산보연 Data forecasting 07월 29일 Version - Lung cancer Data

작성자 : 이은경

<What to do>

: 바뀐 Data level 별로 반응변수는 "질병 발생 건수", 설명변수는 "YEAR"로 하여 3가지 모형 적합(1. 단순선형회귀모형 / 2. 1차 spline model / 3. quadratic regression)

: train data와 validation data는 YEAR 기준으로 split / train data는 2000년 ~ 2014년, validation data는 2015년 ~ 2018년 자료로 지정

: train data로 3가지 모형 적합 후, validation data의 질병 발생 건수 예측

: validation data의 연도에 따른 실제 질병 발생 건수 알고 있으므로 예측값과 비교

: Performance criteria는 MAPE로 사용 $(\frac{1}{4}\sum_{t=2015}^{2018}|y_{it}-\widehat{y_{it}}|/y_{it})\times 100$ (이때, i는 각 사업장, t는 연도 의미) / 후에 진행될 model ensemble 위해 RMSE도 계산

: Data level 별로 적합한 model들의 performance를 시각적으로 표현하기 위해 plots 생성

: 추적 인년 합계가 매우 적어 추세가 불안정한 사업장 "T", "U"는 제외

<Result>

1) Lung cancer data

① Level 1 data

①-1. "YEAR" 변수로만 grouping한 data : lv1_lung_total

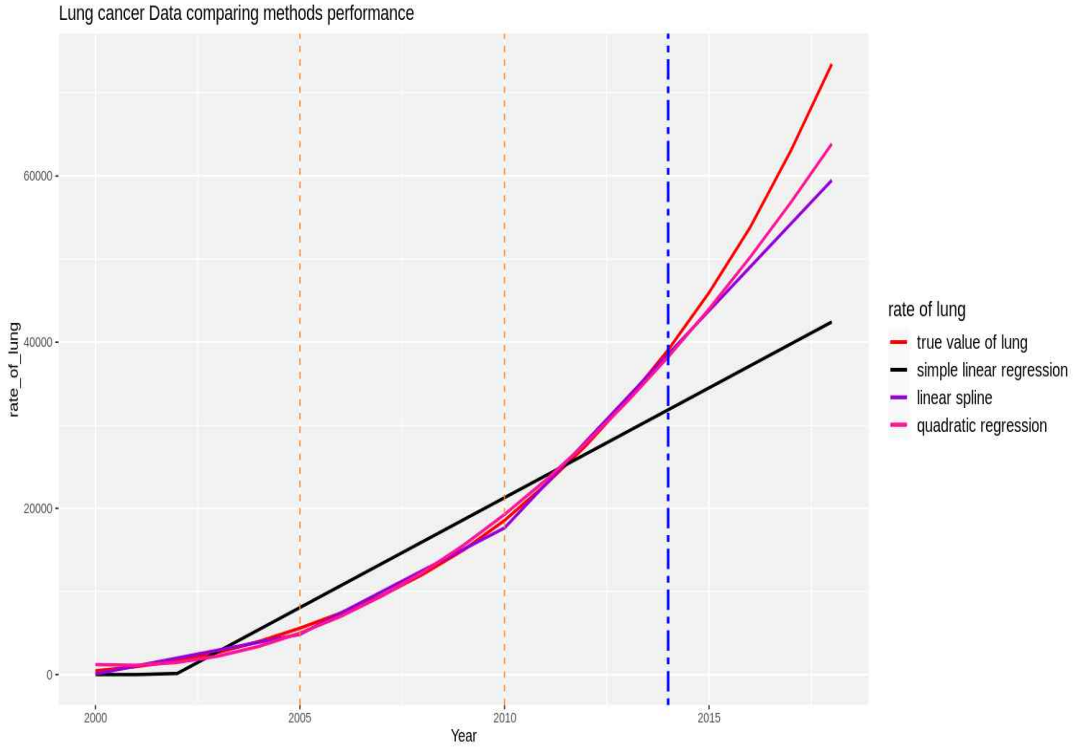
: 해당 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 계산한 MAPE 값은 아래 표와 같다.

NAME	단순선형회귀	1차 spline	Polynomial 회귀
X	33.74	11.63	8.46

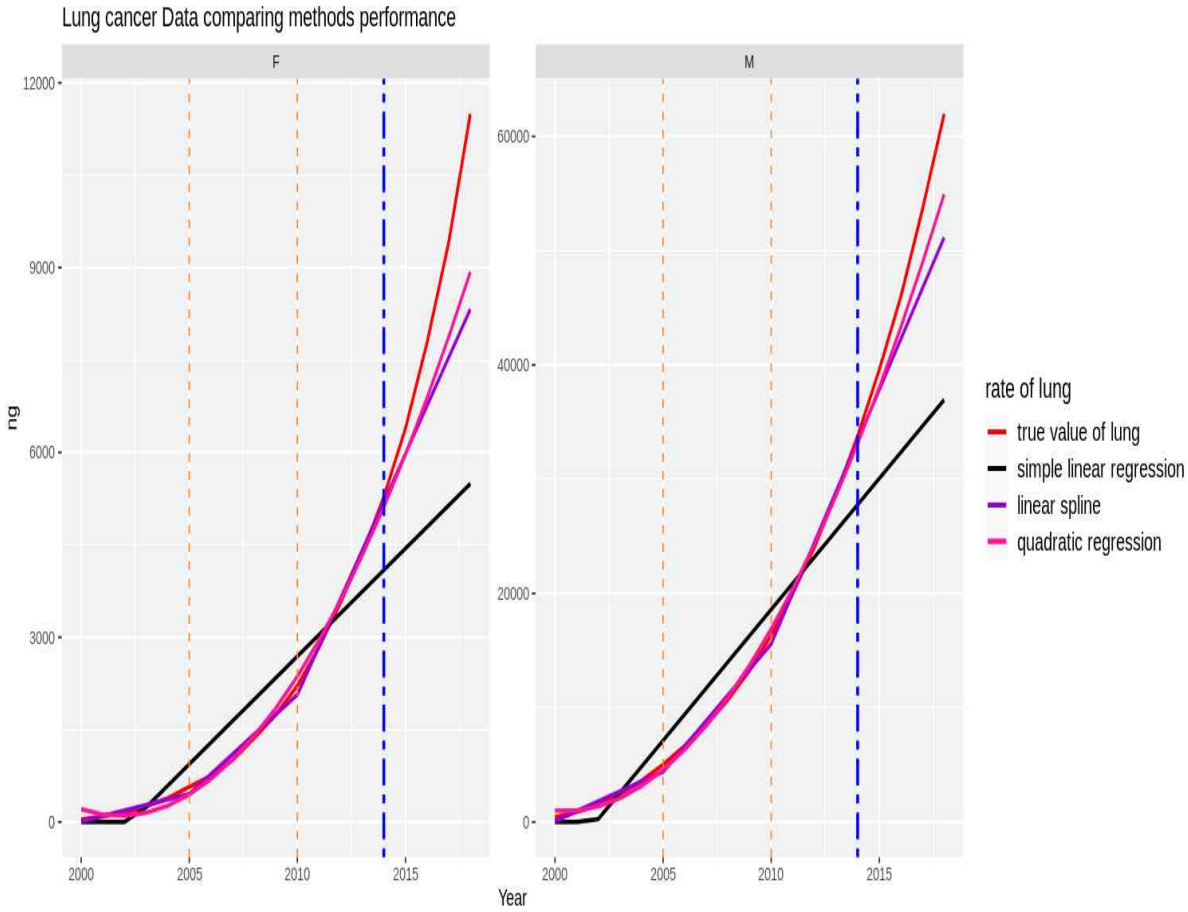
: $YEAR^2$이 설명변수로 추가된 다중선형회귀모형의 MAPE 값이 가장 작은 것을 확인할 수 있다.

: 반응변수가 "누적 발생 건수"인 경우에는 연도에 따른 변화 추세가 지수함수 혹은 이차함수 형태를 보인다.

: 그래프의 파란색 실선은 YEAR=2014년을 의미한다.



①-2. “YEAR”, “SEX” 변수로 grouping한 데이터 : lv1_lung_SEX



: 그래프의 파란색 실선은 YEAR=2014년을 의미한다.

: 해당 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 계산한 MAPE 값은 아래 표와 같다.

NAME	단순선형회귀	1차 spline	Polynomial 회귀
XF	41.63	16.76	14.19
XM	32.36	10.73	7.45

: 성별에 상관없이 $YEAR^2$이 설명변수로 추가된 다중선형회귀모형의 MAPE 값이 가장 작은 것을 확인할 수 있다.

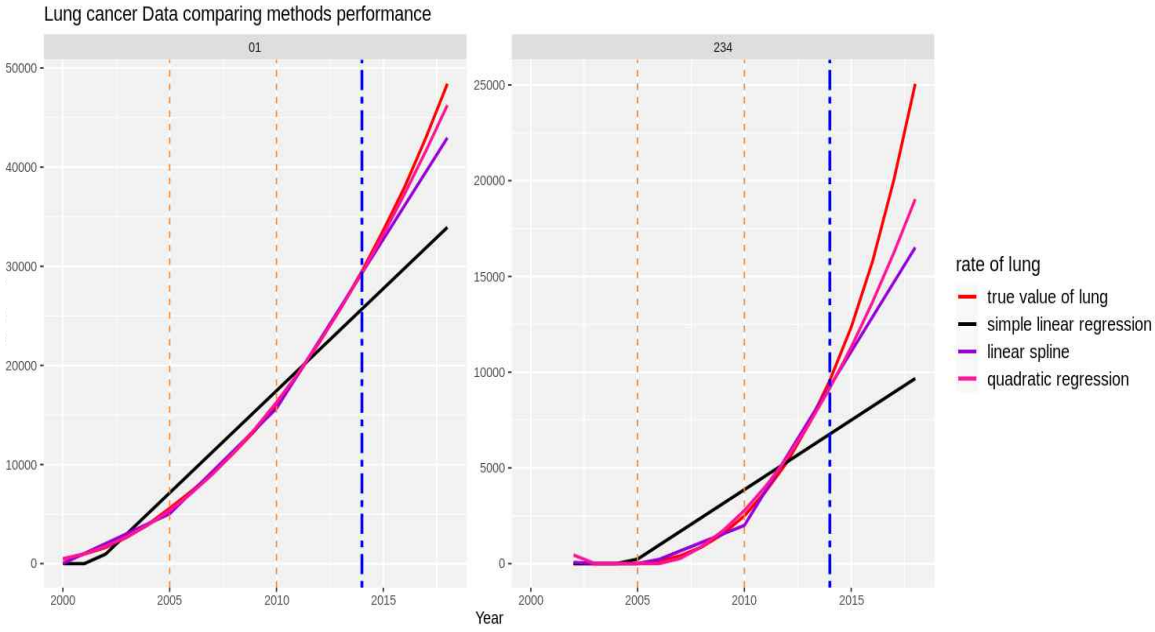
: 그래프의 y축을 비교해보면, 같은 연도일 때, 남성의 누적 질병 발생 건수가 여성의 누적 질병 발생 건수보다 많다는 것을 알 수 있다.

