산보연 과제 07월 09일 Version - Forecasting

작성자 : 이은경

< What to do >

: 사업장 중분류(UP2)별 설명변수를 YEAR, 반응변수를 질병 통합 누적 발생률로 하여 단순선형회귀모형 적합

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

: train data로 단순선형회귀모형 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)

: 단순선형회귀를 적합한 모형의 performance를 시각적으로 표현하기 위해 plots 생성

[Result 제시]

① Lung Cancer Data

: MAPE를 오름차순으로 정렬

UP2	MAPE
39	65.01
2	47.17
34	44.31
36	26.49
1	23.28
87	20.25
45	19.27
37	18.17
85	17.61
59	16.45
18	16.33
76	15.41
62	12.15
95	11.86
90	11.84
99	11.12
3	10.76
10	10.76
91	10.25
56	10.08
70	10.05
49	9.75
94	9.70
68	9.21
86	8.99
35	8.81
26	8.75
6	**8.61**
50	8.51
55	8.39
19	8.35
58	8.15
73	8.04
41	8.02

UP2	MAPE
75	8.02
46	7.88
71	7.36
65	7.35
24	7.27
30	7.11
21	6.12
64	5.81
96	5.74
38	5.70
13	5.24
29	5.17
61	4.92
84	4.89
51	4.79
63	4.43
60	4.41
14	4.35
22	3.60
66	3.59
7	3.53
15	3.46
23	3.40
31	3.34
27	3.25
12	3.12
5	3.01
32	2.92
47	2.88
74	2.85
72	2.67
17	2.43
42	2.23
25	2.10

UP2	MAPE
33	2.08
52	2.04
16	2.01
20	1.81
11	1.40
8	-99
97	-99
98	-99

: Lung cancer data에서 2000년 ~ 2014년 자료를 가지고 통합 누적 발생률에 대해 YEAR을 설명변수로 하여 단순선형회귀모형을 적합하고, 2015년 ~ 2018년 자료의 통합 누적 발생률을 예측한 결과, 사업장 "39"(환경 정화업)의 MAPE가 65.01로 가장 큰 값을 보인다. 이전, EDA 과정에서 사업장 "39" 외에 사업장 "6"(광업 - 금속, 철, 비금속)이 간접 SIR, 통합 누적 발생률 값이 컸었는데 이 사업장의 MAPE는 8.61로 의외의 결과를 보였다. 그리고 사업장 "8"(광업 - 광업, 자원, 원유), "97", "98"(가구 내 고용활동)은 MAPE가 "NaN" 값을 보이는 것으로 확인되었다.

