산보연 Data forecasting 07월 15일 Update Version

작성자 : 이은경

< What to do >

: Data level 별(level1, level2 – UP1으로 grouping, level3 – UP2로 grouping) 설명변수를 YEAR, 반응변수를 질병 통합 누적 발생률로 하여 단순선형회귀모형, 1차 spline 두 가지 모형 적합

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

: train data로 각각의 단순선형회귀모형, spline model 적합 후, validation data의 통합 누적 발생률 예측

: validation data의 연도에 따른 실제 통합 누적 발생률 알고 있으므로 예측값과 비교

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

(통합 누적 발생률 true value 중 0이 있어 MAPE값이 "NaN"값이 나오는 사업장이 있는 경우, 이 사업장의 MAPE는 "-99"로 대체)

: Data level별 적합한 단순선형회귀모형, spline model의 performance를 시각적으로 표현하기 위해 plots 생성

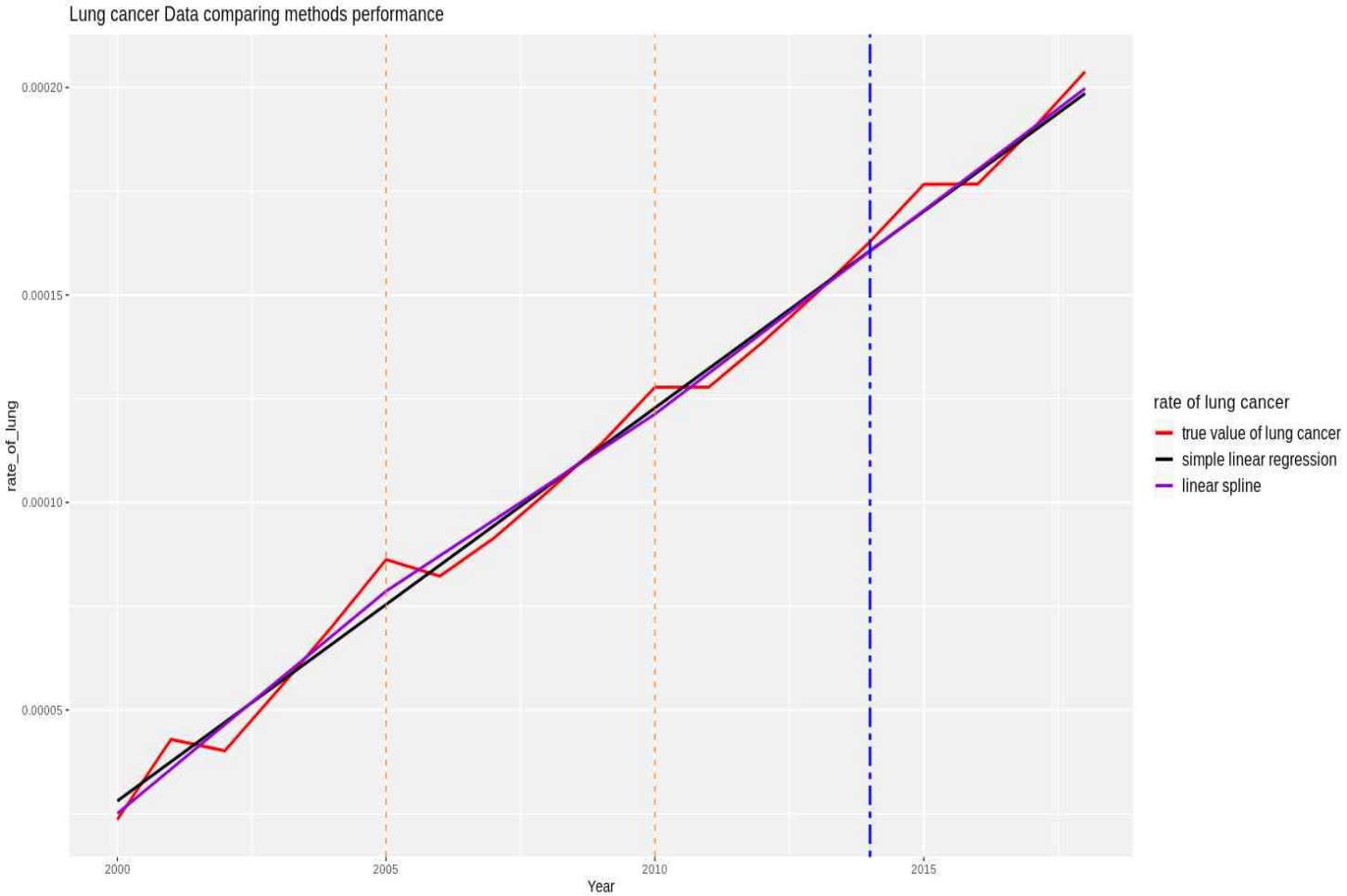
[Result 제시]

1) Lung Cancer Data

① Level 1 data

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.



: 각 model별 MAPE는 아래 표와 같다.

	단순선형회귀모형	1차 spline function
MAPE	2.05	1.92

: 그래프나, MAPE 결과를 보아도 level1 data에서는 단순선형회귀모형이나 1차 spline function 두 모형 모두 비슷한 performance를 보이는 것을 알 수 있다. 그럼에도 불구하고, 1차 spline function이 조금 더 작은 MAPE를 보이는 이유는 실제 통합 누적 발생률 추세에서 어느 순간 증가하다가 감소하는 경향을 보이는데, 이를 조금 더 잘 반영했기 때문이라고 추측한다.