: 실제 누적 발생 건수 변화 추세와 모형이 예측한 추세를 비교해보면, 1차 spline function과 polynomial regression 모형은 train set에서 실제 추세와 거의 가깝게 적합이 되었다.(단순선형회귀모형의 경우, 지수적으로 증가하는 추세를 학습하지 못한 것으로 파악됨) 이때, polynomial regression 모형이 2015년 이후에도 지수적으로 증가하는 추세를 더 잘 예측하는 것으로 추측 된다.

: 같은 항목에 대해 leukmia data와 비교를 해보면, 폐암의 누적 발생 건수가 더 많다는 점을 확인할 수 있다.

---------- 다음 페이지로 ----------

①-3. “YEAR”, “CAL2” 변수로 grouping한 데이터 : lv1_lung_CAL



: **그래프의 파란색 실선은 YEAR=2014**년을 의미한다.

: 해당 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 계산한 MAPE 값은 아래 표와 같다.

NAME	단순선형회귀	1차 spline	Polynomial 회귀
X01	23.71	6.73	2.66
X234	51	22.36	16.22

: 입사 시기에 상관없이 $YEAR^2$이 설명변수로 추가된 다중선형회귀모형의 MAPE 값이 가장 작은 것을 확인할 수 있다.

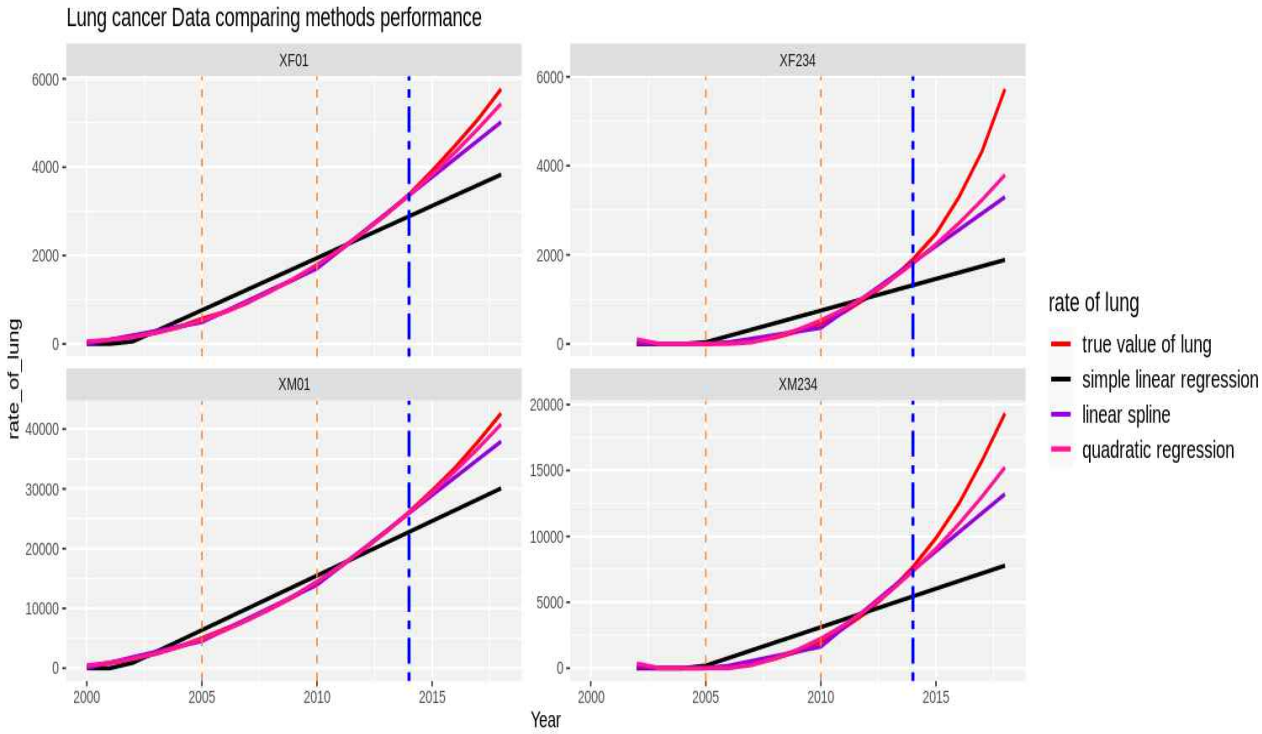
: 그래프의 y축을 비교해보면, 같은 연도일 때, 입사 시기가 2000년 이전인 근로자 집단의 누적 질병 발생 건수가 입사 시기가 2000년 이후인 근로자 집단의 누적 질병 발생 건수보다 많다는 것을 알 수 있다.

: 입사 시기가 2000년 이전인 근로자 집단의 경우, Polynomial regression 모형의 성능이 매우 좋은 것으로 보인다. 반면, 입사 시기가 2000년 이후인 근로자 집단은 3가지 모형의 성능이 모두 그다지 좋지 않은 것으로 파악된다. 실제 추세 변화를 보면, 2000년에서 2010년까지는 발생 건수가 없었지만, 이후로 급격히 증가하는 것으로 보이는데, 이 부분을 반영할 수 있는 또 다른 모형 고려가 필요하다고 생각한다.

: 같은 항목에 대해 lung data와 비교를 해보면, 누적 질병 발생 건수도 더 많고, 2000년 이후 입사한 근로자 집단에 대해 백혈병의 누적 발생 건수는 2005년을 기점으로 급격히 증가하는 반면, 폐암의 누적 발생 건수는 2010년을 기점으로 급격히 증가한다.

---------- 다음 페이지로 ----------

①-4. “YEAR”, “SEX”, “CAL2” 변수로 grouping한 데이터 : lv1_lung_SEX_CAL2



: **그래프의 파란색 실선은 YEAR=2014**년을 의미한다.

: 해당 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 계산한 MAPE 값은 아래 표와 같다.

NAME	단순선형회귀	1차 spline	Polynomial 회귀
XF01	27.03	8.12	3.83
XF234	54.74	26.97	21.53
XM01	23.27	6.54	2.5
XM234	49.96	21.07	14.72

: **파란색으로 강조한 부분**은 각 집단에서 가지는 MAPE 최솟값을 뜻한다.

: 그래프와 MAPE 파악 결과를 살펴보면, 여성이면서 입사 시기가 2000년 이후인 근로자 집단의 누적 질병 누적 발생 건수 추세 파악이 현 모형들로는 어려워 보인다.
---------- Leukemia data의 경우에서도 공통적으로 나온 결론

---------- Level 1 Data 파악 결과 정리 ----------

1. 폐암 발생 건수는 지수적으로 증가하는 형태를 보인다.
2. 입사 시기가 2000년 이후인 근로자 집단의 경우, 2000년부터 2010년 사이에는 질병 발생 이벤트가 없었으나, 이후부터 지수적으로 급격히 증가함을 확인하였다.
3. 폐암 누적 발생 건수 추세의 차이는 성별 보다는 입사 시기에 의한 차이가 더 두드러져 보인다.
4. 반응변수가 “누적 통합 질병 발생률”이 아닌 “질병 누적 발생 건 수”이다 보니, 시간이 지남에 따라 추세가 감소하는 부분이 전혀 없어 회귀모형이 이전보다 좋은 성능을 보이는 것을 확인하였다. - 모든 level data에서 보여 지는 공통사항
5. 백혈병의 누적 발생 건수가 더 적음에도 불구하고, 반응변수가 질병 통합 누적 발생률일때와 다르게 폐암의 누적 발생 건수를 예측하는 것이 더 어렵다.

---

② Level 2 Data

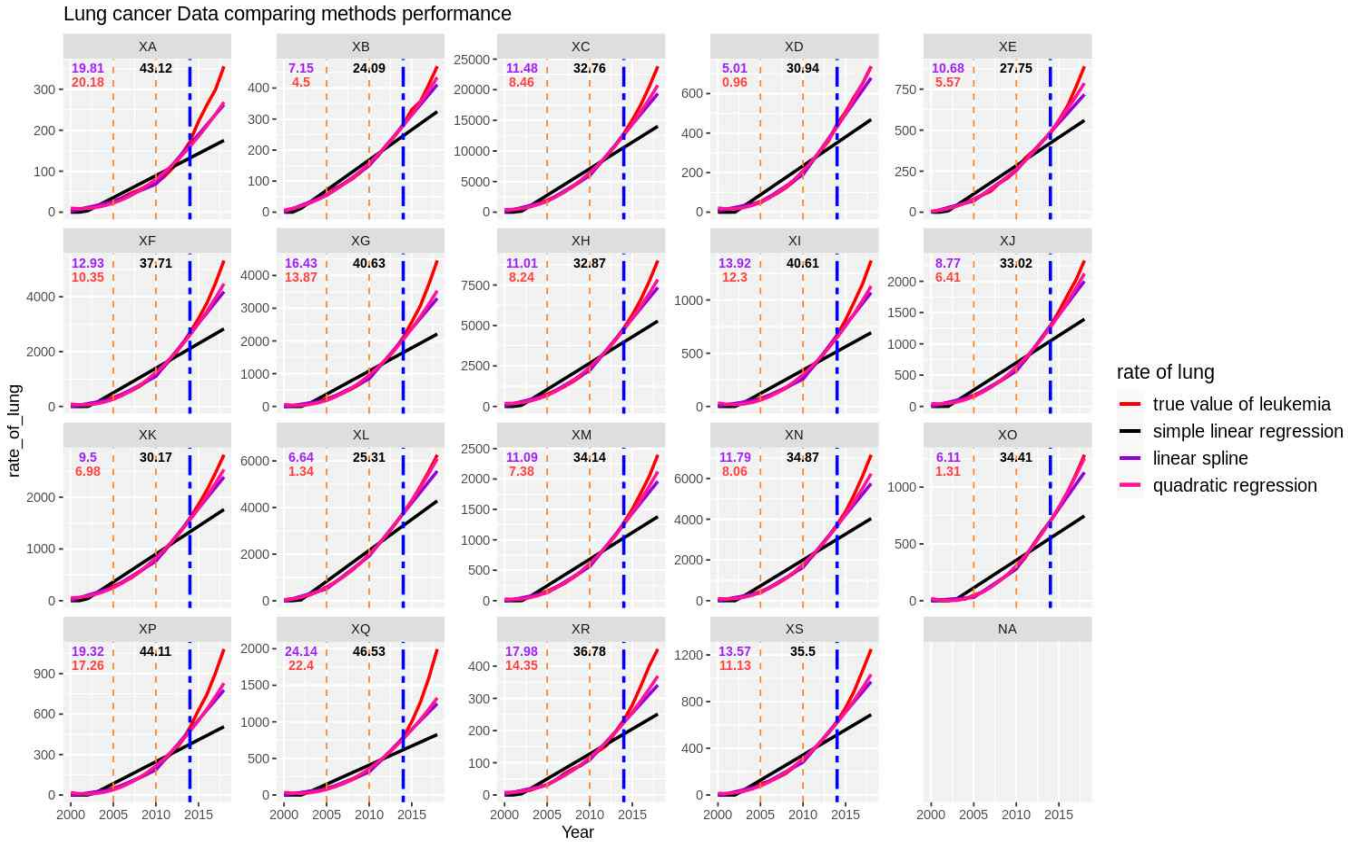
②-1. “UP1”, “YEAR” 변수로 grouping한 데이터 : lv2_lung_total

