객체 별 서로 다른 직장 가입자임을 나타내는 것인가?

Question2)

→ 해당 설명 부분 이해하지 못함.

2005년부터는 고용직업분류가 만들어짐. 2차.
직업코드에는 고용직업분류(고용노동부 개발, 기술직/사무직처럼 직형태로 분류),
vs 표준직업분류(통계청 개발, 숙련자/고숙련자 느낌)
4 digits까지는 둘이 호환됨

Question3) “MRR_STATUS”는 혼인상태를 나타내는 변수인데, 이 변수값이 결측인 것은 어떻게 해석하면 되는가? / 그리고, 결측이 아닐 때 “MRR_STATUS” 변수가 가지는 unique한 값은
‘1’, ‘2’, ‘3’, ‘4’, ‘9’이다. 값이 의미하는 것은 무엇인가?

	혼인상태
1	1
2	2
3	3
4	4
5	9

Question4) “DTH_EDU”(교육 정도) 변수, “NATIONALITY_CLASS_F”(국적 구분) 변수 값이 결측인 경우는 어떻게 해석해야 하는지?