가설 설정

- **가설 A**
  - H_0(귀무가설): 시도의 노인인구수 대비 요양기관 비율은 전국과 차이가 없다
  - H_1(대립가설): 시도의 노인인구수 대비 요양기관 비율은 전국과 차이가 있다
  - 전국은 모든 시도의 평균 비율을 의미함

  - 전국의 요양 기관을 시도별로 구분하고 해당 지역의 노인인구수를 이용하여 노인인구수 대비 요양기관의 비율을 계산한다.
  - 만약 시도 간 비율 차이가 없다면 현재 요양 서비스 공급의 분배가 적절히 이루어지고 있는 것이다.
  - 하지만, 비율 차이가 유의미하게 나타난다면 특정 지역에 추가적인 자원(예산) 배분이 필요할 가능성이 있다.
  - 가설 A에서는 OLS + Cluster-Robust SE를 이용하여 시도 간 비율 차이가 존재하는지 검정할 계획이다.

- **가설 B**
  - H_0(귀무가설): 시도별 노인인구 대비 요양기관 비율의 변화 추세는 전국 평균 추세와 차이가 없다.
  - H_1(대립가설): 시도별 노인인구 대비 요양기관 비율의 변화 추세는 전국 평균 추세와 차이가 있다.

  - 전국의 노인인구수 대비 요양기관의 비율을 이용해서 연도별 변화를 나타내는 회귀직선을 도출한다.
  - 각 시도별에서도 회귀직선을 도출하여 기울기의 차이를 비교함으로써 크게 변하는 지역을 탐색한다.
  - 가설 B에서는 회귀분석 기반의 비교를 통해 특이한 변화 패턴을 보이는 지역을 확인할 계획이다.

데이터 분석 방법 소개

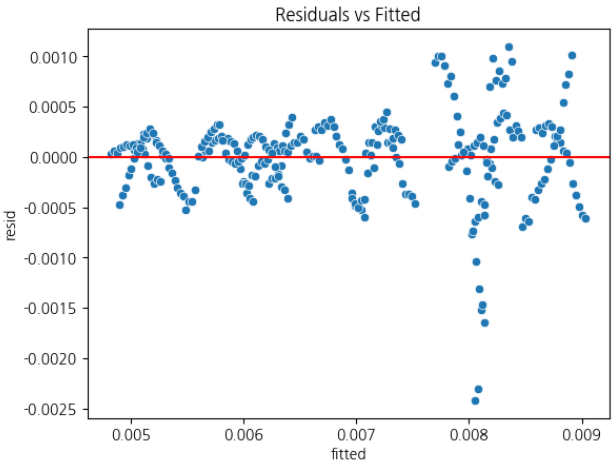
- **OLS + Cluster-Robust SE - 가설 A에 대한 분석 방법**
  - OLS + Cluster-Robust SE란 무엇인가?
    - OLS(최소제곱법)
      - 회귀모형에서 계수를 추정하는 가장 기본적인 방법
      - 변수 간의 선형관계를 이용해 종속변수의 변화를 설명하는 통계 모델
    - Cluster-Robust SE(클러스터-로버스트 표준오차)
      - 단순 OLS의 표준오차(SE)가 독립성, 정규성, 등분산성 등의 가정을 위반할 때 표준오차를 더 신뢰성 있게 계산하는 방법
    - 시도를 클러스터 단위로 지정하여 같은 시도 내 연도들의 상관관계를 보정하는 구조
  - OLS + Cluster-Robust SE를 선택한 이유
    - 단순 ANOVA
      - 기본 데이터의 경우 n=1인 17개의 그룹에 대한 데이터라서 분산분석이 불가능
      - 전처리한 데이터의 경우 그룹 내의 데이터가 독립적이지 않으므로 사용 불가능
    - 혼합설계 ANOVA
      - 두 종류의 요인을 동시에 다루는 분산분석
      - between-subject 요인(시도) + within-subject 요인(년도)에 대해 분석
      - 단, 혼합설계 ANOVA의 가정 중 정규성 가정이 충족되지 않아 적용이 불가능
  - OLS + Cluster-Robust SE의 주요 가정
    - 선형성
      - 독립변수와 종속변수의 관계가 선형
      - 잔차 vs 적합값 플롯을 통해 확인
        - 잔차가 0을 기준으로 무작위로 분포하는지
    - 외생성
      - 독립변수가 오차항과 상관되지 않아야 함
        - 독립변수에 어떤 값이 들어가더라도 평균 오차는 0이어야 함
      - 잔차 + 년도 / 잔차 + 시도명 플롯을 통해 확인
        - 특정 연도에서만 잔차가 계속 양수/음수인지
        - 특정 시도에서만 잔차가 한 방향으로 치우쳐져 있는지