: “UP1”, “YEAR”로 grouping한 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 계산한 MAPE 값은 아래 표와 같다.

NAME	단순선형회귀	1차 spline	Polynomial 회귀
XA	43.12	19.81	20.18
XB	24.09	7.15	4.5
XC	32.76	11.48	8.46
XD	30.94	5.01	0.96
XE	27.75	10.68	5.57
XF	37.71	12.93	10.35
XG	40.63	16.43	13.87
XH	32.87	11.01	8.24
XI	40.61	13.92	12.3
XJ	33.02	8.77	6.41
XK	30.17	9.5	6.98
XL	25.31	6.64	1.34
XM	34.14	11.09	7.38
XN	34.87	11.79	8.06
XO	34.41	6.11	1.31
XP	44.11	19.32	17.26
XQ	46.53	24.14	22.4
XR	36.78	17.98	14.35
XS	35.5	13.57	11.13

: 각 NAME 별로 MAPE 최솟값을 **파란색으로 강조**하였다.

: 대부분 Polynomial 회귀모형이 좋은 성능을 보이는 것을 확인할 수 있다.→ 이를 통해 각 사업장의 질병 누적 발생 건수가 선형으로 증가하지 않는다는 점을 추측할 수 있다.



: **그래프의 파란색 실선은 YEAR=2014년**을 의미한다.

: 그래프 안, **주황색 글씨는 Polynomial 회귀 적합 통해 얻은 MAPE 값**, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

: 반응변수가 “누적 질병 발생 건수”일 때 대분류 기준, 모든 사업장이 불안정한 추세를 보이지 않는다는 점이 눈에 띈다. 또한, 사업장 “A”(농업, 임업, 어업), 사업장 “P”(교육 서비스업), 사업장 “Q”(보건업 및 사회복지 서비스업)은 2014년 이후 누적 질병 발생 건수가 급격히 증가한다는 사실도 주목할만하다.

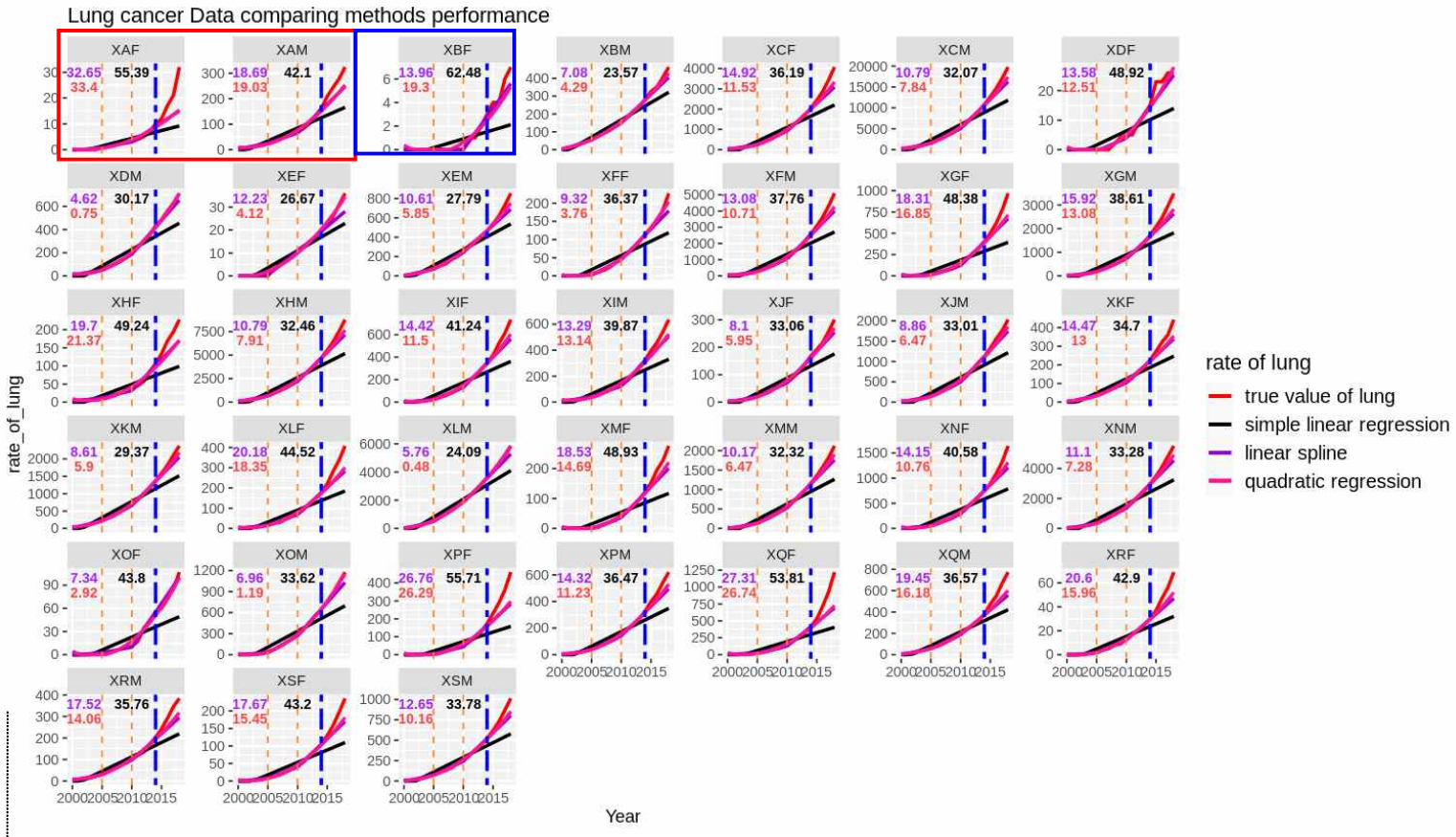
②-2. “UP1”, “SEX”, “YEAR” 변수로 grouping한 데이터 : lv2_lung_SEX

: “UP1”, “SEX”, “YEAR”로 grouping한 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 계산한 MAPE 값은 아래 표와 같다.

: 각 사업장별로 모형 별 MAPE를 전부 기록하지 않고, 최솟값만 표에 제시한다. (그림 통해 모든 모형 적합 통해 얻은 MAPE 확인 가능)

NAME	Female	Male
XA	32.65	18.69
XB	13.96	4.29
XC	11.53	7.84
XD	0.75	4.12
XE	4.12	5.85
XF	3.76	10.71
XG	16.85	13.08
XH	19.7	7.91
XI	11.5	13.14
XJ	5.95	6.47
XK	13	5.9
XL	18.35	0.48
XM	14.69	6.47
XN	10.76	7.28
XO	2.92	1.19
XP	26.29	11.23
XQ	26.74	16.18
XR	15.96	14.06
XS	15.45	10.16

: 같은 사업장에서 대부분 남성 근로자 집단을 대상으로 모형을 적합했을 때 더 좋은 성능을 보이는 것을 확인할 수 있다. (추적 인년 합계가 더 크기 때문이라고 추측한다.)



: **그래프의 파란색 실선은 YEAR=2014**년을 의미한다.

: 그래프 안, **주황색 글씨는 Polynomial 회귀 적합 통해 얻은 MAPE 값**, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

: **빨간색으로 강조한 부분**은 사업장 내 모든 성별 집단에 대해 1차 spline function이 제일 좋은 성능을 보이는 사업장을 표시한 것이다. 해당 사업장은 “A”(농업)이다. ---- 백혈병 데이터에서도 동일한 결과 확인 가능.