---

② Level 2 data

: UP1 별 단순선형회귀모형 적합 후 얻은 MAPE, 1차 spline function 적합 후 얻은 MAPE를 작성한 표이다.

: 표의 순서는 UP1 기준으로 grouping 한 후, 2018년 기준으로 계산한 추적 인년 합계의 순위이다. (높은 순에서 낮은 순으로 정렬)

UP1	단순선형회귀모형 적합 후 얻은 MAPE	1차 spline function 적합 후 얻은 MAPE
C	9.28	3.79
G	22.37	10.02
Q	2.31	2.35
F	4	2.18
N	10.02	4.79
H	7.73	3.24
J	7.31	2.59
K	6.01	3.72
M	7.14	1.97
L	9.62	5.12
I	1.86	4.01
P	5.2	1.1
S	9.7	1.3
O	5.42	2.21
D	6.13	2.63
R	4.89	6.33
E	17.61	6.66
A	11.12	9.97
“ ”	6.4	11.52
B	2.03	2.81
U	-99	-99
T	11.12	6.01

: 두 모형 적합 통해 얻은 MAPE 비교 후, **더 낮은 MAPE 값**을 가지는 부분에 **빨간색**으로 표시하였다.

: 사업장 “U”(국제 및 외국기관)은 실제 폐암 발생률 값 중 0이 있어 MAPE가 “NaN”값이 나왔다.

: 전체적으로 UP1 기준으로 grouping 한 데이터에서는 1차 spline function의 성능이 더 좋은 것을 확인할 수 있다.

---------- 다음 페이지로 ----------

: 적합 결과를 예측한 결과를 시각화

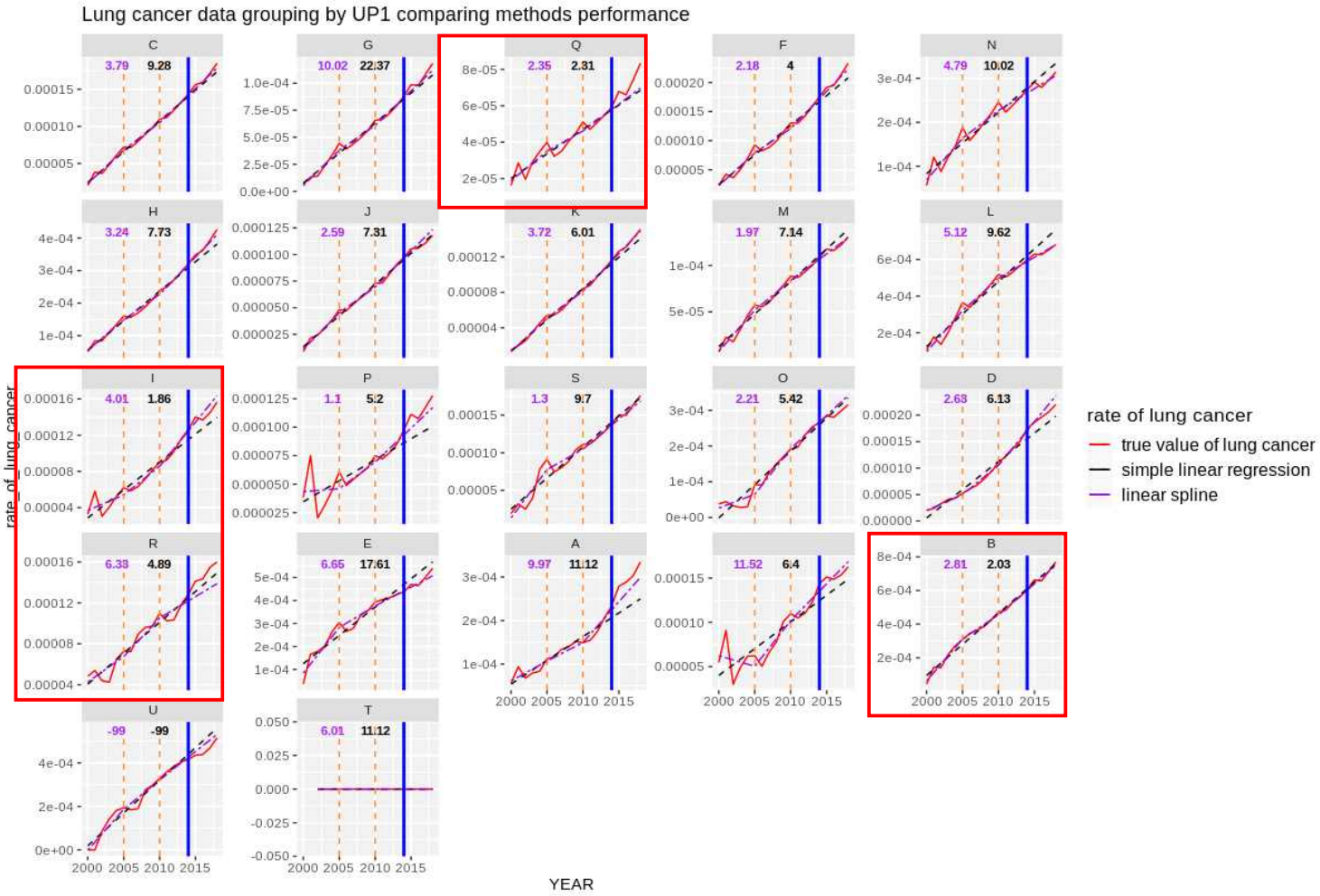
1) Version 1

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP1 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

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



: 그래프를 보면, 2014년 이전의 데이터, 즉, train data는 1차 spline function이 훨씬 더 실제 폐암 통합 누적 발생률 변화를 잘 학습한 것으로 보인다.