- 다중공선성
  - 독립변수들끼리 서로 강하게 상관되어 있어서 각 변수의 개별 효과를 구분하기 어려운 상태인지 아닌지 확인
    - 완전한 다중공선성이 존재하면 OLS 추정 불가
    - 다중공선성 자체는 OLS의 주요 가정이 아니지만, 계수의 불완정성을 유발할 수 있기 때문에 VIF로 점검
  - VIF는 다른 독립변수들과의 상관 때문에 해당 변수의 분산이 얼마나 부풀려졌는지를 나타내는 지표
    - VIF 5 미만인지 확인할 것
- 오차의 등분산성
  - 기본 OLS SE에서는 이걸 가정해야 하지만, Cluster-Robust SE를 사용하므로 쓸 필요 없음
- 클러스터 간 독립성
  - 클러스터 간 오차항은 독립 -> 가정하고 진행
  - 같은 시도 내에서는 독립일 필요 없음
- 클러스터 수가 충분히 많아야 함
  - 이론적으로 30개 이상이어야 함
  - 경제학, 사회학에서는 10~20개도 허용

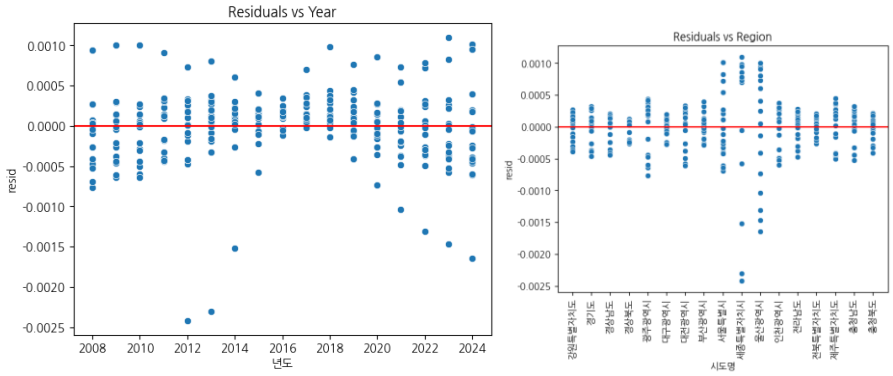
- **회귀 분석 - 가설 B에 대한 검정 방법**
  - 연도에 따른 노인인구 대비 요양기관 비율의 변화를 분석하기 위해 연도를 독립변수(설명변수), 비율을 종속변수(반응변수)로 하는 단순 선형 회귀모형을 적합
    - 전국의 비율을 이용해 회귀모형을 적합하여 전체의 평균적인 비율의 변화를 추정
    - 각 시도별 비율을 이용해 회귀모형을 적합하여 시도별 비율의 변화를 추정
    - 전국 회귀모형의 기울기와 시도별 회귀모형의 기울기를 z-test로 비교하여 변화 속도가 크게 높거나/낮은 지역을 확인한다.
      - z-test: 두 값의 차이가 통계적으로 유의한지 판단하기 위한 z-통계량 기반의 검정
      - z-통계량: 차이를 표준 오차를 이용해 표준화한 값
  - 회귀 분석에서의 주요 가정
    - 불편성: 설명변수가 주어졌을 때 오차항들의 기댓값은 0
      - 설명변수가 년도인 단순회귀에서는 관측 오류나 내생성 문제가 사실상 없어서 자동으로 충족됨
    - 등분산성: 설명변수가 주어졌을 때 오차항들의 분산이 일정
      - 단일 설명변수 + 길지 않은 시계열에서는 영향이 미미
    - 독립성: 오차항들은 서로 독립이어야 함
      - 가정하고 진행
    - 정규성: 오차항들은 정규분포를 따름
      - z-test를 사용하기 때문에 정규성 가정이 거의 필요 없음
      - 표본 크기가 시도당 15년 정도면 근사적으로 성립
    - 위의 4가지 가정은 현재 시행할 예정인 기울기 비교에서는 반드시 확인해야 할 필수 가정이 아니기 때문에 생략함