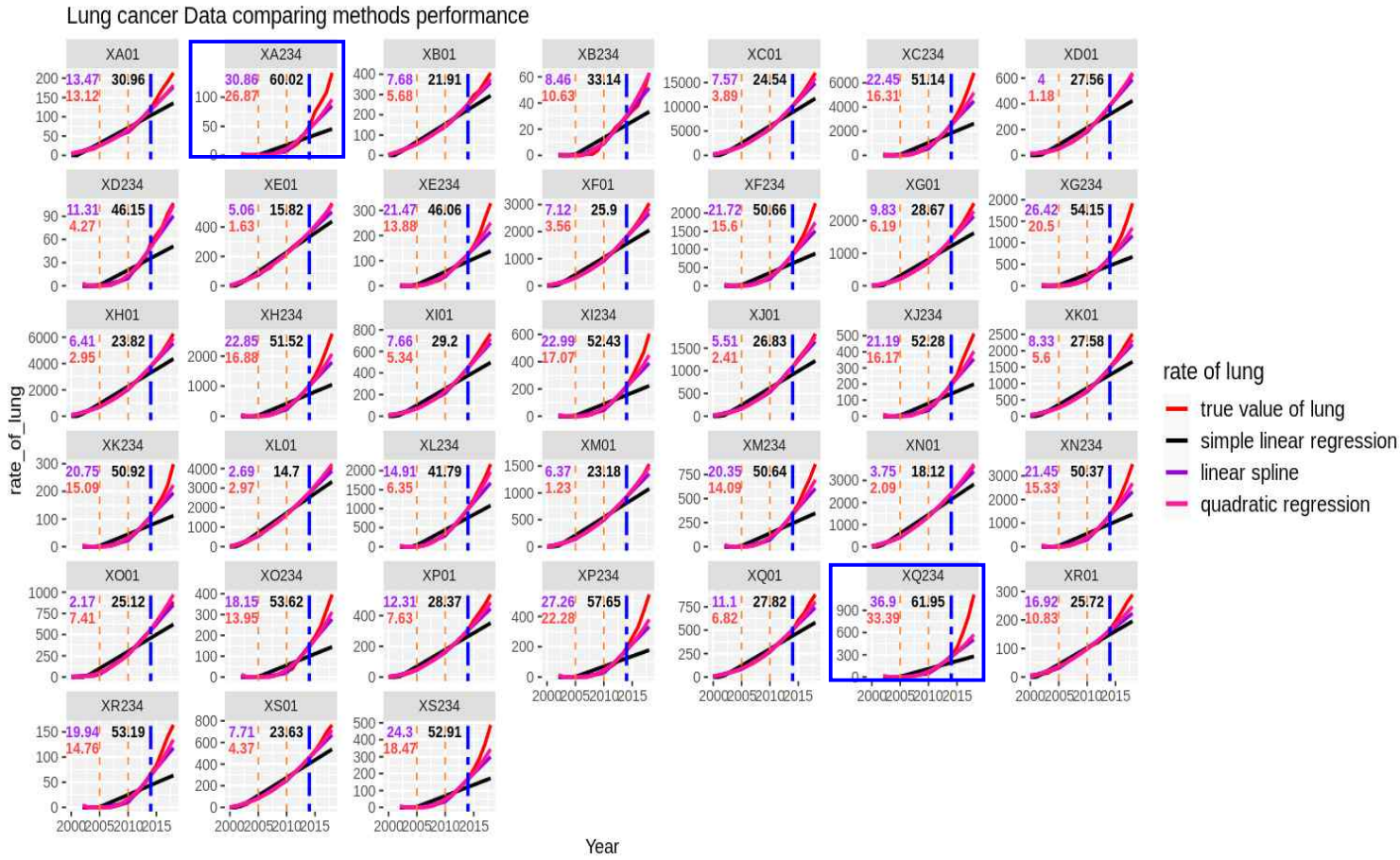
: **파란색으로 강조한 부분**은 연도에 따른 질병 누적 발생 건수의 추세가 계단 함수를 보이는 사업장을 표시한 것이다. 해당 사업장은 “B”(광업)이고, 해당 사업장 내에서는 여성 근로자 집단에 대해서만 특별한 모형을 띈다.

②-3. “UP1”, “CAL2”, “YEAR” 변수로 grouping한 데이터 : lv2_lung_CAL2

: “UP1”, “SEX”, “YEAR”로 grouping한 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 계산한 MAPE 값은 아래 표와 같다.

: 각 사업장별로 모형 별 MAPE를 전부 기록하지 않고, 최솟값만 표에 제시한다. (그림 통해 모든 모형 적합 통해 얻은 MAPE 확인 가능)

NAME	01	234
XA	13.12	26.87
XB	5.68	8.46
XC	3.89	16.31
XD	1.18	4.27
XE	1.63	13.88
XF	3.56	15.6
XG	6.19	20.5
XH	2.95	16.88
XI	5.34	17.07
XJ	2.41	16.17
XK	5.6	15.09
XL	2.97	6.35
XM	1.23	14.09
XN	2.09	15.33
XO	2.17	13.95
XP	7.63	22.28
XQ	6.82	33.39
XR	10.83	14.76
XS	4.37	18.47



: 같은 사업장에서 대부분 입사 시기가 2000년 이전인 집단을 대상으로 모형을 적합했을 때 더 좋은 성능을 보이는 것을 확인할 수 있다.

: **그래프의 파란색 실선은 YEAR=2014년**을 의미한다.

: 그래프 안, **주황색 글씨는 Polynomial 회귀 적합 통해 얻은 MAPE 값**, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

: **파란색으로 강조한 부분**은 연도에 따른 질병 누적 발생 건수의 추세가 2014년 이후 급격한 증가 추세를 보이는 사업장을 표시한 것이다. 해당 사업장은 각각 “A”(농업, 어업, 임업), “Q”(보건업 및 사회복지 서비스업)이고, 해당 사업장 내에서는 모두 입사 시기가 2000년 이후인 근로자 집단에 대해서만 해당 사항을 가진다.

②-4. “UP1”, “SEX”, “CAL2”, “YEAR” 변수로 grouping한 데이터 : lv2_lung_SEX_CAL2

: “UP1”, “SEX”, “CAL2”, “YEAR”로 grouping한 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 계산한 MAPE 값은 아래 표와 같다.

: 각 사업장별로 모형 별 MAPE를 전부 기록하지 않고, 최솟값만 표에 제시한다. (그림 통해 모든 모형 적합 통해 얻은 MAPE 확인 가능)

NAME	F01	F234	M01	M234
XA	38.96	24.93	11.34	27.41
XB	31.65	58.98	5.25	8.46
XC	4.61	21.06	3.76	15.03
XD	12.55	61.28	0.86	1.59
XE	5.14	11.11	1.39	14.25
XF	6.56	3.74	3.9	16.38
XG	0.97	24.16	7.17	18.98
XH	14.79	19.99	2.7	16.74
XI	4.84	14.8	5.79	20.23
XJ	1.91	14.72	2.89	16.54
XK	11.31	13.31	4.71	15.81
XL	3.64	27.78	2.61	4.01
XM	3.8	27.26	2	10.99
XN	1.15	14.76	2.29	15.44
XO	20.54	13.01	3.2	14.42
XP	15.5	26.69	4.62	16.93
XQ	5.21	33.88	7.93	31.76
XR	14.57	10.97	10.39	15.87
XS	5.38	20.48	4.2	17.76

: 백혈병 데이터와 다르게 MAPE 값이 “NaN”을 가지는 사업장은 없다.

: MAPE 값 경향을 보면 대부분 사업장에 대해 남성보다는 여성 근로자 집단이, 입사 시기가 2000년 이전인 집단보다는 2000년 이후인 근로자 집단의 MAPE 값이 크다는 것을 확인할 수 있다.

: **그래프의 파란색 실선은 YEAR=2014**년을 의미한다.

: 그래프 안, **주황색 글씨는 Polynomial 회귀 적합 통해 얻은 MAPE 값**, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

: **파란색으로 강조한 부분**은 연도에 따른 질병 누적 발생 건수의 추세가 계단 함수를 보이는 사업장을 표시한 것이다. 해당 사업장은 “B”(광업)이고, 해당 사업장 내 여성인 근로자 집단에서 현상이 두드러져 보인다.

: **빨간색으로 강조한 부분**은 2014년 이후 질병의 누적 발생 건수가 급격히 증가하는 추세를 보이는 사업장 / 집단을 뜻한다. 해당 사업장은 “D”(전기, 가스, 증기 및 공기 조절 공급업), “L”(부동산업), “M”(전문, 과학 및 기술 서비스업), “Q”(보건업 및 사회복지 서비스업)이며, 세 집단의 공통점은 입사 시기가 모두 2000년 이후라는 점이다.

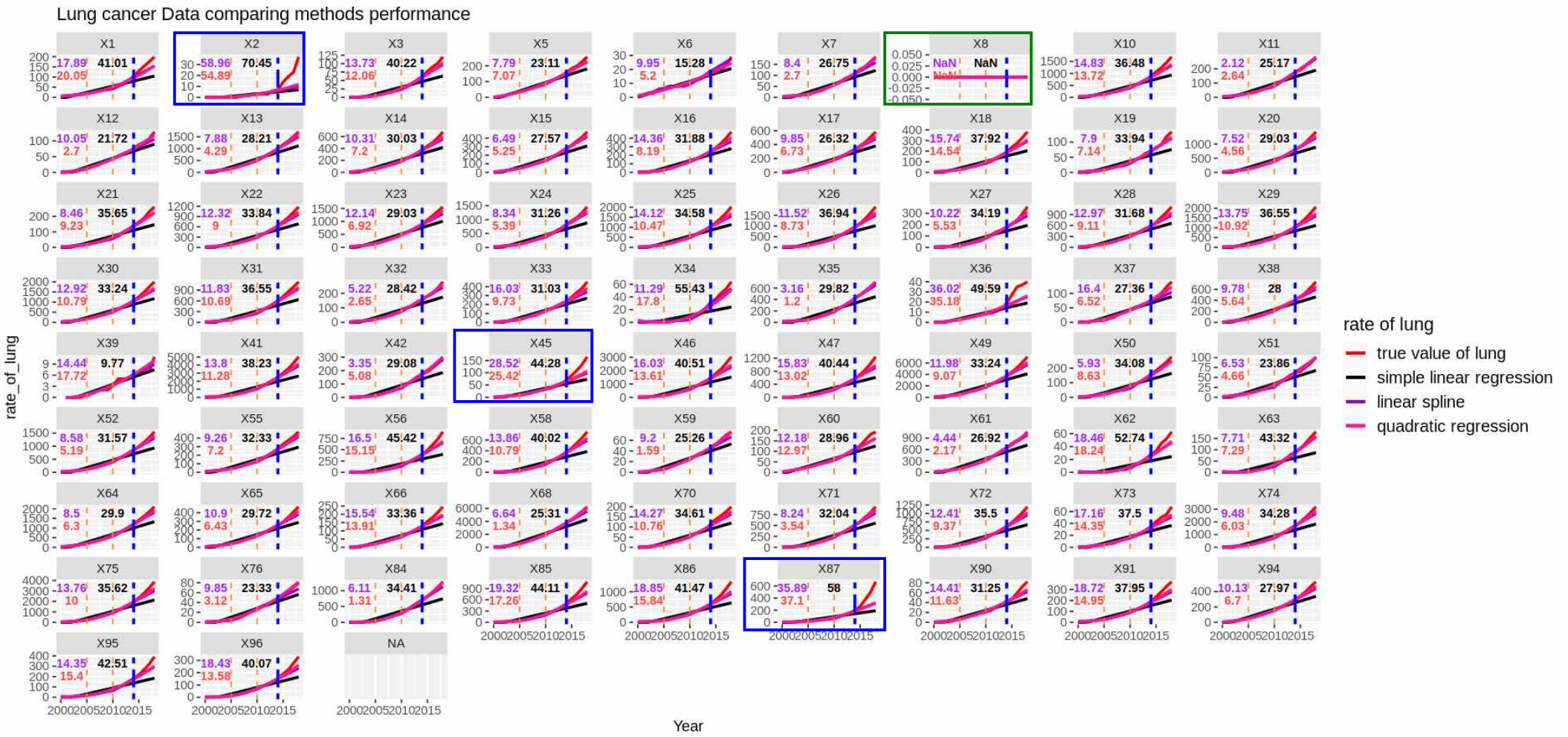
---------- Level 2 Data 파악 결과 정리 ----------