: 1차 spline function보다 단순선형회귀모형이 더 좋은 그래프에 **빨간색 테두리**를 쳐놓았다. 해당 사업장에 대한 정보는 아래 표와 같다.

UP1	사업장명	단순선형회귀모형이 더 좋은 성능을 보이는 이유 추측
R	예술, 스포츠, 여가 관련 서비스업	실제 통합 누적 발생률과 적합값 차이가 1차 spline model에서 더 두드러짐.
I	숙박 및 음식점업	2014년 이후 통합 누적 발생률이 잠시 감소하는 부분이 있는데 1차 spline model은 지속적으로 발생률이 증가한다고 예상.
B	광업	2016년 ~ 2017년 구간 외에 다른 구간에서 실제 통합 누적 발생률과 적합값 차이가 1차 spline model에서 더 두드러짐.
Q	보건업 및 사회복지 서비스업	추측이 되지 않음.

2) Version 2

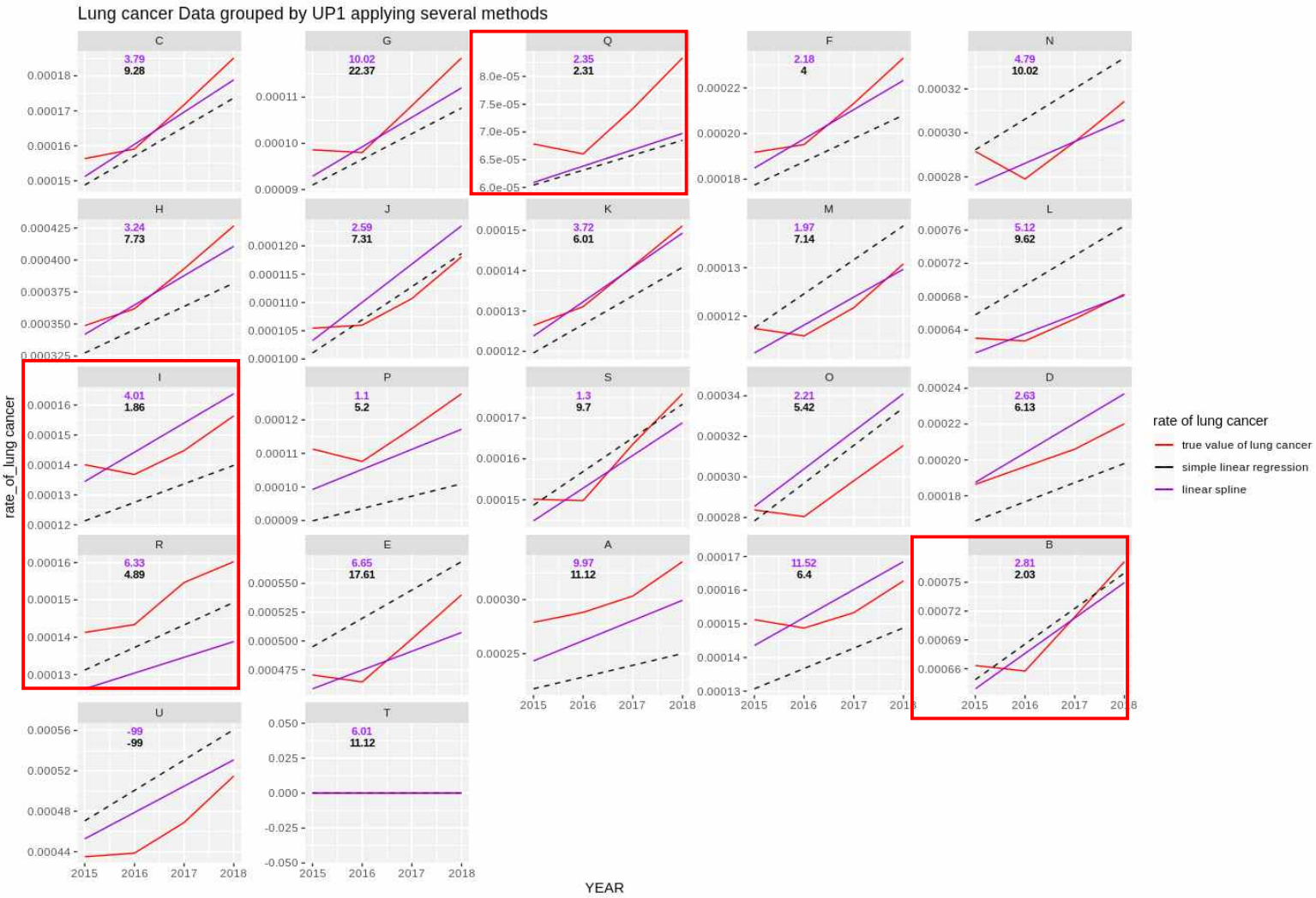
: Version 1의 그래프에서 validation data 시각화 부분이 눈에 띄지 않아, 2015년 ~ 2018년에 대해 각 사업장(UP1) 별 실제 통합 누적 발생률과 적합한 모형 통해 예측한 값 plotting

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP1 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

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



: 1차 spline function보다 단순선형회귀모형이 더 좋은 그래프에 **빨간색 테두리**를 쳐놓았다.

---------- 다음 페이지로 ----------

③ Level 3 data

: UP2 별 단순선형회귀모형 적합 후 얻은 MAPE, 1차 spline function 적합 후 얻은 MAPE를 작성한 표이다.

