

# 경영경제데이터분석

## Team Project 분석 보고서

담뱃값 인상에 따른 흡연율 차이 분석 및 흡연율, 담배 판매량 예측

5조



**KYUNG HEE**  
UNIVERSITY



이름	2019100873 박건우
	2019100890 이승재
	2019100895 이윤서
	2020103907 강민영
	2020103910 권수영
	2021103961 곽동엽

# 목차

## 1. 서론

## 2. 분석 기법 소개

### 2.1 RDD

### 2.2 DID

### 2.3 SARIMA

## 3. 담뭇값 인상에 따른 흡연을 차이 분석 (단기적 관점)

### 3.1 인과성 분석(RDD) 개요

### 3.2 분석 준비

#### 3.2.1 데이터 준비

#### 3.2.2 흡연을 데이터, 담배 판매량 데이터로 대체

#### 3.2.3 RDD 가정 성립 확인

### 3.3 RDD 분석

#### 3.3.1 대역폭 및 모형 선택

#### 3.3.2 RDD 추정 및 결과 해석

## 4. 담뭇값 인상에 따른 흡연을 차이 분석 (장기적 관점)

### 4.1 인과성 분석(DID) 개요

### 4.2 데이터 준비

### 4.3 병행추세 가정 검증

### 4.4 분석 및 결과 해석

## 5. 담배 판매량 변화 예측

### 5.1 배경

### 5.2 데이터 준비

### 5.3 시계열 성분 분해

### 5.4 단위근 검정

#### 5.4.1 DF 검정

#### 5.4.2 ADF 검정

### 5.5 ACF, PACF를 통한 AR, MA 모수 추정

5.5.1 원본 시계열 데이터 ACF 확인

5.5.2 계절성 차분 이후, ACF 관찰

5.5.3 계절성 차분 + 일반 차분 이후 ACF, PACF 관찰

5.6 SARIMA 모형 설계

5.6.1 사용자 정의 모델 설계

5.6.2 auto.arima를 활용한 자동 모델 설계 (Seasonal=TRUE)

5.7 2024년 담배 판매량 예측

6. 종합 결론

7. 참고문헌

8. 부록

## 1. 서론

흡연은 전 세계적으로 주요한 공중 보건 문제 중 하나로, 흡연에 따른 건강상의 위험은 개인뿐만 아니라 사회 전체에 심각한 영향을 미친다. 폐암, 심혈관 질환, 만성 폐쇄성 폐질환 등의 각종 질병이 흡연과 직간접적으로 관련이 있으며, 이러한 건강 문제로 인한 경제적 손실 또한 막대하다. 이에 따라 각국 정부는 흡연율을 줄이기 위한 다양한 정책을 시행해 왔다. 그 중에서도 담뱃값 인상은 가장 일반적이고 효과적인 정책 중 하나로 평가받고 있다. 담뱃값 인상을 통해 담배 구매력을 낮추고, 흡연율을 감소시키려는 시도는 많은 연구에서 그 효과를 입증한 바 있다. 한국 정부 역시 이러한 글로벌 흐름에 따라 2015년 1월, 담뱃값을 대폭 인상하는 정책을 시행하였다. 이는 담배 소비를 줄이고 국민 건강을 증진시키기 위한 중대한 조치였다.

본 프로젝트는 이와 같은 맥락에서 담뱃값 인상 정책이 실제로 흡연율에 단기적, 장기적으로 어떠한 영향을 미쳤는지를 분석하기 위해 RDD(Regression Discontinuity Design) 기법과 DID(Difference in Differences) 기법을 활용하여 담뱃값 인상 전후의 흡연율 변화를 정밀하게 평가하고, 정책의 인과적 효과를 확인하고자 한다.

나아가 이러한 분석 결과를 바탕으로 시계열 분석을 진행하고자 한다. 2024년 1월, 총선 후 기존 한 갑에 4500원이던 담뱃값이 8000원, 1만원으로 인상될 수도 있다는 내용의 기사가 기재되었다. 우리 조는 2023년까지의 담배 판매량 데이터를 바탕으로 시계열 분석을 진행하여 2024년 담배 판매량을 예측하고 이때 실제적인 흡연율 감소 효과가 있는지를 분석하고자 한다.

본 프로젝트의 결과는 흡연율 감소를 위한 정책적 함의를 제공하는 데 기여할 것으로 기대된다. 특히, 담뱃값 인상과 같은 가격 정책이 실제로 흡연율을 낮추는 데 효과적이지에 대한 실증적 근거를 제시함으로써, 향후 정책 결정 과정에 중요한 참고자료가 될 수 있을 것이다. 또한, 향후 담배 판매량 예측을 통해 보건 정책의 방향성을 설정하는 데 유용한 정보를 제공할 수 있을 것이다.