데이터 분석

- **가설 A**
  - OLS + Cluster-Robust SE의 주요 가정
    - 선형성



    -
      - 잔차-적합값 플롯에서 일부 관측치가 약간 튀는 부분은 있으나 전반적으로 잔차가 0을 중심으로 일정 범위 내에서 무작위적으로 분포하고 있어 선형성을 충족한다고 판단함
    - 외생성



    -
      - 잔차-연도 플롯과 잔차-시도명 플롯을 확인한 결과
        - 특정 연도에서 잔차가 지속적으로 양수, 음수로 치우치는 패턴이 없고
        - 특정 시도에서만 잔차가 한 방향으로 편중되는 양상도 나타나지 않으며
        - 전체적으로 잔차가 구조적 패턴 없이 분포함
        - 따라서 독립변수와 오차항이 상관되어 있다는 증거가 없어 외생성 가정을 충족한다고 판단함

◆ 다중공선성

```
===== VIF 결과 =====
       variable            VIF
0         const  172641.646971
1            년도       1.007396
2       시도명_경기도       1.880702
3      시도명_경상남도       1.880702
4      시도명_경상북도       1.880702
5     시도명_광주광역시       1.880702
6     시도명_대구광역시       1.880702
7     시도명_대전광역시       1.880702
8     시도명_부산광역시       1.880702
9     시도명_서울특별시       1.880702
10  시도명_세종특별자치시       1.691607
11    시도명_울산광역시       1.880702
12    시도명_인천광역시       1.880702
13     시도명_전라남도       1.880702
14  시도명_전북특별자치도       1.880702
15  시도명_제주특별자치도       1.880702
16     시도명_충청남도       1.880702
17     시도명_충청북도       1.880702
```

- 년도 변수와 시도명 변수들에 대한 VIF가 모두 5 미만으로 나타나 회귀 계수 추정이 안정적임을 확인하였음

◆ 클러스터 간 독립성

- 클러스터 간에 독립성이 있다고 가정하고 진행함

◆ 클러스터 수가 충분히 많아야 함

- 17개의 그룹으로 사회학 분야에서 허용하는 수치

■ OLS + Cluster-Robust SE 진행 결과

◆ 가설

- H_0(귀무가설): 시도의 노인인구수 대비 요양기관 비율은 전국과 차이가 없다
- H_1(대립가설): 시도의 노인인구수 대비 요양기관 비율은 전국과 차이가 있다