: 표의 순서는 UP2 기준으로 grouping 한 후, 2018년 기준으로 계산한 추적 인년 합계의 순위이다. (높은 순에서 낮은 순으로 정렬)

UP2	단순선형회귀모형 적합 후 얻은 MAPE	1차 spline function 적합 후 얻은 MAPE
46	7.88	3.76
26	8.81	1.77
41	8.04	3.21
75	8.02	5.18
86	9.21	6.56
49	10.05	3.62
64	5.81	1.36
30	7.11	3.23
29	5.17	2.8
47	2.88	2.88
58	8.35	1.6
68	9.7	1.3
25	2.1	5.68
85	17.61	6.65
10	10.76	3.72
72	2.67	2.43
28	2.08	3.84
87	20.25	17.62
71	7.36	3.94
20	1.81	3.08
13	5.24	1.79
74	2.85	3.53
22	3.6	2.37
52	2.04	3.22
56	10.15	5.92
61	4.92	4.35
31	3.34	1.68
24	7.27	2.45
14	4.35	2.47
23	3.4	5.72
84	4.89	6.33
65	7.35	3.47
35	8.99	6.98
27	3.25	4.01
95	11.86	2.43
55	8.51	3.4
94	9.75	2.12
18	16.33	5.97
66	3.59	6.84
17	2.43	3.09
21	6.12	5.98
91	10.25	11.72
33	2.06	10.58
70	10.08	5.24
15	3.46	2.56
96	5.74	6.75
45	19.27	22.27
63	4.43	15.42
38	5.7	2.17
11	1.4	12.91

UP2	단순선형회귀모형 적합 후 얻은 MAPE	1차 spline function 적합 후 얻은 MAPE
32	2.92	5.42
16	2.01	9.59
62	12.15	3.19
60	4.41	4.95
42	2.23	16.2
73	8.15	8.67
51	4.79	20.96
50	8.61	8.1
1	23.28	7.96
59	16.45	6.1
90	11.84	11.6
19	8.39	5.62
34	44.31	18.84
76	15.41	9.57
99	11.12	6.01
7	3.53	3.93
37	18.17	16.44
5	3.01	3.16
3	10.76	3.57
12	2.12	7.03
36	26.49	33.15
2	47.17	59.78
39	65.01	16.5
6	8.75	21.59
97	-99	-99
8	-99	-99
98	-99	-99

: 두 모형 적합 통해 얻은 MAPE 비교 후, **더 낮은 MAPE 값**을 가지는 부분에 **빨간색**으로 표시하였다.

: 전반적으로 1차 spline function의 성능이 더 좋아 보인다.

: 단순선형회귀모형이 더 좋은 성능을 보이는 사업장들 보면, 추적 인년 합계가 적은 경향을 보인다.

: 사업장 “97”, “8”, “98”은 MAPE 값이 “NaN”이다.

: 적합 결과를 예측한 결과를 시각화

1) Version 1

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP2 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

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

: 1차 spline function보다 단순선형회귀모형의 성능이 더 좋은 사업장에 대해 **빨간색 테두리**를 그렸다.

: 해당 사업장에 대한 정보를 취합해보면 다음과 같다.

---

1) 대분류 "B"에 속해있는 중분류 사업장 전부("05", "06", "07")가 단순선형회귀모형 적합 때 좋은 성능을 보인다.

2) 단순선형회귀모형이 더 좋은 성능을 보이는 사업장 대부분(40%)은 대분류 기준으로 모두 "C"(제조업)에 포함된다.

: 전반적으로 1차 spline function의 성능이 더 좋은 경향을 보인다.