## 2. 분석 기법 소개

### 2.1 RDD

Regression Discontinuity Design (RDD)는 특정 기준값을 중심으로 데이터의 변화를 분석하여 인과적 효과를 추정하는 기법이다. RDD는 정책 시행 전후의 데이터를 비교함으로써, 정책의 인과적 영향을 평가하는 데 유용하다. 본 프로젝트에서는 2015년 담뱃값 인상을 기준으로, 그 이전과 이후의 월간 담배 판매량 변화를 분석하여 담뱃값 인상이 흡연율에 미친 영향을 평가하였다. 이를 통해 담뱃값 인상의 단기적 효과를 명확히 파악하고자 한다.

### 2.2 DID

Difference-in-Differences (DID) 분석은 시간에 따른 처치군과 통제군의 결과 차이를 비교하여 인과적 효과를 추정하는 기법이다. DID는 자연 실험 환경에서 정책 효과를 평가하는 데 널리 사용된다. 본 프로젝트에서는 DID 분석을 통해 담뱃값 인상 전후의 흡연율 변화를 평가하였다. 처

치군으로는 담뱃값 인상을 시행한 한국을, 통제군으로는 같은 시기 담뱃값 변화가 없었던 일본을 설정하였다. 이를 통해 담뱃값 인상이 흡연율에 미친 장기적 효과를 비교 분석하고자 한다.

## 2.3 SARIMA

Seasonal Autoregressive Integrated Moving Average (SARIMA) 모델은 계절적 변동을 포함한 시계열 데이터 분석 및 예측 기법이다. SARIMA는 시계열 데이터의 추세, 계절성, 불규칙성을 모두 반영하여 미래 값을 예측하는 데 유용하다. 본 프로젝트에서는 SARIMA 모델을 사용하여 2024년 담배 판매량을 예측하였다.

## 3. 담뱃값 인상에 따른 흡연율 차이 분석 (단기적 관점)

### 3.1 인과성 분석(RDD) 개요

본 연구는 2015년 담뱃값 인상이 한국 성인 흡연율에 미친 단기적 영향을 분석하는 것을 목적으로 한다. 이를 위해 RDD 기법을 활용하여 담뱃값 인상 정책의 인과적 효과를 평가하고자 한다.

2015년 1월 1일, 한국 정부는 담배 가격을 기존 2500원에서 4500원으로 대폭 인상하는 정책을 시행하였다. 이 정책은 흡연율 감소를 목표로 하였으며, 그 효과를 평가하기 위해 담뱃값 인상 전후의 데이터를 분석할 필요가 있다.

본 연구에서는 다음과 같은 요소를 중심으로 RDD 분석을 수행한다:

- ✓ 처치: 2015년 담뱃값 인상
- ✓ 할당 변수: 연월 (2013년 7월부터 2014년 6월까지의 월별 데이터)
- ✓ 결과 변수: 월간 담배 판매량
- ✓ 임계값: 2015년 (담배 가격이 2500원에서 4500원으로 인상된 시점)

RDD 분석에서는 할당 변수를 기준으로 데이터를 나누어, 임계값 전후의 결과 변수를 비교함으로써 정책의 인과적 효과를 평가한다. 본 연구의 경우, 2015년을 기준으로 그 이전과 이후의 월간 담배 판매량 변화를 분석하게 된다. 이를 통해 담뱃값 인상이 실제로 흡연율에 어떠한 영향을 미쳤는지를 명확하게 파악할 수 있다.

### 3.2 분석 준비

#### 3.2.1 데이터 준비

2015년 담뱃값 인상이 한국 성인 흡연율에 미친 영향을 분석하기 위해 한국담배협회에서 제공하는 '2012년 1월 ~ 2015년 7월까지 월별 담배 판매량'과 기획재정부에서 발표한 '2022년 상반기 담배 시장 동향' 데이터를 취합하여 활용하였다. 해당 데이터는 연구 목적에 맞게 다음과 같은 변수로 구성되어 있다:

- ✓ month: 연월 (개월 기준 소수점 표기)
- ✓ increase: 담뱃값 인상 연월
- ✓ month\_d: 연월 - 담뱃값 인상 연월

- ✓ treat: 담뱃값 인상 여부
- ✓ smoke: 담배 판매량 (단위: 억 갑)

분석에 사용된 데이터의 수집 기간은 2012년 1월부터 2016년 12월까지로 설정하였다. 이 기간을 설정한 이유는 다음과 같다. 2017년 5월 궐련형 전자담배가 출시되었고, 2016년 말에는 경고그림과 문구 제도가 본격적으로 시행되었다. 이러한 요인들은 담뱃값 인상 외에도 흡연율 변화에 영향을 미칠 수 있기 때문에, 담뱃값 인상만으로 인한 흡연율 변화를 명확히 분석하기 위해 2017년 이후의 데이터는 제외하였다.