1. 폐암 누적 발생 건수는 지수적으로 증가하는 형태를 보인다.
2. 같은 사업장 내에서 남성보다는 여성 근로자 집단이, 입사 시기가 2000년 이전보다는 2000년 이후인 근로자 집단의 질병 누적 발생 건수 추세가 더 불안정하다.
3. 반응변수가 "누적 통합 누적 발생률"일 때와는 다르게, 백혈병 데이터보다 폐암 데이터에 대해 적합한 모형의 성능이 더 좋지 않다.
4. 반응변수가 "누적 통합 질병 발생률"이 아닌 "질병 누적 발생 건수"이다 보니, 시간이 지남에 따라 추세가 감소하는 부분이 전혀 없어 회귀모형이 이전보다 좋은 성능을 보이는 것을 확인하였다. - 모든 level data에서 보이는 공통사항

---

③-1. "UP2", "YEAR" 변수로 grouping한 데이터 : lv3_lung_total

: "UP2", "YEAR"로 grouping한 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 예측한 값을 실제값과 비교한 그래프는 아래와 같다. (UP2 기준으로 사업장 수가 많아 표로 MAPE 제시는 생략)



: **그래프의 파란색 실선은 YEAR=2014**년을 의미한다.

: 그래프 안, **주황색 글씨는 Polynomial 회귀 적합 통해 얻은 MAPE 값**, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

1) MAPE 값이 "NaN"을 가지는 사업장 : **초록색으로 강조**한 부분

NAME	사업장명
X8	광업 지원 서비스업

: lv1 data를 살펴보면, lv1_lung_SEX_CAL2에서 "XBF01", "XBF234"가 NAME인 사업장 / 집단에서 MAPE값이 "NaN"값을 가진 결과를 알 수 있다. 이를 통해 추측컨대, 대분류 기준 "B"에 속하는 중분류 사업장 05 ~ 08 중 사업장 08의 MAPE 값이 "NaN"이 원인이 되어 앞의 결과가 발생한 것이라고 볼 수 있다.

2) 연도에 따른 질병 누적 발생 건수의 추세가 2014년 이후로 급격히 증가하여 모형의 성능이 좋지 않은 사업장 : **파란색으로 강조**한 부분

NAME	사업장명
X2	임업
X45	자동차 및 부품 판매업
X87	사회복지 서비스업

: 위의 lv2 data 파악 결과와 비교해보았을 때 겹치는 사업장은 "2"(임업), 사업장 "87"(사회복지 서비스업)이다.

: 이 케이스에서도 세 모형의 성능이 모두 좋지 않게 나온다는 점에 주목해야 한다.

------ 백혈병 데이터와 다르게 질병 누적 발생 건수가 계단함수 형태를 보이는 사업장은 없다.

---------- 다음 페이지로 ----------

③-1. “UP2”, “SEX”, “YEAR” 변수로 grouping한 데이터 : lv3_lung_SEX

: “UP2”, “SEX”, “YEAR”로 grouping한 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 예측한 값을 실제값과 비교한 그래프는 아래와 같다. (UP2 기준으로 사업장 수가 많아 표로 MAPE 제시는 생략)

: 그래프의 파란색 실선은 YEAR=2014년을 의미한다.

: 그래프 안, 주황색 글씨는 Polynomial 회귀 적합 통해 얻은 MAPE 값, 보라색 글씨는 1차 spline function 적합 통해 얻은 MAPE 값, 검은색 글씨는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

③-1-1. SEX가 Female인 경우

1) MAPE 값이 “NaN”을 가지는 사업장 : **초록색으로 강조**한 부분

NAME	사업장명
X2F	임업
X8F	광업 지원 서비스업
X34F	산업용 기계 및 장비 수리업
X39F	환경 정화 및 복원업

: MAPE 값이 “NaN”값을 가지는 사업장의 그래프를 보면 2가지 경우가 있다. 2014년 이후에도 질병 누적 발생 건수가 0건인 경우와 2014년까지는 질병 누적 발생 건수가 0건이었다가 2014년 이후 질병 누적 발생 건수가 수직으로 증가하는 경우이다. 후자로는 “X6F”사업장이 해당 된다.

2) 연도에 따른 질병 누적 발생 건수의 추세가 계단 함수 모형을 띄는 사업장 : **빨간색으로 강조**한 부분

NAME	사업장명
X3F	어업
X5F	석탄, 원유 및 천연가스 광업
X7F	비금속 광물 광업; 연료용 제외
X19F	코크스 연탄 및 석유정제품 제조업
X36F	수도업
X37F	하수, 폐수 및 분뇨 처리업
X42F	전문직별 공사업
X50F	수상 운송업
X51F	항공 운송업
X76F	임대업; 부동산 제외

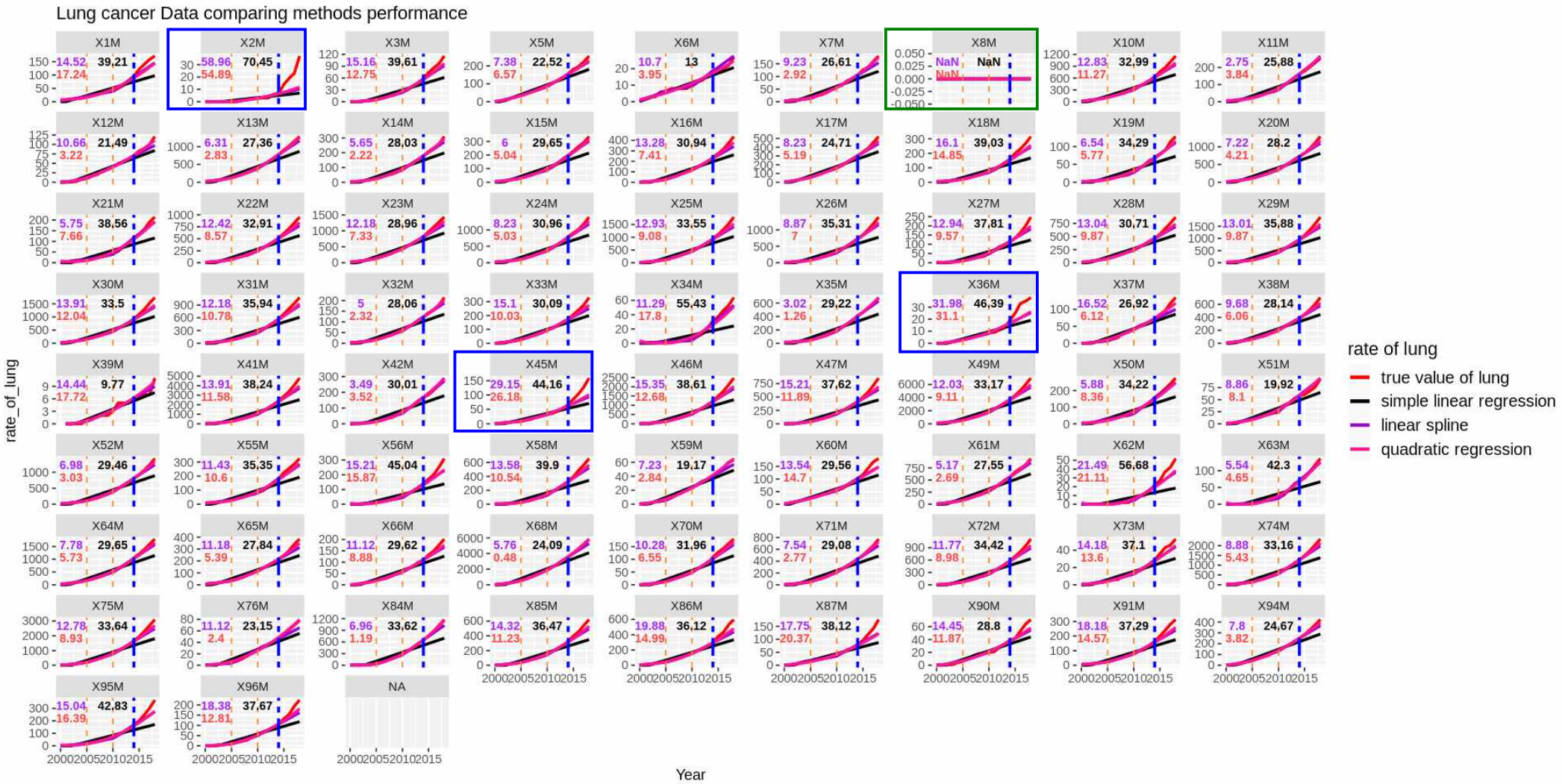
: 옆 표에 제시된 사업장들의 그래프를 보면 MAPE가 20 이하인 모형도 있지만, 대부분 모든 모형에 대해 큰 MAPE 값을 보인다. 계단 함수의 특성상, 일정 구간은 상수함수 형태이고 한 시점에서 급격한 증가가 일어나게 되므로 MAPE 값이 크게 나타나는 현상은 당연해 보이며, 이 경우에는 1차 spline function이 다른 모형에 비해 더 나은 성능을 보이는 것 같아 보인다.

3) 연도에 따른 질병 누적 발생 건수의 추세가 2014년 이후로 급격히 증가하여 모형의 성능이 좋지 않은 사업장 : **파란색으로 강조**한 부분

NAME	사업장명
X1F	농업
X52F	창고 및 운송관련 서비스업
X66F	금융 및 보험관련 서비스업
X70F	연구개발업
X73F	기타 전문, 과학 및 기술 서비스업
X87F	사회복지 서비스업

: 옆 표에 제시된 사업장들의 그래프를 보면 2014년 이전까지는 완만한 추세 곡선 / 직선을 보이다가 2014년 이후 누적 질병 발생 건수가 급격하게 증가하는 바람에 적합한 모형의 성능이 매우 떨어지는 것을 알 수 있다. 이 부분을 반영할 수 있는 다른 모형 혹은 추가 변수 생성이 필요해 보인다.