◆ 분석 결과

-

```
===== OLS + Cluster Robust SE =====
                            OLS Regression Results
==============================================================================
Dep. Variable:                   rate   R-squared:                       0.892
Model:                            OLS   Adj. R-squared:                  0.885
Method:                 Least Squares   F-statistic:                     752.8
Date:                Sat, 29 Nov 2025   Prob (F-statistic):          6.98e-15
Time:                        17:02:34   Log-Likelihood:                 1802.8
No. Observations:                 285   AIC:                            -3570.
Df Residuals:                     267   BIC:                            -3504.
Df Model:                          17
Covariance Type:              cluster
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       -0.0486      0.038     -1.287      0.217      -0.129       0.031
C(시도명, Sum)[S.강원특별자치도]    -0.0016    2.2e-06   -740.896      0.000      -0.002      -0.002
C(시도명, Sum)[S.경기도]          0.0004    2.2e-06    189.844      0.000       0.000       0.000
C(시도명, Sum)[S.경상남도]        -0.0010    2.2e-06   -464.495      0.000      -0.001      -0.001
C(시도명, Sum)[S.경상북도]        -0.0018    2.2e-06   -834.780      0.000      -0.002      -0.002
C(시도명, Sum)[S.광주광역시]       0.0014    2.2e-06    618.938      0.000       0.001       0.001
C(시도명, Sum)[S.대구광역시]       0.0012    2.2e-06    525.540      0.000       0.001       0.001
C(시도명, Sum)[S.대전광역시]       0.0019    2.2e-06    875.474      0.000       0.002       0.002
C(시도명, Sum)[S.부산광역시]      -0.0007    2.2e-06   -305.133      0.000      -0.001      -0.001
C(시도명, Sum)[S.서울특별시]       0.0018    2.2e-06    822.017      0.000       0.002       0.002
C(시도명, Sum)[S.세종특별자치시]     0.0013   3.53e-05     36.205      0.000       0.001       0.001
C(시도명, Sum)[S.울산광역시]       0.0010    2.2e-06    470.997      0.000       0.001       0.001
C(시도명, Sum)[S.인천광역시]    -3.106e-05    2.2e-06    -14.091      0.000   -3.57e-05   -2.64e-05
C(시도명, Sum)[S.전라남도]        -0.0018    2.2e-06   -802.537      0.000      -0.002      -0.002
C(시도명, Sum)[S.전북특별자치도]    -0.0004    2.2e-06   -198.024      0.000      -0.000      -0.000
C(시도명, Sum)[S.제주특별자치도]     0.0003    2.2e-06    135.816      0.000       0.000       0.000
C(시도명, Sum)[S.충청남도]        -0.0012    2.2e-06   -534.197      0.000      -0.001      -0.001
년도                          2.752e-05   1.87e-05      1.469      0.161   -1.22e-05    6.72e-05
==============================================================================
Omnibus:                      108.096   Durbin-Watson:                   0.449
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              648.662
Skew:                          -1.411   Prob(JB):                    1.40e-141
Kurtosis:                       9.831   Cond. No.                     8.38e+05
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The condition number is large, 8.38e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

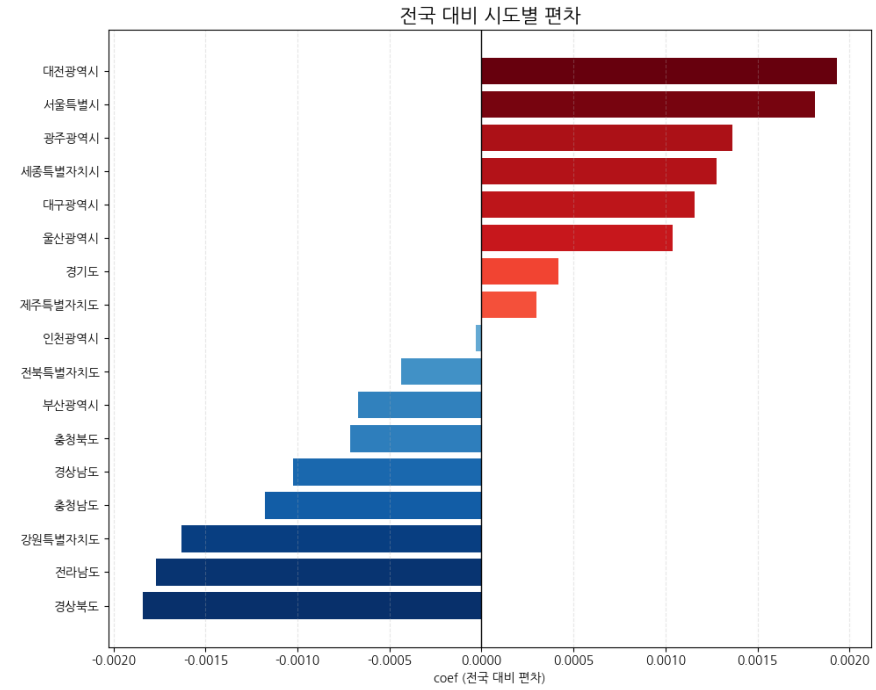
- 해석하기 어렵고 17번째 시도인 충청북도가 사라지기 때문에 그것까지 출력한 테이블 생성