---------- 다음 페이지로 ----------

2) Version 2

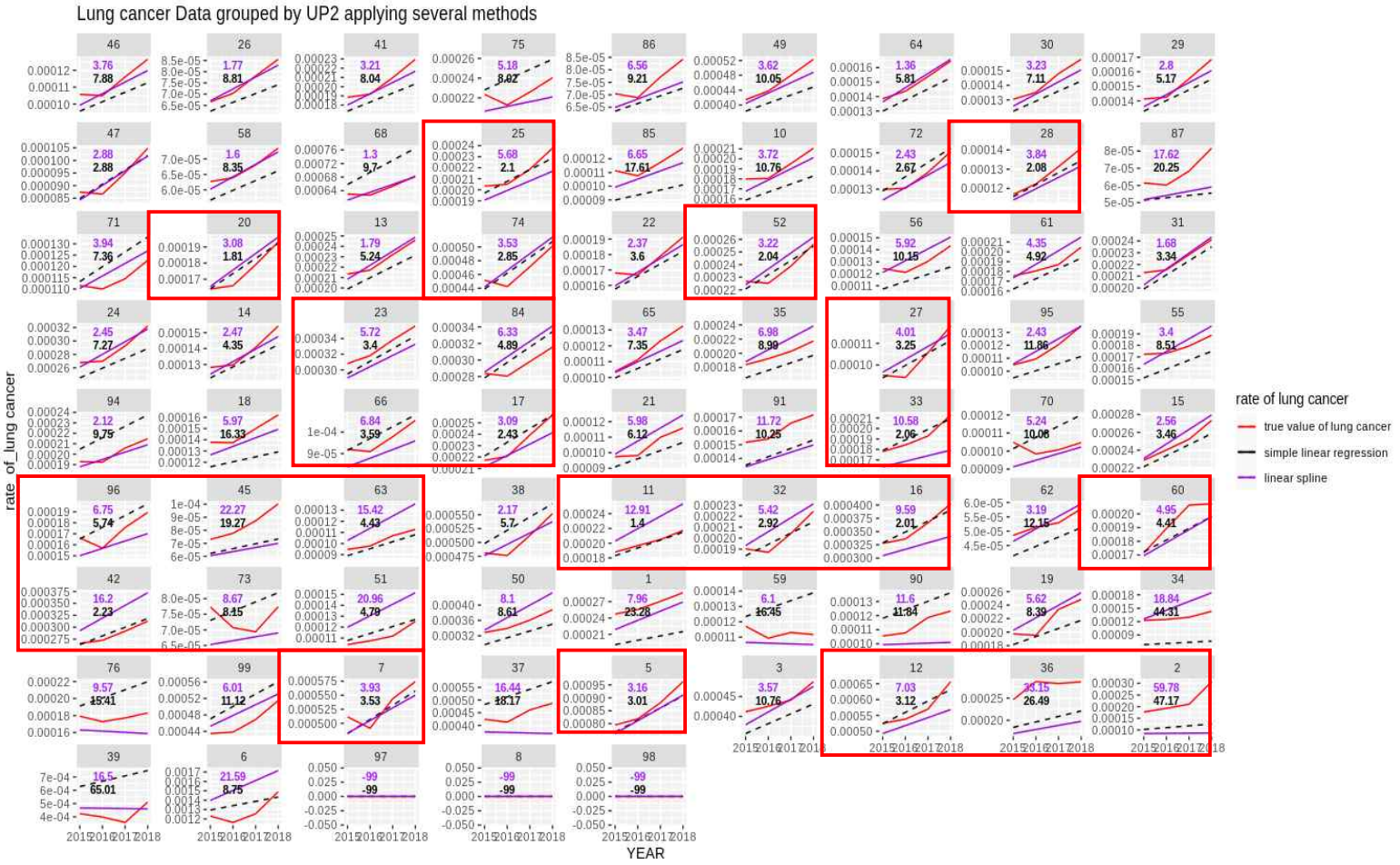
: Version 1의 그래프에서 validation data 시각화 부분이 눈에 띄지 않아, 2015년 ~ 2018년에 대해 각 사업장(UP2)별 실제 통합 누적 발생률과 적합한 모형 통해 예측한 값 plotting

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP2 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

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



: 1차 spline function보다 단순선형회귀모형의 성능이 더 좋은 사업장에 대해 **빨간색 테두리**를 그렸다.

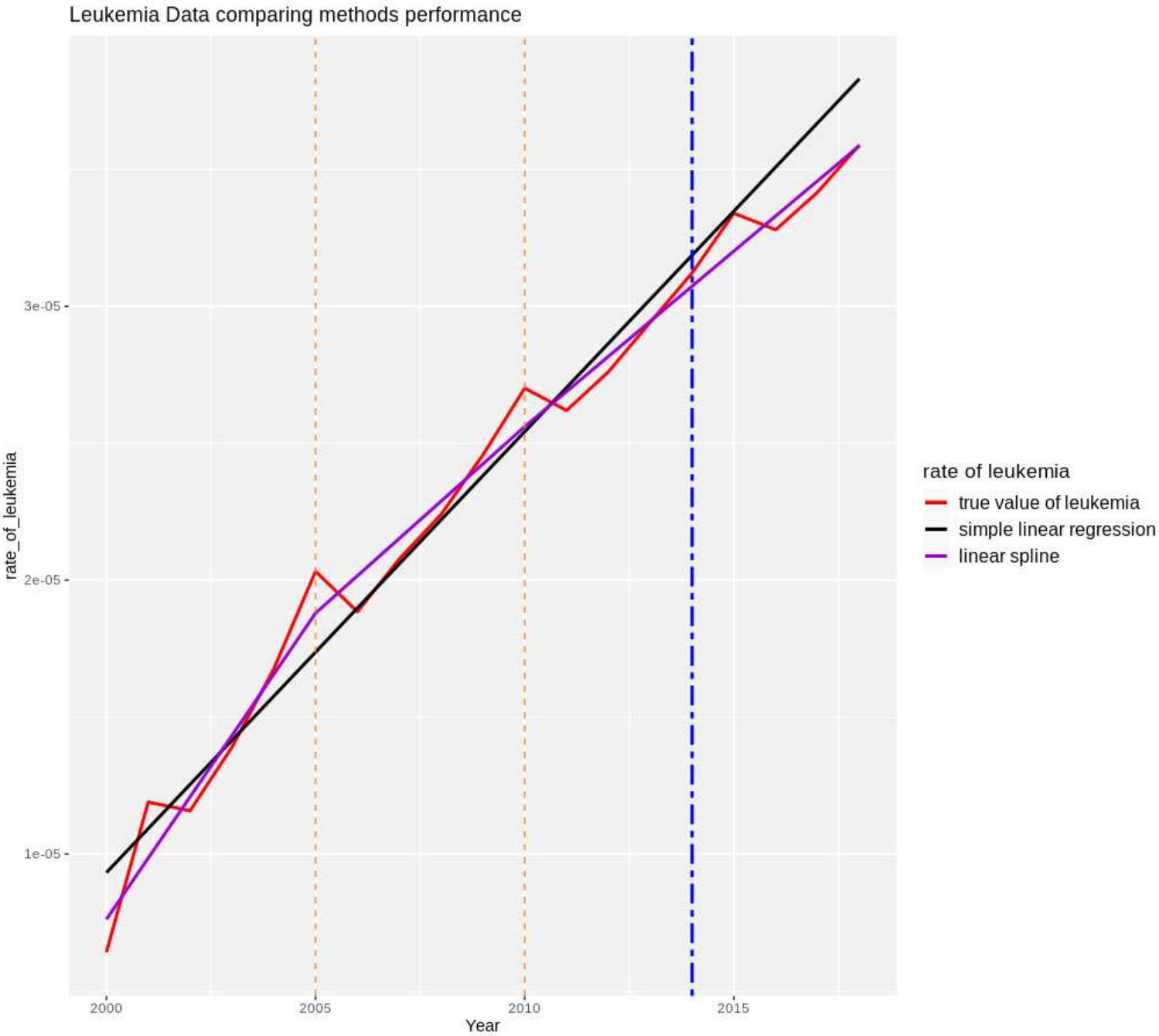
---------- 다음 페이지로 ----------

1) Leukemia Data

① Level 1 data

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.