③-1-2. SEX가 Male인 경우



1) MAPE 값이 “NaN”을 가지는 사업장 : 초록색으로 강조한 부분

NAME	사업장명
X8F	광업 지원 서비스업

: 다른 집단 / 데이터(leukemia data)에 비해서 MAPE값이 “NaN”인 사업장의 수가 현저하게 적은 것을 알 수 있다.

2) 연도에 따른 질병 누적 발생 건수의 추세가 2014년 이후로 급격히 증가하여 모형의 성능이 좋지 않은 사업장 : **파란색으로 강조**한 부분

NAME	사업장명
X2M	임업
X36M	수도업
X45M	자동차 및 부품 판매업

: 옆 표에 제시된 사업장들의 그래프를 보면 2014년 이전까지는 완만한 추세 곡선 / 직선을 보이다가 2014년 이후 누적 질병 발생 건수가 급격하게 증가하는 바람에 적합한 모형의 성능이 매우 떨어지는 것을 알 수 있다. 이 부분을 반영할 수 있는 다른 모형 혹은 추가 변수 생성이 필요해 보인다.

------ 폐암 데이터의 경우, 남성 근로자 집단은 여성 근로자 집단에 비해 훨씬 더 안정적인 추세를 보이며, 현재 고려하고 있는 모형으로도 설명이 가능한 사업장이 많다.

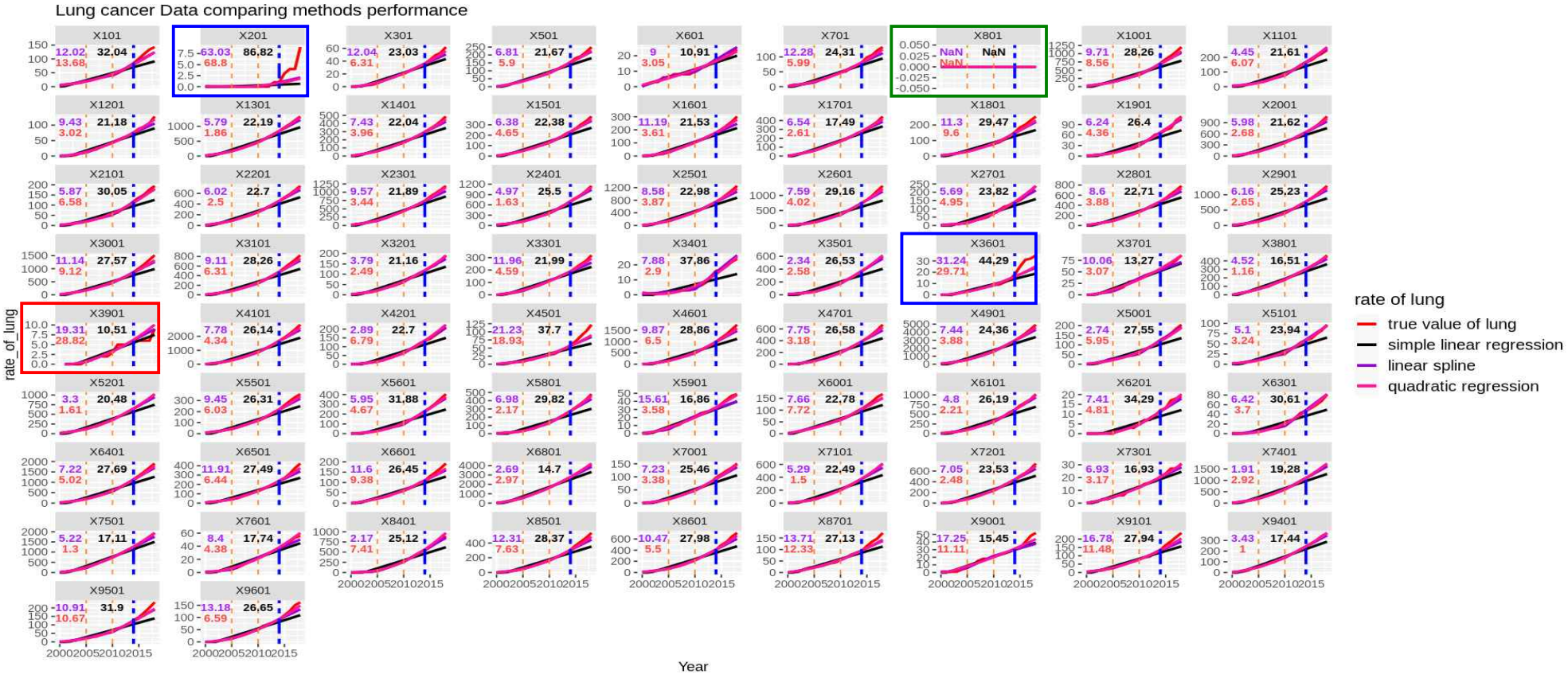
③-2. "UP2", "CAL2", "YEAR" 변수로 grouping한 데이터 : lv3_leukemia_SEX

: "UP2", "CAL2", "YEAR"로 grouping한 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 예측한 값을 실제값과 비교한 그래프는 아래와 같다. (UP2 기준으로 사업장 수가 많아 표로 MAPE 제시는 생략)

: **그래프의 파란색 실선은 YEAR=2014**년을 의미한다.

: 그래프 안, **주황색 글씨는 Polynomial 회귀 적합 통해 얻은 MAPE 값**, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

③-2-1. CAL2가 "01"인 경우



1) MAPE 값이 “NaN”을 가지는 사업장 : **초록색으로 강조**한 부분

NAME	사업장명
X801	광업 지원 서비스업

: MAPE 값이 “NaN”값을 가지는 사업장의 그래프를 보면 2가지 경우가 있다. 2014년 이후에도 질병 누적 발생 건수가 0건인 경우와 2014년까지는 질병 누적 발생 건수가 0건이었다가 2014년 이후 질병 누적 발생 건수가 수직으로 증가하는 경우이다. 사업장 “8”의 경우는 전자에 속한다.

2) 연도에 따른 질병 누적 발생 건수의 추세가 계단 함수 모형을 띄는 사업장 : **빨간색으로 강조**한 부분

NAME	사업장명
X3901	환경 정화 및 복원업

3) 연도에 따른 질병 누적 발생 건수의 추세가 2014년 이후로 급격히 증가하여 모형의 성능이 좋지 않은 사업장 : **파란색으로 강조**한 부분

NAME	사업장명
X201	임업
X3601	수도업

: 옆 표에 제시된 사업장들의 그래프를 보면 2014년 이전까지는 완만한 추세 곡선 / 직선을 보이다가 2014년 이후 누적 질병 발생 건수가 급격하게 증가하는 바람에 적합한 모형의 성능이 매우 떨어지는 것을 알 수 있다. 이 부분을 반영할 수 있는 다른 모형 혹은 추가 변수 생성이 필요해 보인다.

---------- 폐암 데이터의 경우, 입사 시기가 2000년 이전인 근로자 집단 대상으로 질병의 누적 발생 건수 추세가 많이 안정되어 있음을 알 수 있다. 위의 세 항목에 해당하는 사업장이 별로 없고, 그래프를 보면 시간에 따른 추세 모형이 “YEAR” 변수만으로도 설명이 되는 부분이 많이 있는 것을 알 수 있다.

---------- 다음 페이지로 ----------

③-2-2. CAL2가 “234”인 경우

1) MAPE 값이 “NaN”을 가지는 사업장 : 초록색으로 강조한 부분

NAME	사업장명
X8234	광업 지원 서비스업
X12234	담배 제조업
X39234	환경 정화 및 복원업

: MAPE 값이 “NaN”값을 가지는 사업장의 그래프를 보면 2가지 경우가 있다. 2014년 이후에도 질병 누적 발생 건수가 0건인 경우와 2014년까지는 질병 누적 발생 건수가 0건이었다가 2014년 이후 질병 누적 발생 건수가 수직으로 증가하는 경우이다. 후자로는 “X12234”, “X39234” 사업장이 해당 된다.

: **파란색으로 강조한 부분**은 입사 시기가 2000년 이전인 근로자 집단에서도 동일한 현상을 보이는 사업장을 의미한다.

2) 연도에 따른 질병 누적 발생 건수의 추세가 계단 함수 모형을 띄는 사업장 : **빨간색으로 강조**한 부분

NAME	사업장명
X6234	어업
X51234	석탄, 원유 및 천연가스 광업

: 계단 함수의 특성상, 일정 구간은 상수함수 형태이고 한 시점에서 급격한 증가가 일어나게 되므로 MAPE 값이 크게 나타나는 현상은 당연해 보인다.

3) 연도에 따른 질병 누적 발생 건수의 추세가 2014년 이후로 급격히 증가하여 모형의 성능이 좋지 않은 사업장 : **파란색으로 강조**한 부분

NAME	사업장명
X1234	농업
X2234	**임업**
X5234	석탄, 원유 및 천연가스 광업
X19234	코크스, 연탄 및 제조업
X36234	**수도업**
X37234	하수, 폐수 및 분뇨업
X45234	자동차 및 부품 판매업
X60234	방송업
X70234	연구 개발업
X86234	보건업
X87234	사회복지 서비스업

: 옆 표에 제시된 사업장들의 그래프를 보면 2014년 이전까지는 완만한 추세 곡선 / 직선을 보이다가 2014년 이후 누적 질병 발생 건수가 급격하게 증가하는 바람에 적합한 모형의 성능이 매우 떨어지는 것을 알 수 있다. 이 부분을 반영할 수 있는 다른 모형 혹은 추가 변수 생성이 필요해 보인다.

: **파란색으로 강조한 부분**은 입사 시기가 2000년 이전인 근로자 집단에서도 동일한 결과를 보였던 사업장을 의미한다.

: 2000년 이전에 입사한 근로자 집단의 질병 누적 발생 건수 추세와 비교해보았을 때, 2014년 이후 추세가 급격히 증가하는 사업장이 훨씬 더 많다는 점을 알 수 있다.