```
===== 전국 기준 시도별 Effect Coding 계수 + p-value (17개 완성) =====
        시도명       coef        p-value
0       경상북도 -0.001840  4.438050e-46
1       전라남도 -0.001769  1.111805e-43
2   강원특별자치도 -0.001633  4.341595e-39
3       충청남도 -0.001177  4.806469e-24
4       경상남도 -0.001024  2.608999e-19
5       충청북도 -0.000714  7.993526e-11
6      부산광역시 -0.000673  7.606833e-10
7   전북특별자치도 -0.000436  4.604276e-05
8      인천광역시 -0.000031  7.683652e-01
9   제주특별자치도  0.000299  4.836063e-03
10        경기도  0.000418  9.177405e-05
11     울산광역시  0.001038  9.696633e-20
12     대구광역시  0.001158  1.922066e-23
13  세종특별자치시  0.001277  2.858486e-22
14     광주광역시  0.001364  4.320977e-30
15     서울특별시  0.001812  3.945798e-45
16     대전광역시  0.001930  4.261925e-49
```

- 모든 시도에서 p-value가 0.05 미만이기 때문에 모든 시도는 전국과 비교할 때 비율이 통계적으로 유의미한 차이가 있다고 판단할 수 있다.
- 이때 17번째 시도인 충청북도의 경우 p-value를 Wald test를 이용해 따로 계산하였으며 이로 인해 얻을 수 있는 p-value가 가지는 의미는 다른 p-value가 가지는 의미와 동일하다.
  - Wald test: 회귀모형의 계수들이 0인지 확인하는 검정

- ◆ Sum Coding에서는 모든 시도 계수의 합 = 0 제약 때문에 16개의 시도만 출력되고 1개의 시도는 누락됨
- ◆ 모든 시도 계수의 합 = 0을 이용해서 누락된 계수를 계산
- ◆ 계수를 계산한 후 Wald test를 적용하여 그 계수가 0인지 검정