: 각 model별 MAPE는 아래 표와 같다.

	단순선형회귀모형	1차 spline function
MAPE	5.33	1.74

: 그래프나, MAPE 결과를 보아도 level1 data에서는 폐암 데이터와 다르게 1차 spline function이 더 나은 performance를 보이는 것을 알 수 있다. 백혈병 통합 누적 발생률이 주춤하는 경향을 1차 spline function이 더 잘 학습한 것으로 보인다.

② Level 2 data

: UP1 별 단순선형회귀모형 적합 후 얻은 MAPE, 1차 spline function 적합 후 얻은 MAPE를 작성한 표이다.

: 표의 순서는 UP1 기준으로 grouping 한 후, 2018년 기준으로 계산한 추적 인년 합계의 순위이다. (높은 순에서 낮은 순으로 정렬)

UP1	단순선형회귀모형 적합 후 얻은 MAPE	1차 spline function 적합 후 얻은 MAPE
C	4.89	1.02
G	10.94	3.62
Q	6.39	6.77
F	6.94	1.43
N	10.91	6.33
H	2.31	4.78
J	12.95	4.81
K	3.06	6.03
M	5.17	3.76
L	2.32	5.37
I	5.45	3.6
P	9.61	5.04
S	14.32	3.75
O	32.77	5.06
D	16.52	20.04
R	11.9	11.09
E	29.79	5.24
A	38.96	7.32
“ ”	10.21	16.1
B	13.14	3.93
U	3.57	10.27
T	100	100

: 두 모형 적합 통해 얻은 MAPE 비교 후, **더 낮은 MAPE 값**을 가지는 부분에 **빨간색**으로 표시하였다.

: 추적 인년 합계의 순위와 모형의 성능은 연관성이 있을 가능성은 낮은 것으로 판단된다.

: 전체적으로 UP1 기준으로 grouping 한 데이터에서는 1차 spline function의 성능이 더 좋은 것을 확인할 수 있다.