---------- 다음 페이지로 ----------

③-3. “UP2”, “SEX”, “CAL2”, “YEAR” 변수로 grouping한 데이터 : lv3_leukemia_SEX_CAL2

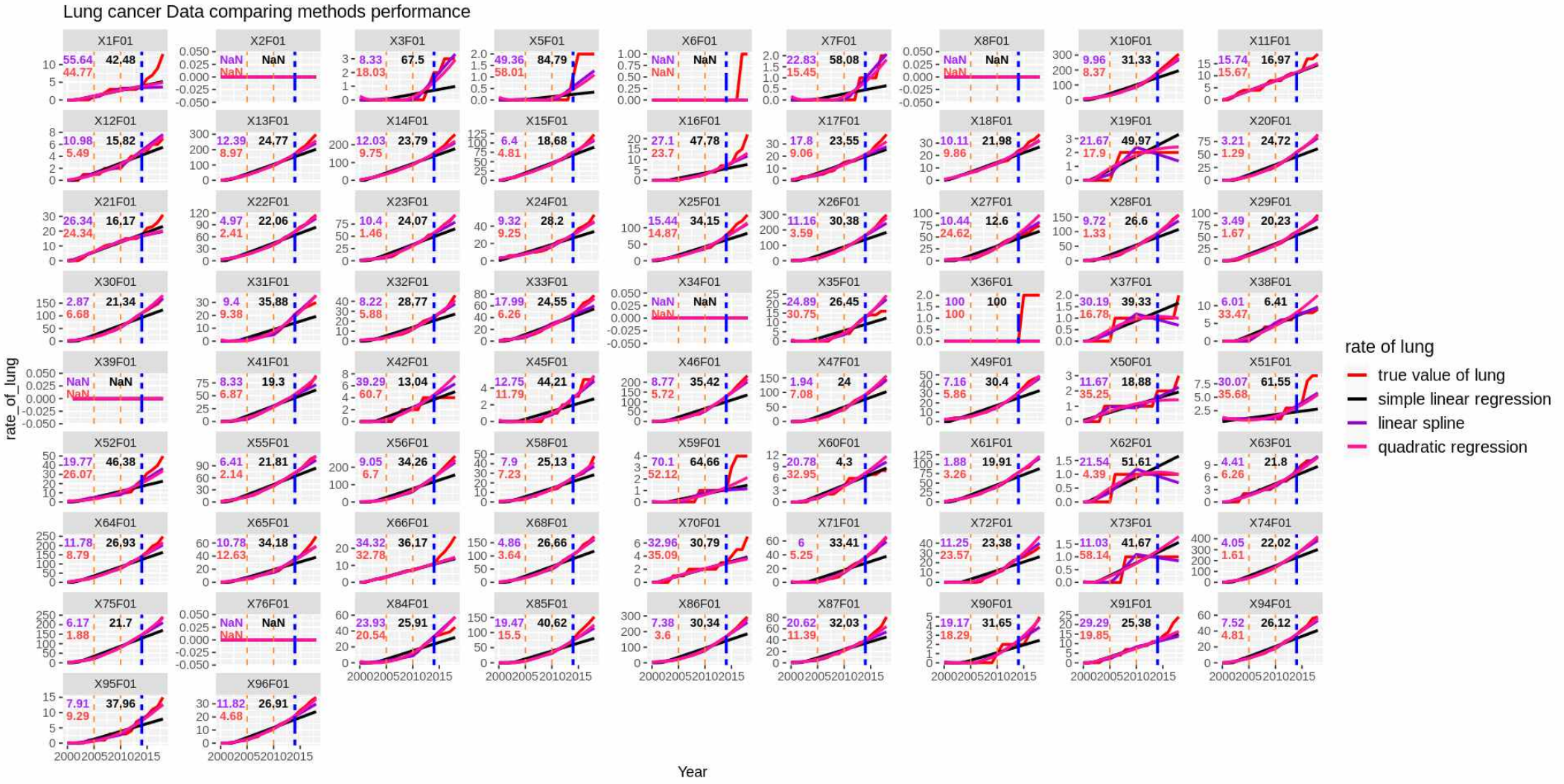
: “UP2”, “SEX”, “CAL2”, “YEAR”로 grouping한 데이터에 대해 3가지 모형을 적합하고 validation set을 대상으로 예측한 값을 실제값과 비교한 그래프는 아래와 같다. (UP2 기준으로 사업장 수가 많아 표로 MAPE 제시는 생략)

: 총 4가지 subset으로 나누어 그래프 제시 (“F01”, “F234”, “M01”, “M234”)

: **그래프의 파란색 실선은 YEAR=2014**년을 의미한다.

: 그래프 안, **주황색 글씨는 Polynomial 회귀 적합 통해 얻은 MAPE 값**, **보라색 글씨**는 **1차 spline function 적합 통해 얻은 MAPE 값**, **검은색 글씨**는 **단순선형회귀모형 적합 통해 얻은 MAPE 값**을 의미한다.

③-3-1. Female + CAL2 = “01”



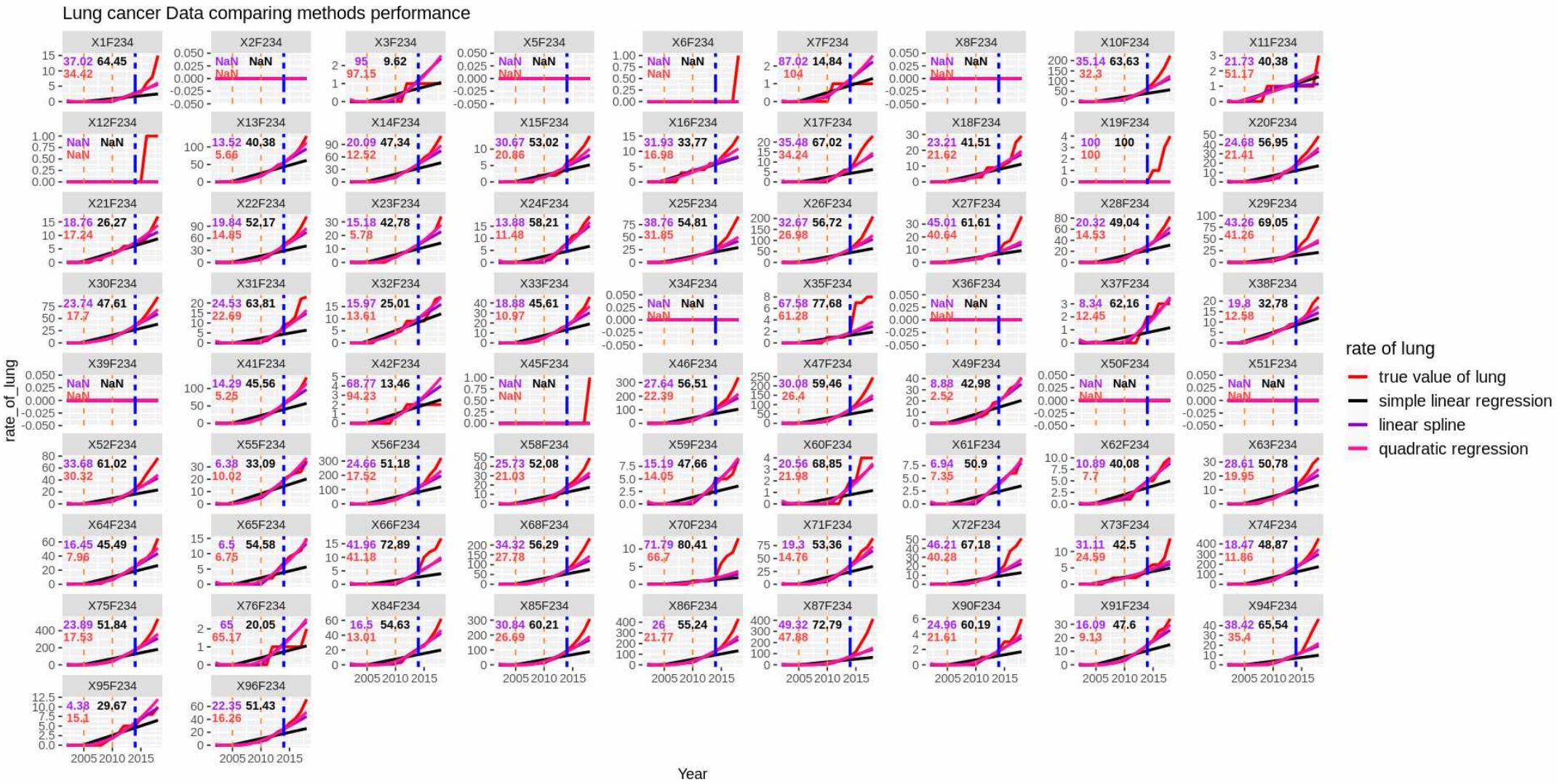
Type	NAME	사업장명
MAPE값이 “NaN”인 사업장	X2F01	임업
	X6F01	금속 광업
	X8F01	광업 지원 서비스업
	X34F01	산업용 기계 및 장비 수리업
	X36F01	수도업
	X39F01	환경 정화 및 복원업
	X76F01	임대업; 부동산 제외
누적 질병 발생 건수 추세가 계단 함수 모형을 보이는 사업장	X3F01	어업
	X5F01	석탄, 원유 및 천연가스 광업
	X7F01	비금속광물 광업; 연료용 제외
	X19F01	코크스, 연탄 및 석유 정제품 제조업
	X36F01	수도업
	X37F01	하수, 폐수 및 분뇨 처리업
	X50F01	수상 운송업
	X51F01	항공 운송업
	X59F01	영상, 오디오 기록물 제작 및 배급업
	X62F01	컴퓨터 프로그래밍, 시스템 통합 및 관리업
	X73F01	기타 전문, 과학 및 기술 서비스업
	X90F01	창작, 예술 및 여가관련 서비스업
2014년 이후 누적 질병 발생 건수 추세가 급격히 증가하는 사업장	X1F01	농업
	X16F01	목재 및 나무제품 제조업; 가구제외
	X51F01	항공 운송업
	X66F01	금융 및 보험관련 서비스업
	X70F01	연구개발업
	X91F01	스포츠 및 오락관련 서비스업

--- 여성이면서 입사 시기가 2000년 이전인 근로자 집단의 경우, 누적 질병 발생 건수의 추세가 계단 함수 모형을 띄는 사업장이 제일 많았다.