### 3.2.2 흡연율 데이터, 담배 판매량 데이터로 대체

담뱃값 인상의 단기적 영향을 분석하기 위해 월간 데이터를 확보하는 것이 적합하다고 판단하였다. 그러나 흡연율 데이터는 연 단위로 조사되기 때문에 월 단위 데이터를 확보할 수 없었다. 이에 따라, 월 단위 데이터가 존재하는 담배 판매량 데이터를 대신 사용하기로 하였다.

연율과 담배 판매량 데이터의 대체 가능성을 검증하기 위해, 연도별 현재 흡연율과 연도별 담배 판매량 합계, 그리고 연도별 매일 흡연율과 연도별 담배 판매량 합계 간 상관계수를 분석하였다. 이를 통해 담배 판매량이 흡연율을 대체할 수 있다는 타당성을 검토하였으며, 그 결과는 다음과 같다:

- ✓ 현재흡연율과 담배판매량 합계 간의 상관계수: 0.7906
- ✓ 매일흡연율과 담배판매량 합계 간의 상관계수: 0.8944

분석 결과, 현재흡연율과 담배 판매량 합계 간의 상관계수는 0.7906으로 나타났으며, 매일흡연율과 담배 판매량 합계 간의 상관계수는 0.8944로 나타났다. 이는 담배 판매량이 흡연율을 상당 부분 반영할 수 있음을 시사한다.

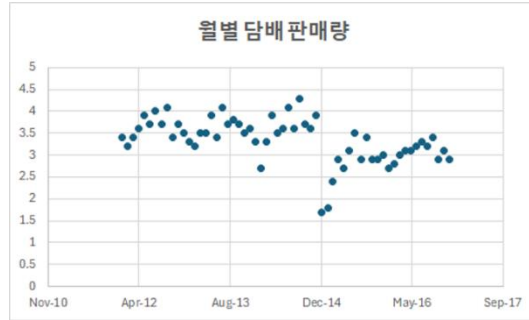
또한, '담뱃세 인상정책의 흡연억제 효과 - 시계열자료를 통한 실증분석'에서도 담배 판매량을 통해 흡연이 감소했는지를 확인하였다. 이러한 선행연구는 담배 판매량 데이터가 흡연율을 대체하여 사용할 수 있음을 뒷받침하는 근거로 작용한다.

### 3.2.3 RDD 가정 성립 확인

RDD 분석을 수행하기 위해서는 몇 가지 중요한 가정이 성립해야 한다. 담뱃값 인상이 흡연율에 미친 영향을 분석하기 위해 다음과 같은 RDD 가정 성립 여부를 확인하였다.

#### 1) 할당 변수와 결과 변수의 연속성:

할당 변수인 연월과 결과 변수인 월간 담배 판매량이 임계값을 기준으로 연속적인 분포를 가져야 한다. 이를 확인하기 위해 그래프를 통해 임계값(2015년)을 중심으로 한 월간 담배 판매량의 분포를 시각적으로 분석하였다. 아래 그래프는 임계값을 기준으로 자연스럽고 연속적인 분포를 보임을 확인할 수 있다.



## 2) 조작 가능성의 부재:

담뱃값 인상 정책은 정부에 의해 결정되고 예고되었기 때문에 정책 시행 시점이 외부에 의해 조작될 수 없다.

## 3) 임계값 주변의 동질성:

- ✓ 흡연자 특성: 담뱃값 인상 직전과 직후의 흡연자 집단은 연령, 성별, 소득 수준 등 주요 인구통계학적 특성이나 흡연 습관에서 큰 차이가 없을 것으로 예상된다.
- ✓ 외부 요인: 담뱃값 인상 시점 전후에 흡연율에 영향을 미칠 수 있는 다른 정책 변화나 사회적 이슈가 없었다.

위와 같은 가정 성립을 통해, RDD 분석의 타당성을 확보할 수 있었다.