---------- 다음 페이지로 ----------

: 적합 결과를 예측한 결과를 시각화

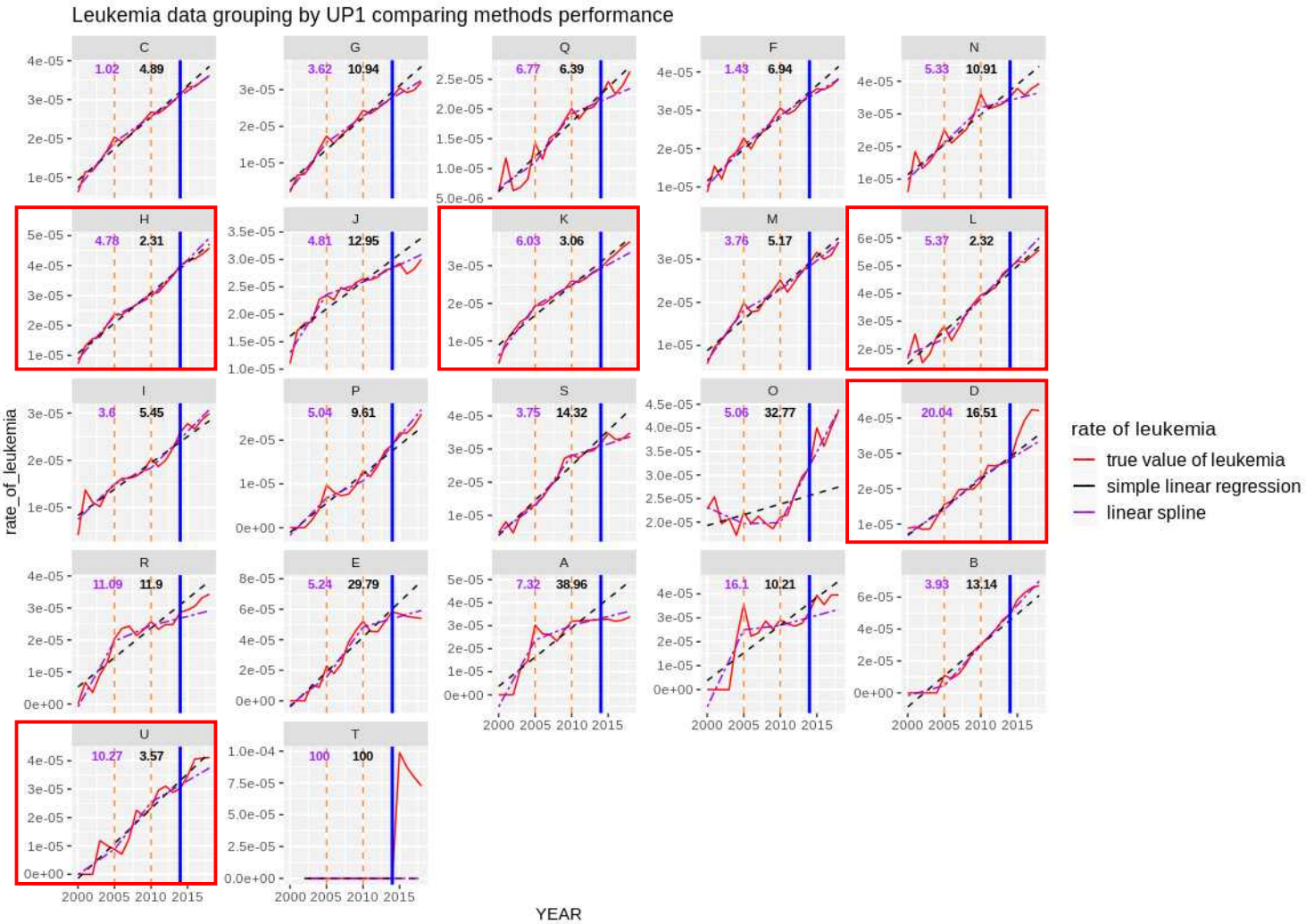
1) Version 1

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP1 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

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



: 1차 spline function보다 단순선형회귀모형이 더 좋은 그래프에 **빨간색 테두리**를 쳐놓았다. 해당 사업장에 대한 정보는 아래 표와 같다.

UP1	사업장명	단순선형회귀모형의 성능이 더 나은 이유 추측
H	운수 및 창고업	2014년 이후 통합 누적 발생률 추세가 살짝 감소하는데, 이에 가깝게 예측한 것이 단순선형회귀모형이다.
K	금융 & 보험업	2014년 이후 통합 누적 발생률 추세가 살짝 증가하는데, 이에 가깝게 예측한 것이 단순선형회귀모형이다.
L	부동산업	2014년 이후 통합 누적 발생률 추세가 살짝 감소하는데, 이에 가깝게 예측한 것이 단순선형회귀모형이다.
D	전기, 가스, 증기 및 공기 조절 공급업	2014년 이후 통합 누적 발생률 추세가 급격히 증가하는데, 이에 가깝게 예측한 것이 단순선형회귀모형이다.
U	국제 및 외국기관	2014년 이후 통합 누적 발생률 추세가 살짝 증가했다가 다시 감소하는데, 이에 가깝게 예측한 것이 단순선형회귀모형이다.

2) Version 2

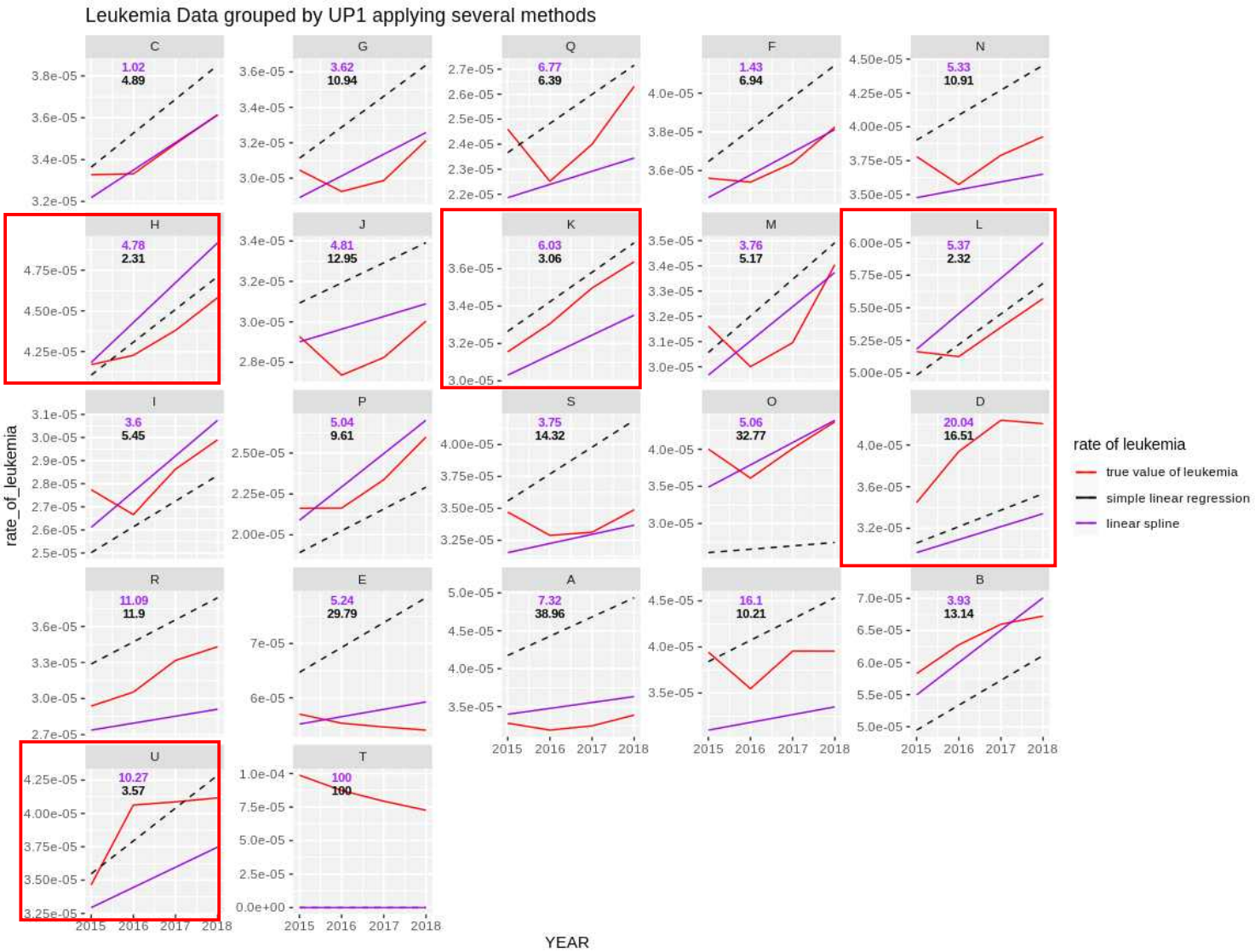
: Version 1의 그래프에서 validation data 시각화 부분이 눈에 띄지 않아, 2015년 ~ 2018년에 대해 각 사업장(UP1)별 실제 통합 누적 발생률과 적합한 모형 통해 예측한 값 plotting