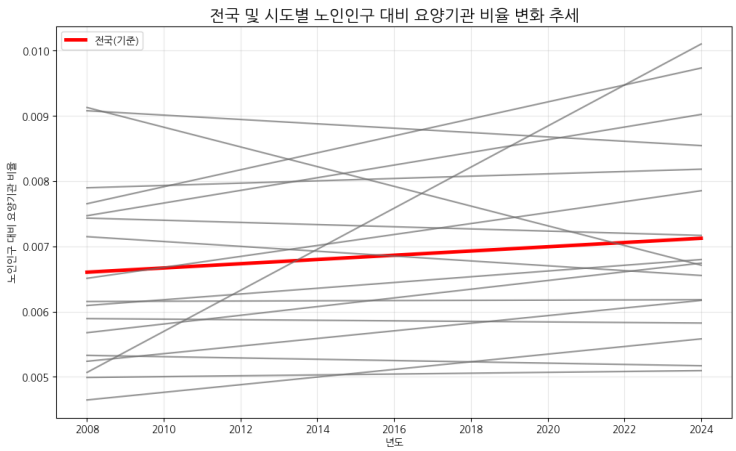
◆ 결과 시각화



- ●
- ● 전국 대비 시도별 편차를 정렬하여 만든 그래프
- ● 대전광역시, 서울특별시는 다른 시도에 비해 유독 더 높은 비율을 보였으며
- ● 경상북도, 전라남도, 강원특별자치도는 다른 시도에 비해 유도 더 낮은 비율을 보였다.

- **가설 B**
  - 가설
    - H_0(귀무가설): 시도별 노인인구 대비 요양기관 비율의 변화 추세는 전국 평균 추세와 차이가 없다.
    - H_1(대립가설): 시도별 노인인구 대비 요양기관 비율의 변화 추세는 전국 평균 추세와 차이가 있다.
  - 회귀모형

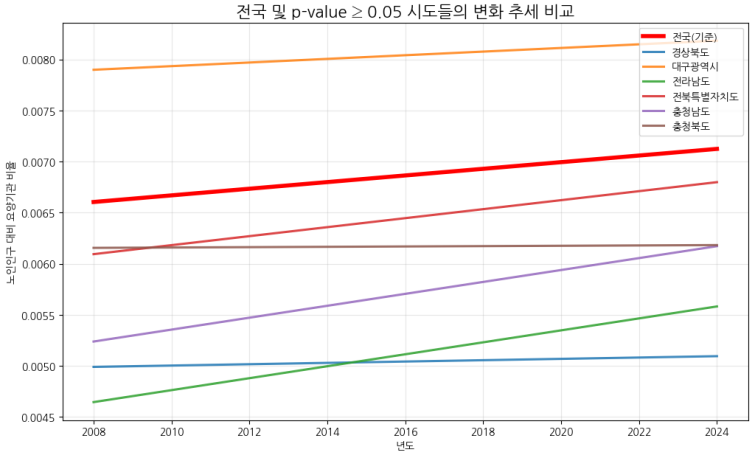
	시도명	beta_r	se_r
0	전국	0.000032	0.000016
1	강원특별자치도	-0.000010	0.000006
2	경기도	-0.000017	0.000007
3	경상남도	-0.000004	0.000008
4	경상북도	0.000007	0.000005
5	광주광역시	0.000097	0.000012
6	대구광역시	0.000018	0.000006
7	대전광역시	-0.000033	0.000006
8	부산광역시	0.000067	0.000003
9	서울특별시	0.000130	0.000006
10	세종특별자치시	0.000315	0.000049
11	울산광역시	-0.000151	0.000012
12	인천광역시	-0.000037	0.000007
13	전라남도	0.000059	0.000008
14	전북특별자치도	0.000044	0.000006
15	제주특별자치도	0.000084	0.000009
16	충청남도	0.000058	0.000012
17	충청북도	0.000002	0.000008



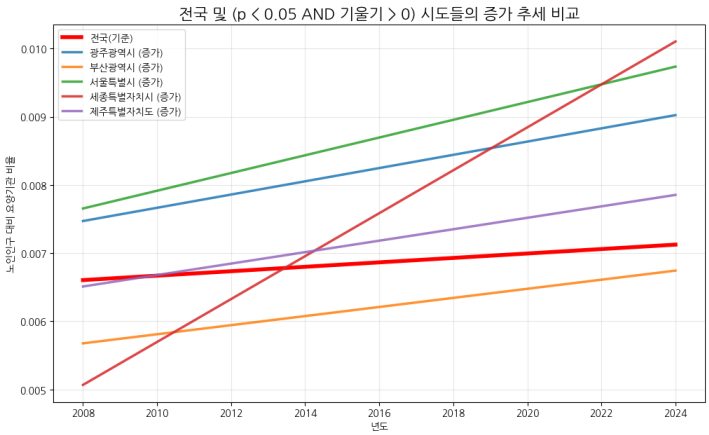
    - $\widehat{rate} = beta_r * year + intercept$
  - 분석 결과

	시도명	beta_r	se_r	z-stat	p-value	p-value < 0.05
0	강원특별자치도	-0.000010	0.000006	-2.509398	1.209370e-02	True