--- 이전에 파악한 결과와 비교해보면, 입사 시기가 2000년 이전인 근로자 집단의 경우 누적 질병 발생 건수 추세가 안정된 경우가 대부분이었는데, 여성이라는 성별 관련 층화가 추가되면서 불안정함이 보이는 사업장이 보인다.

---------- 다음 페이지로 ----------

③-3-2. Female + CAL2 = “234”



---------- 다음 페이지로 ----------

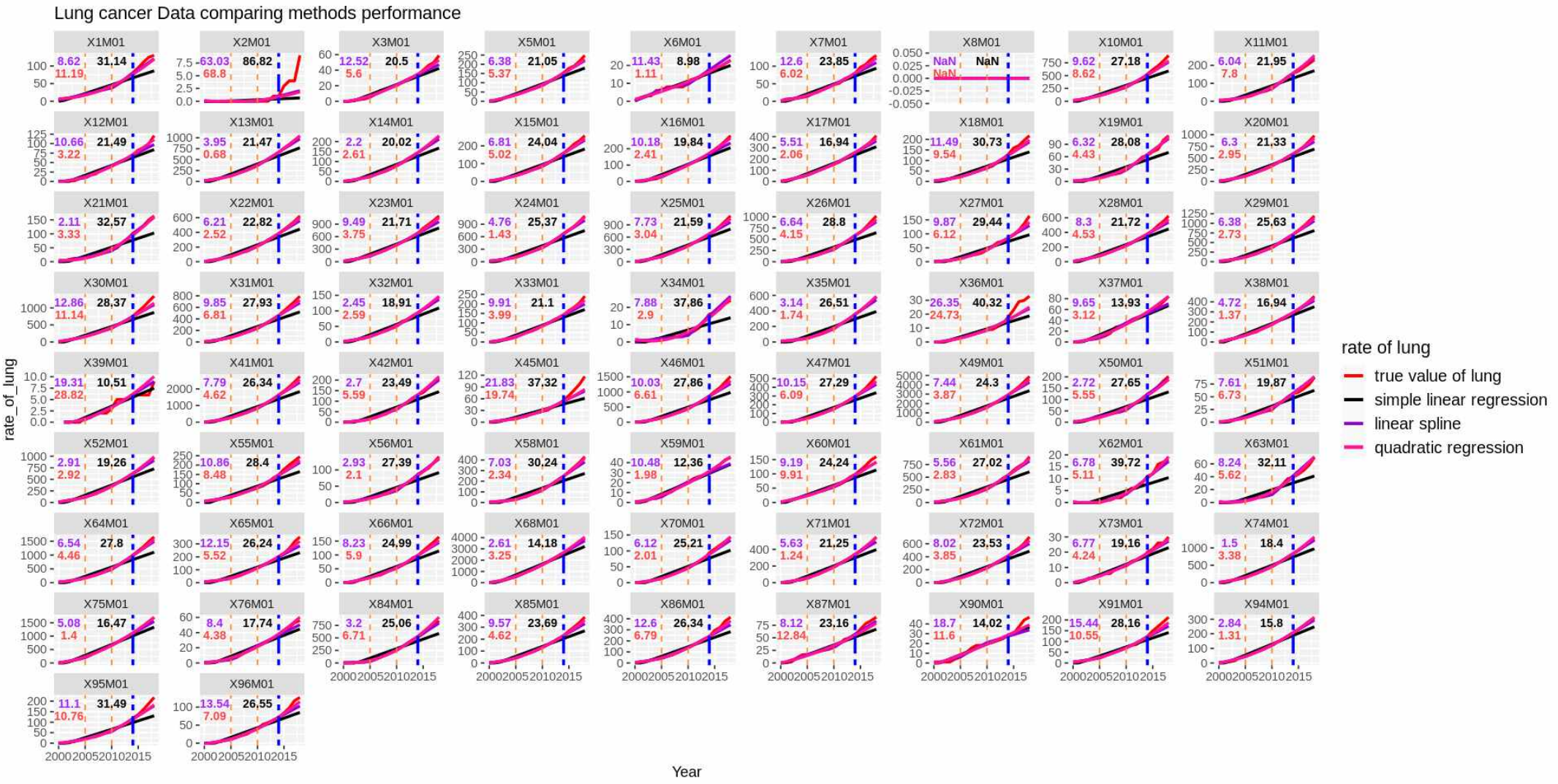
Type	NAME	사업장명
MAPE값이 “NaN”인 사업장	**X2F234**	**임업**
	X5F234	석탄, 원유 및 천연가스 광업
	X6F234	**금속 광업**
	X8F234	**광업 지원 서비스업**
	X12F234	담배 제조업
	X34F234	**산업용 기계 및 장비 수리업**
	X36F234	**수도업**
	X39F234	**환경 정화 및 복원업**
	X45F234	자동차 및 부품 판매업
	X50F234	수상 운송업
	X51F234	항공 운송업
누적 질병 발생 건수 추세가 계단 함수 모형을 보이는 사업장	**X3F234**	**어업**
	X7F234	**비금속 광물, 광업; 연료용 제외**
	X11F234	음료 제조업
	X42F234	전문직별 공사업
	X60F234	방송업
	X65F234	보험 및 연금업
	X76F234	임대업; 부동산 제외
2014년 이후 누적 질병 발생 건수 추세가 급격히 증가하는 사업장	**X1F234**	**농업**
	X10F234	식료품 제조업
	X17F234	펄프, 종이 및 종이제품 제조업
	X25F234	금속 가공제품 제조업; 기계 및 가구 제외
	X26F234	전자 부품, 컴퓨터, 영상, 음향 및 통신장비 제조업
	X27F234	의료, 정밀, 광학 기기 및 시계 제조업
	X29F234	기타 기계 및 장비 제조업
	X35F234	전기, 가스, 증기 및 공기 조절 공급업
	X52F234	창고 및 운송관련 서비스업
	X66F234	**금융 및 보험관련 서비스업**
	X68F234	부동산업
	X70F234	**연구개발업**
	X72F234	건축 기술, 엔지니어링 및 기타 과학기술 서비스업
	X73F234	기타 전문, 과학 및 기술 서비스업
	X85F234	교육 서비스업
	X87F234	사회복지 서비스업
	X94F234	협회 및 단체

--- 여성이면서 입사 시기가 2000년 이후인 근로자 집단의 경우, 2014년 이후 누적 질병 발생 건수의 추세가 급격히 증가하는 사업장이 많다.

--- 이전에 파악한 결과와 비교해보면, 입사 시기가 2000년 이후인 근로자 집단의 경우에도 누적 질병 발생 건수 추세가 2014년 이후 급격히 증가하는 경우가 많았는데, 여성이라는 성별 관련 층화가 추가되면서 추세의 불안정함이 더 커진 사업장이 보인다.

--- 파란색으로 강조한 부분은 여성이면서 입사 시기가 2000년 이전인 근로자 집단에 대해 파악한 결과와 동일한 내역에 속하는 사업장을 의미한다.

③-3-3. Male + CAL2 = “01”



Type	NAME	사업장명
MAPE값이 “NaN”인 사업장	X8M01	광업 지원 서비스업
누적 질병 발생 건수 추세가 계단 함수 모형을 보이는 사업장	X39M01	환경 정화 및 복원업
2014년 이후 누적 질병 발생 건수 추세가 급격히 증가하는 사업장	X2M01	임업
	X36M01	수도업
	X45M01	자동차 및 부품 판매업

---- 폐암 데이터의 경우, 남성 근로자 집단, 입사 시기가 2000년 이전인 근로자 집단 각각 전체를 보았을 때에도 누적 질병 발생 건수 추세의 불안정함이 거의 없었다. 층화를 더 추가했을 때에도 추세의 불안정함이 다른 집단에 비해 많이 적다는 것을 알 수 있다.

③-3-4. Male + CAL2 = “234”

Type	NAME	사업장명
MAPE값이 “NaN”인 사업장	**X8M234**	**광업 지원 서비스업**
	X12M234	담배 제조업
	X39M234	환경 정화 및 복원업
누적 질병 발생 건수 추세가 계단 함수 모형을 보이는 사업장	X6M234	금속 광업
	X51M234	항공 운송업
	X65M234	우편 및 통신업
2014년 이후 누적 질병 발생 건수 추세가 급격히 증가하는 사업장	X1M234	농업
	X2M234	**임업**
	X5M234	석탄, 원유 및 천연가스 광업
	X11M234	음료 제조업
	X19M234	코크스, 연탄 및 석유정제품 제조업
	X33M234	기타 제품 제조업
	X36M234	**수도업**
	X37M234	하수, 폐수 및 분뇨 처리업
	X45M234	**도매 및 상품 중개업**
	X60M234	방송업
	X86M234	보건업
	X87M234	사회복지 서비스업

---- 폐암 데이터의 경우, 남성 근로자 집단, 입사 시기가 2000년 이후인 근로자 집단 각각 전체를 보았을 때 누적 질병 발생 건수 추세의 불안정함은 입사 시기 2000년 이후인 집단에서만 보였다. 층화를 더 추가했을 때에는 추세의 불안정함이 더 커지진 않았지만, 남성이면서 입사 시기가 2000년 이전인 집단에 비해 누적 질병 발생 건수 추세의 불안정함은 더 크다.

---------- Level 3 data total 정리 ----------

1. 사업장을 중분류(UP2) 기준으로 나누었을 때, 반응변수를 “질병 누적 통합 발생률”로 했을 때의 파악 결과와 비슷하게 남성보다는 여성 근로자 집단이, 입사 시기가 2000년 이전보다 이후인 근로자 집단의 추세가 더 불안정하다. - leukemia data와의 공통점
2. 특히, 입사 시기가 2000년 이후인 집단의 질병 누적 발생 건수 추세는 2014년 이후로 급격히 증가하는 경향을 많이 보인다. - leukemia data와의 공통점
3. 눈에 띄는 사업장은 광업 쪽(“6”, “8”)과 담배 제조업(“12”)이며, 특히 사업장 “8”을 주목할 필요가 있다.

------ Leukemia data VS Lung cancer data ------

1. 반응변수가 통합 누적 발생률 일때는 같은 모형임에도 폐암 데이터를 적합했을 때 더 좋은 성능을 보였는데, 반응변수를 질병 누적 발생 건수로 변경하니 같은 모형임에도 백혈병 데이터를 적합해서 얻은 MAPE 값이 더 낮은 경우가 많았다.