: 결과를 그래프로 표현

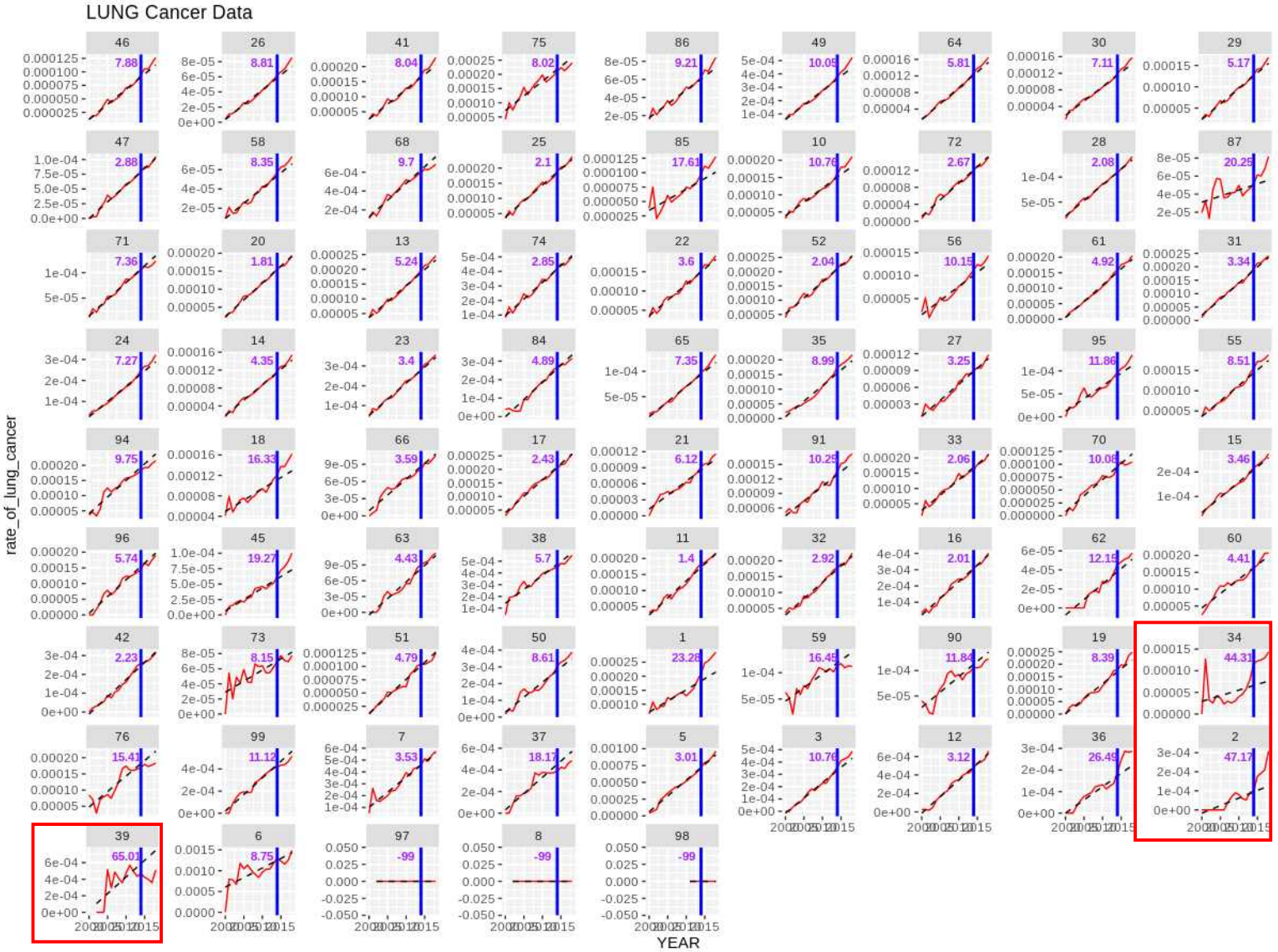
1) Version 1

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 단순선형회귀모형으로 **예측한 값**이다.

: **파란색 수직선**은 YEAR=2014를 의미하는 선이다.

: 각 그래프 안 **보라색 글씨는 각 사업장의 MAPE**를 의미한다.

: 그래프의 순서는 2018년 추적 인년 합계 순위이다. (오름차순 정렬)



: MAPE가 큰 사업장 TOP3를 빨간색 테두리로 표시하였다.

: 해당 사업장의 정보는 아래 표에 작성하였다.

UP2	사업장명	MAPE	MAPE가 큰 이유 추측
39	환경 정화업	65.01	통합 누적 발생률 추세가 peak를 찍고 하락
2	임업	47.17	통합 누적 발생률 추세가 2014년 이후 급격히 증가
34	산업용 기계 및 장비 수리업	44.31	통합 누적 발생률 추세가 2014년 이후 급격히 증가