# 3.3 RDD 분석

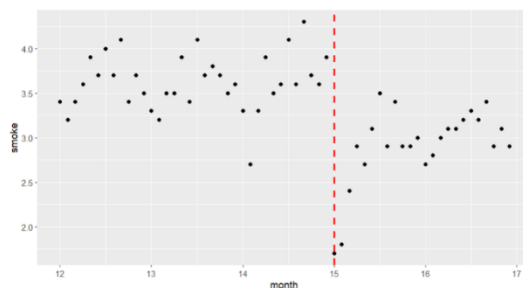
## 3.3.1 대역폭 및 모형 선택

대역폭 설정을 위해 다음과 같은 선행 연구와 시각화 결과를 확인하였다.

### 1) 선행연구 확인:

- ✓ '담배가격 인상에 따른 청소년 흡연율 변화', 민경문 외 2인(2018): 2013년~2015년까지의 데이터를 분석.
- ✓ '담배가격 인상에 따른 청소년의 흡연 및 금연 시도 관련요인 분석', 김환희(2017): 2014~2017년까지의 데이터를 분석.

### 2) 시각화 결과 확인



<2012년 1월~2016년 12월까지의 결과 시각화>

선행 연구와 시각화 결과를 기반으로, 본 연구에서는 6개월, 1년, 1년 6개월로 대역폭을 설정하여 분석을 진행하기로 결정하였다.

모형 선택의 경우, 각각 RDD 회귀 모형을 선형, 2차항 비선형 모형, 3차항 비선형 모형으로 분석하기로 결정하였다. 이러한 모형을 선택한 근거는 다양한 형태의 데이터 분포를 고려하여 최적의 모델을 찾기 위함이다. 선형 모형은 단순하지만 직관적인 해석이 가능하며, 비선형 모형은 데이터의 복잡한 패턴을 포착할 수 있어 다양한 형태의 변화에 유연하게 대응할 수 있다.

### 3.3.2 RDD 추정 및 결과 해석

- 대역폭: 6개월 & 모형: RDD 회귀모형 추정 결과

Variables	Coefficient	Std.Error	P-value
Intercept	3.7067	0.2332	0.000
Treat	-2.0102	0.2891	0.000
Month_d	-0.5486	0.7187	0.465
Treat : Month_d	4.1057	0.9162	0.002

→ 담뱃값 인상이라는 처치 효과의 추정치가 유의함, LATE = -2.0102

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 2.01억갑 감소했음을 나타낸다.

- 대역폭: 6개월 & 모형: 2차항 비선형 RDD 회귀모형 추정 결과

Variables	Coefficient	Std.Error	P-value
Intercept	3.8400	0.5004	0.000
Treat	-2.1971	0.5568	0.006
Month_d	0.6514	3.9284	0.873
Month_d^2	2.0571	6.5924	0.764
Treat : Month_d	3.6771	4.5457	0.445
Treat : Month_d^2	-3.6000	7.9231	0.663

→ 담뱃값 인상이라는 처치 효과의 추정치가 1% 유의수준에서 유의함, LATE = -2.1971

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 2.2억갑 감소했음을 나타낸다.

- 대역폭: 6개월 & 모형: 3차항 비선형 RDD 회귀모형

Variables	Coefficient	Std.Error	P-value
Intercept	4.16667	1.17406	0.016
Treat	-2.55714	1.21527	0.089
Month_d	5.58254	16.05275	0.742
Month_d^2	21.65714	61.64165	0.740
Month_d^3	22.40000	69.89952	0.762
Treat : Month_d	0.07937	17.11885	0.997



Treat : Month_d^2	-30.40000	68.19362	0.674
Treat : Month_d^3	-12.80000	79.69771	0.879

→ 담뱃값 인상이라는 처치 효과의 추정치가 10% 유의수준에서 유의함, LATE = -2.55714

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 2.56억갑 감소했음을 나타낸다.

- 대역폭: 1년 & 모형: RDD 회귀모형

Variables	Coefficient	Std.Error	P-value
Intercept	4.0864	0.2448	0.000
Treat	-1.8259	0.3216	0.000
Month_d	0.8517	0.3992	0.045
Treat : Month_d	0.1504	0.5335	0.781

→ 담뱃값 인상이라는 처치 효과의 추정치가 유의함, LATE = -1.8259

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 1.83억갑 감소했음을 나타낸다.

- 대역폭: 1년 & 모형: 2차항 비선형 RDD 회귀모형

Variables	Coefficient	Std.Error	P-value
Intercept	3.6750	0.2871	0.000
Treat	-2.0387	0.3497	0.000
Month_d	-1.2638	1.2186	0.313
Month_d^2	-1.9528	1.0950	0.091
Treat : Month_d	6.3515	1.5315	0.001
Treat : Month_d^2	-2.1327	1.4137	0.148

→ 담뱃값 인상이라는 처치 효과의 추정치가 유의함, LATE = -2.0387

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 2.04억갑 감소했음을 나타낸다.