1	경기도	-0.000017	0.000007	-2.783976	5.369700e-03	True
2	경상남도	-0.000004	0.000008	-2.054440	3.993308e-02	True
3	경상북도	0.000007	0.000005	-1.558529	1.191079e-01	False
4	광주광역시	0.000097	0.000012	3.267031	1.086817e-03	True
5	대구광역시	0.000018	0.000006	-0.859220	3.902192e-01	False
6	대전광역시	-0.000033	0.000006	-3.823150	1.317574e-04	True
7	부산광역시	0.000067	0.000003	2.112905	3.460893e-02	True
8	서울특별시	0.000130	0.000006	5.770191	7.918151e-09	True
9	세종특별자치시	0.000315	0.000049	5.467657	4.560244e-08	True
10	울산광역시	-0.000151	0.000012	-9.284564	0.000000e+00	True
11	인천광역시	-0.000037	0.000007	-4.000600	6.318210e-05	True
12	전라남도	0.000059	0.000008	1.453827	1.459942e-01	False
13	전북특별자치도	0.000044	0.000006	0.676613	4.986513e-01	False
14	제주특별자치도	0.000084	0.000009	2.813976	4.893286e-03	True
15	충청남도	0.000058	0.000012	1.300058	1.935812e-01	False
16	충청북도	0.000002	0.000008	-1.739735	8.190556e-02	False

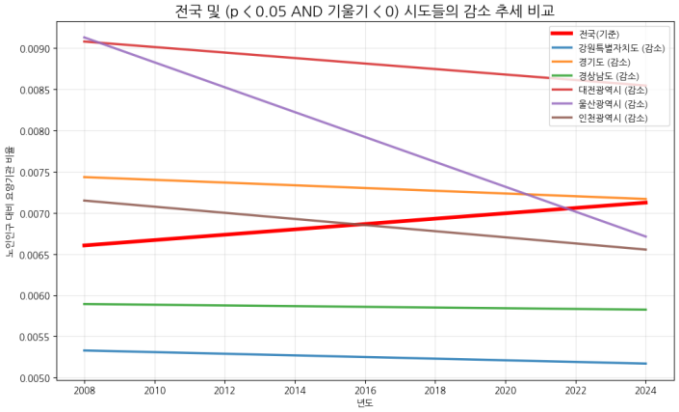
    - 시도별 회귀직선의 기울기와 전국 회귀직선의 기울기 사이 z-test 수행
    - 수행결과



- 시도의 비율 변화 추세는 전국의 비율 변화 추세와 차이가 없는 지역은 경상북도, 대구광역시, 전라남도, 전북특별자치도, 충청남도, 충청북도이다.
  - 전국은 상승하는 추세
  - 대구광역시와 전북특별자치도, 충청남도, 전라남도는 비슷하게 상승하는 추세