---------- 다음 페이지로 ----------

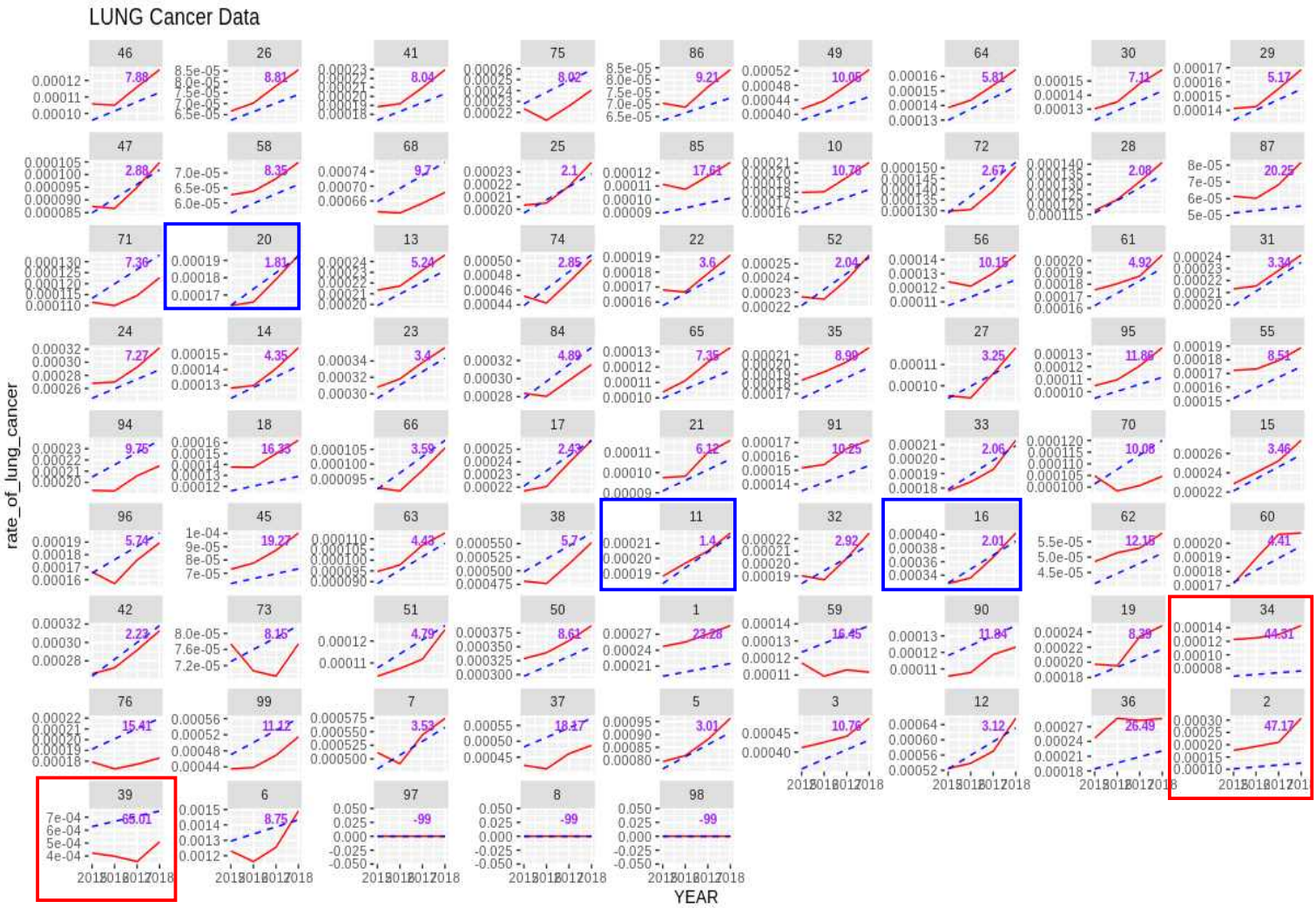
2) Version 2

: 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, **파란색 점선**은 적합한 단순선형회귀모형으로 **예측한 값**이다.

: 각 그래프 안 **보라색 글씨는 각 사업장의 MAPE**를 의미한다.

: 그래프의 순서는 2018년 추적 인년 합계 순위이다. (오름차순 정렬)



: MAPE가 가장 작은 사업장 TOP3는 **파란색** 테두리로, MAPE가 가장 큰 사업장 TOP3는 **빨간색** 테두리로 표시해 놓았다. MAPE가 큰 사업장들의 그래프를 보면, 실제 통합 누적 발생률 변화 추세와 단순선형회귀모형이 예측한 추세와 차이가 많이 나는 것을 볼 수 있다. YEAR 변수만으로 통합 누적 발생률 변화를 설명하긴 어려워 보인다. 해당 사업장에 대한 정보는 다음과 같다.