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP1 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

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



: 1차 spline function보다 단순선형회귀모형이 더 좋은 그래프에 **빨간색 테두리**를 쳐놓았다.

---------- 다음 페이지로 ----------

③ Level 3 data

: UP2 별 단순선형회귀모형 적합 후 얻은 MAPE, 1차 spline function 적합 후 얻은 MAPE를 작성한 표이다.

: 표의 순서는 UP2 기준으로 grouping 한 후, 2018년 기준으로 계산한 추적 인년 합계의 순위이다. (높은 순에서 낮은 순으로 정렬)

UP2	단순선형회귀모형 적합 후 얻은 MAPE	1차 spline function 적합 후 얻은 MAPE
46	14.02	3.76
26	6.23	1.45
41	5.94	2
75	5.61	10.51
86	11.83	4.53
49	2.97	3.73
64	2.16	5.73
30	2.48	3.78
29	2.76	3.94
47	5.55	3.65
58	8.5	8.88
68	2.32	5.37
25	5.78	2.97
85	9.61	5.04
10	17.19	10.97
72	15.18	4.01
28	9.34	3.29
87	17.88	25.85
71	10.01	6.11
20	7.14	3.08
13	19.01	9.72
74	15.82	4.03
22	3.57	9.84
52	3.63	8.45
56	5.84	5.12
61	18.06	4.36
31	18.34	6.7
24	17.66	3.6
14	6.91	3.2
23	7.68	10.17
84	32.77	5.06
65	5.77	6.14
35	14.5	19.59
27	22.43	8.01
95	17.07	8.55
55	5.15	5.78
94	14.61	4.94
18	15.57	10.18
66	4.11	8.55
17	7.54	6.71
21	8.76	6.35
91	18.1	10.45
33	11.32	5.47
70	10.01	13.65
15	7.38	4.03
96	6.39	5.38
45	6.61	8.59
63	36.4	18.92
38	40.72	5.16
11	2.94	13.72
32	24.34	5.46
16	7.43	4.06
62	9.42	24.03
60	23.52	22.47
42	33.62	18.24
73	28.41	20.35
51	15.7	13.91
50	11.4	36.66
1	33.89	2.25
59	12.72	68.26
90	3.35	11.65
19	6.51	34.45
34	6.53	23.79
76	14.43	10.34
99	3.57	10.27
7	3	13.48
37	14.45	9.14
5	19.48	9.46
3	48.83	57.82
12	29.04	18.87
36	57.4	31.15
2	79.28	79.96
39	21.76	152.7
6	-99	-99
97	100	100
8	-99	-99
98	-99	-99

: 두 모형 적합 통해 얻은 MAPE 비교 후, **더 낮은 MAPE 값**을 가지는 부분에 **빨간색**으로 표시하였다.

: 전반적으로 1차 spline function의 성능이 더 좋아 보인다.

: 사업장 “6”, “8”, “98”은 MAPE 값이 “NaN”이다.

: 적합 결과를 예측한 결과를 시각화

1) Version 1

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP2 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

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

: 1차 spline function보다 단순선형회귀모형의 성능이 더 좋은 사업장에 대해 **빨간색 테두리**를 그렸다.

: 해당 사업장에 대한 정보를 취합해보면 다음과 같다.

---

1) 대분류 "K"(금융 및 보험업)에 속해있는 중분류 사업장 전부("64", "65", "66"), 대분류 "L"(부동산업)에 속해있는 중분류 사업장 전부("68"), 대분류 "U"(국제 및 외국기관)에 속해있는 중분류 사업장 전부("99")가 단순선형회귀모형 적합 때 좋은 성능을 보인다.

2) 단순선형회귀모형이 더 좋은 성능을 보이는 사업장 중 많은 사업장(24%)을 포함하고 있는 대분류는 "C"(제조업)이다.

: 전반적으로 1차 spline function의 성능이 더 좋은 경향을 보인다.

2) Version 2

: Version 1의 그래프에서 validation data 시각화 부분이 눈에 띄지 않아, 2015년 ~ 2018년에 대해 각 사업장(UP2) 별 실제 통합 누적 발생률과 적합한 모형 통해 예측한 값 plotting

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 **단순선형회귀모형**으로 **예측한 값**이며, **보라색 실선**은 **1차 spline function 적합 통해 예측한 값**을 의미한다.

: **파란색 수직**은 **YEAR=2014**를 의미하는 선이며, **주황색 수직선**은 **YEAR=2005, 2010년** 1차 spline function의 "knots"를 의미한다.

: 그래프의 순서는 2018년 기준 UP2 기준으로 grouping 한 후 계산한 추적 인년 합계의 순위이다. (높은 값에서 낮은 값 순으로 정렬)

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

: 1차 spline function보다 단순선형회귀모형의 성능이 더 좋은 사업장에 대해 **빨간색 테두리**를 그렸다.