  - 경상북도와 충청북도는 거의 유지하는 추세
- 차이가 있는 지역은 강원특별자치도, 경기도, 경상남도, 광주광역시, 대전광역시, 부산광역시, 서울특별시, 세종특별자치시, 울산광역시, 인천광역시, 제주특별자치도이다.



  - 유의미한 차이가 있고 증가하는 추세인 지역은 광주광역시, 부산광역시, 서울특별시, 세종특별자치시, 제주특별자치도
    - 그중에서도 서울특별시와 세종특별자치시가 크게 증가하고 있음을 확인할 수 있음



■

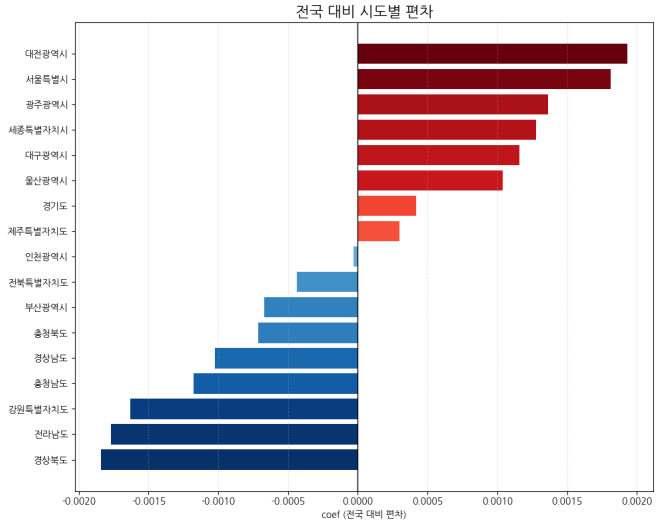
■ 유의미한 차이가 있고 감소하는 추세인 지역은 강원특별자치도, 경기도, 경상남도, 대전광역시, 울산광역시, 인천광역시이다.

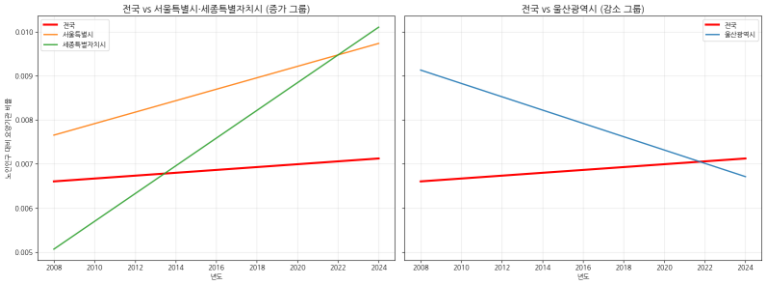
◆ 그중에서도 울산광역시는 크게 감소하고 있음을 확인할 수 있다.

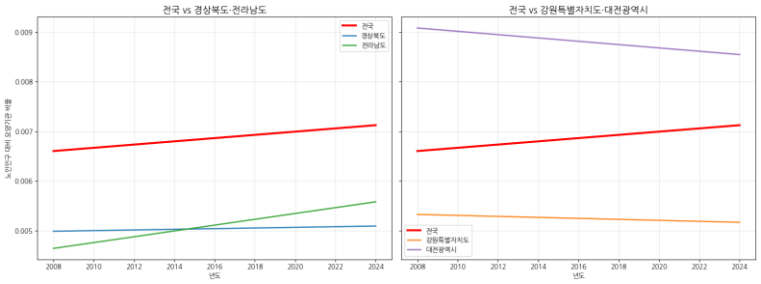
■ 분석 결과 종합

◆ 17개의 시도 중 전국의 변화 비율과 차이를 보이는 시도는 총 11개였으며 그중 5개의 상승하는 지역 중에서는 세종특별자치시와 서울특별시가 큰 차이를 보였으며 6개의 하락하는 지역 중에서는 울산광역시가 크게 감소하고 있다.

데이터 분석 결과

■ 

■ 

■ 

■

■ 종합 분석 결과

- ◆ **서울특별시의 경우 높은 편차를 보이며 비율이 상승하고 있음**
- ◆ **세종특별자치시의 경우 높은 편차를 보이며 비율이 급격히 상승하고 있음**
- ◆ 울산광역시의 경우 높은 편차를 보이지만 비율이 급격히 하락하고 있음
- ◆ **경상북도의 경우 제일 낮은 편차를 보이며 비율이 유지되고 있음**
- ◆ 전라남도의 경우 낮은 편차를 보이며 비율이 조금씩 상승하고 있음
- ◆ 대전광역시는 제일 높은 편차를 보이며 비율은 조금씩 떨어지고 있음
- ◆ **강원특별자치도는 낮은 편차를 보이는데 조금씩 감소하는 모습을 보임**

■ 정리 및 결론

- ◆ 서울특별시, 세종특별자치시는 평균 대비 높은 편차를 보이면서 비율이 상승하고 있음
- ◆ 경상북도는 평균 대비 제일 낮은 편차를 보이면서 비율이 유지되고 있음

◆ 강원특별자치도는 낮은 편차를 보이는데 감소하는 모습을 보임