UP2	사업장명	MAPE
39	환경 정화업	65.01
2	임업	47.17
34	산업용 기계 및 장비 수리업	44.31
16	목재 및 나무 제품 제조업 : 가구 제외	2.01
20	화학물질 및 화학 제품 제조업 : 의약품 제외	1.81
11	음료 제조업	1.4

---------- 다음 페이지로 ----------

[Result 제시]

② Leukemia Data

: MAPE를 오름차순으로 정렬

UP2	MAPE
97	100
2	79.28
36	57.40
3	48.83
38	40.72
63	36.40
1	33.89
42	33.62
84	32.77
12	29.04
73	28.41
32	24.34
60	23.52
27	22.43
39	21.76
5	19.48
13	19.01
31	18.34
91	18.10
61	18.06
87	17.88
24	17.66
10	17.19
95	17.07
74	15.82
51	15.70
18	15.57
72	15.18
94	14.61
35	14.50
37	14.45
76	14.43
46	14.02
59	12.72

UP2	MAPE
86	11.83
50	11.40
33	11.32
70	10.01
71	10.01
85	9.61
62	9.41
28	9.34
21	8.76
58	8.50
23	7.68
17	7.54
16	7.43
15	7.38
20	7.14
14	6.91
45	6.61
34	6.53
19	6.51
96	6.39
26	6.23
41	5.94
56	5.84
25	5.78
65	5.77
75	5.61
47	5.55
55	5.15
66	4.11
52	3.63
22	3.63
99	3.57
90	3.35
7	3.00

UP2	MAPE
49	2.97
11	2.94
29	2.76
30	2.48
68	2.32
64	2.16
6	-99
8	-99
98	-99

: Leukemia Data에 대해 2000년 ~ 2014년 자료를 가지고 통합 누적 발생률에 대해 YEAR을 설명변수로 하여 단순 선형회귀모형을 적합하고, 2015년 ~ 2018년 자료의 통합 누적 발생률을 예측한 결과, 사업장 "97"(가구 내 고용 활동)의 MAPE가 100으로 가장 크게 나왔다. 이 사업장은 이전 EDA 과정에서 간접 SIR과 2018년 기준 백혈병 통합 누적 발생률이 가장 큰 사업장으로 판단되었었다. 반면, 사업장 "6"(광업 - 금속, 철, 비금속) 또한 간접 SIR과 2018년 기준 백혈병 통합 누적 발생률이 2번째로 큰 사업장으로 판단되었었는데, MAPE 값이 "NaN"으로 나와 모형의 적합 정도를 판단하기가 어렵다. 그리고, 사업장 "8"(광업 - 광업, 자원, 원유)과 사업장 "98"(가사 생산 활동)은 폐암 데이터에서와 마찬가지로 MAPE 값이 "NaN"인 것을 확인하였다.

---------- 다음 페이지로 ----------

: 결과를 그래프로 표현

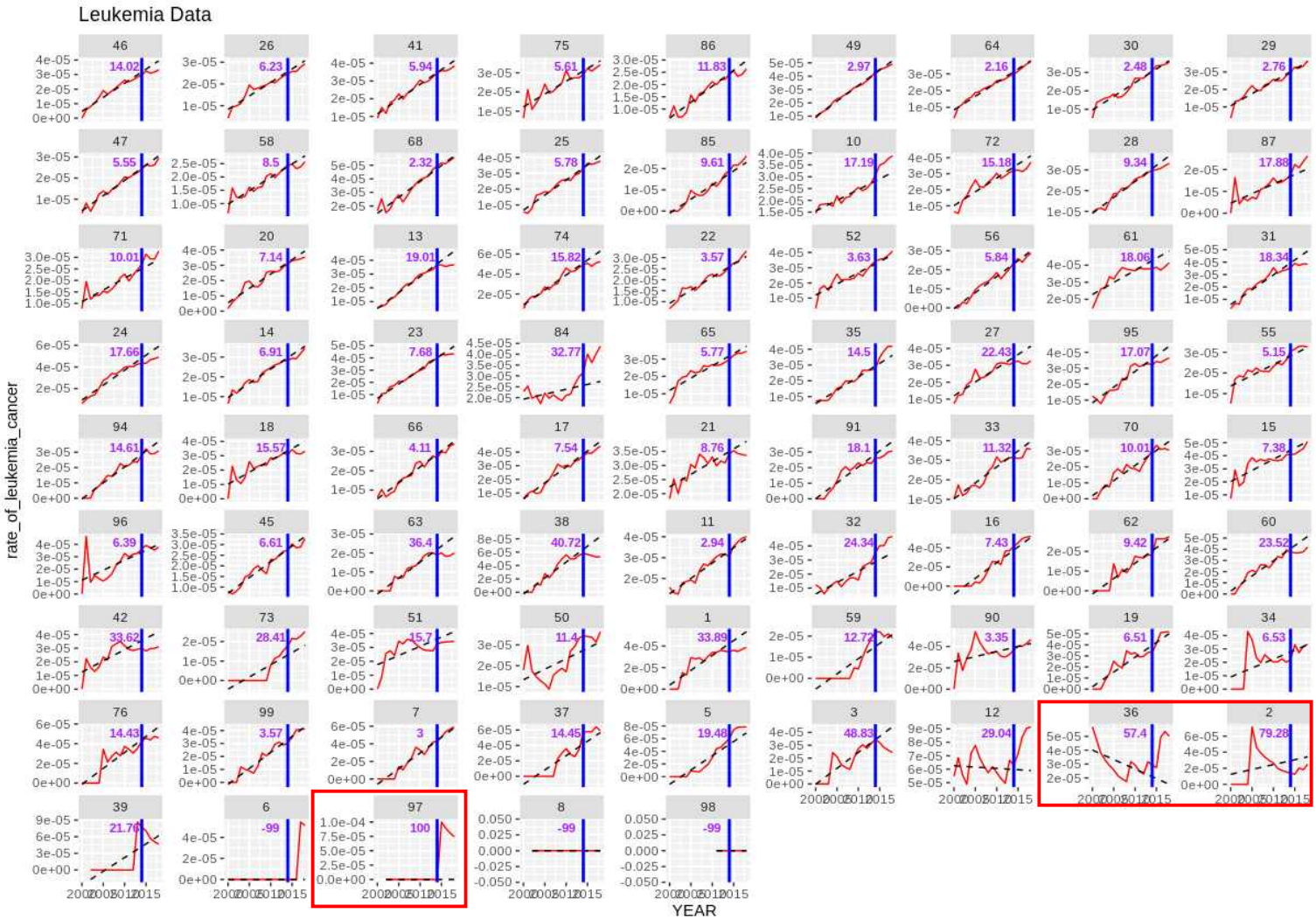
1) Version 1

: **빨간색 실선**이 **실제 통합 누적 발생률 값**이고, **검은색 점선**은 적합한 단순선형회귀모형으로 **예측한 값**이다.

: **파란색 수직선**은 YEAR=2014를 의미하는 선이다.

: 각 그래프 안 **보라색 글씨는 각 사업장의 MAPE**를 의미한다.

: 그래프의 순서는 2018년 추적 인년 합계 순위이다. (오름차순 정렬)



: 이전 EDA 과정에서 통합 누적 발생률의 변화 추세가 불안정하다고 여겼던 사업장들이 예측에서도 마찬가지로 적합한 모형의 성능이 좋지 않는 경향을 보인다.

: MAPE가 큰 사업장 3곳에 대해 빨간색 테두리를 그렸다.

: 해당 사업장의 정보는 아래 표에 제시하였다.

UP2	사업장명	MAPE	MAPE가 큰 이유 추측
97	가구 내 고용활동	100	2014년 이전까지 통합 누적 발생률이 0이었다가 2014년을 기준으로 급격히 증가
2	임업	79.28	단순선형회귀모형으로 peak를 찍고 감소하는 추세 적합 어려움
36	용수 공급업	57.4	2104년 이전까지 통합 누적 발생률이 감소하는 추세였으나, 2014년을 기준으로 통합 누적 발생률이 증가하기 시작함.

---------- 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, **파란색 점선**은 적합한 단순선형회귀모형으로 **예측한 값**이다.

: 각 그래프 안 **보라색 글씨는 각 사업장의 MAPE**를 의미한다.

: 그래프의 순서는 2018년 추적 인년 합계 순위이다. (오름차순 정렬)

: 폐암 데이터와 다르게, 통합 누적 발생률의 추세를 정 반대 방향으로 예측한 경우가 꽤 있다는 점에 주목할 필요가 있다.

: MAPE가 가장 작은 사업장 TOP3는 **파란색** 테두리로, MAPE가 가장 큰 사업장 TOP3는 **빨간색** 테두리로 표시해 놓았다.

: 해당 사업장의 정보는 아래 표와 같다.

UP2	사업장명	MAPE
97	가구 내 고용활동	100
2	임업	79.28
36	용수 공급업	57.4
30	자동차 및 트레일러 제조업	2.48
68	부동산업	2.32
64	금융업	2.16