- 대역폭: 1년 & 모형: 3차항 비선형 RDD 회귀모형

Variables	Coefficient	Std.Error	P-value
Intercept	3.5566	0.4747	0.000
Treat	-1.9953	0.5354	0.002
Month_d	-2.3623	3.6425	0.525
Month_d^2	-4.3892	7.6559	0.574
Month_d^3	-1.4993	4.6585	0.751

Treat : Month_d	8.5813	4.2734	0.061
Treat : Month_d^2	-2.6418	9.3247	0.780
Treat : Month_d^3	3.4629	5.8231	0.560

→ 담뱃값 인상이라는 처치 효과의 추정치가 1% 유의수준에서 유의함, LATE = -1.9953

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 2억갑 감소했음을 나타낸다.

- 대역폭: 1년 6개월 & 모형: RDD 회귀모형

Variables	Coefficient	Std.Error	P-value
Intercept	3.7523	0.1835	0.000
Treat	-1.3518	0.2466	0.000
Month_d	0.1152	0.2035	0.575
Treat : Month_d	0.5017	0.2767	0.079

→ 담뱃값 인상이라는 처치 효과의 추정치가 유의함, LATE = -1.3518

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 1.35억갑 감소했음을 나타낸다.

- 대역폭: 1년 6개월 & 모형: 2차항 비선형 RDD 회귀모형

Variables	Coefficient	Std.Error	P-value
Intercept	4.1853	0.2625	0.000
Treat	-2.1566	0.3335	0.000
Month_d	1.6740	0.7634	0.036
Month_d^2	0.9845	0.4685	0.044
Treat : Month_d	0.5175	0.9934	0.606
Treat : Month_d^2	-2.0343	0.6219	0.003

→ 담뱃값 인상이라는 처치 효과의 추정치가 유의함, LATE = -2.1566

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 2.16억갑 감소했음을 나타낸다.

- 대역폭: 1년 6개월 & 모형: 3차항 비선형 RDD 회귀모형

Variables	Coefficient	Std.Error	P-value
Intercept	3.6534	0.3016	0.000
Treat	-2.0929	0.3592	0.000
Month_d	-1.8851	1.6052	0.250
Month_d^2	-4.4861	2.3228	0.063

Month_d^3	-2.3034	0.9661	0.024
Treat : Month_d	8.4192	1.9790	0.000
Treat : Month_d^2	-4.0003	2.9524	0.186
Treat : Month_d^3	5.6086	1.2528	0.000

→ 담뱃값 인상이라는 처치 효과의 추정치가 유의함, LATE = -2.0929

분석 결과, 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 미치는 영향이 통계적으로 유의미하며, 담뱃값 인상 이후 담배 판매량이 평균적으로 약 2.09억갑 감소했음을 나타낸다.

RDD 분석 결과, 2015년 담뱃값 인상 직후 담배 판매량에 유의미한 변화가 나타났다. 분석에 사용된 대역폭은 6개월, 1년, 1년 6개월이었으며 각각 RDD 회귀모형, 2차항 비선형 RDD 회귀모형, 3차항 비선형 RDD 회귀모형을 사용하여 분석했다. 충분한 관측치가 확보되고 단절 패턴을 명확히 확인할 수 있는 '1년 6개월'의 대역폭과 시각적으로 확인했을 때 데이터의 경향을 잘 반영한 것으로 보이는 '3차항 비선형 RDD 회귀모형'을 선택하였다. 이 때 추정된 LATE 값은 -2.0929로 2015년이라는 컷오프 근처에서 담뱃값 인상이라는 처치가 담배 판매량(흡연율)에 감소하는 방향으로 유의미한 영향을 미친 것으로 나타났다.

## 4. 담뱃값 인상에 따른 흡연율 차이 분석 (장기적 관점)

### 4.1 인과성 분석(DID) 개요

2017년 8월 방영된 '썰전' 230회에서 <담배과세의 효과와 재정><sup>1</sup>을 인용하여 담배 가격에 대한 추가 세수가 2000원 이상이면 세수가 오히려 줄어든다는 주장을 하였다. 지난 2년 간 담배 과세를 증세한 후 1년이 지나면 줄었던 흡연자 수가 원상 복귀되고, 이에 따라 세수가 증가하는 과정이 반복되면서 흡연율 감소라는 본래 목적은 사라지고 증세 효과만 남는 것이 증명되었다.

RDD 기법을 사용한 분석 결과, 단기적인 처치 효과는 분명히 있었으나, 위와 같은 주장이 존재하기 때문에 장기적으로도 처치 효과가 충분했는지 여부를 검증할 필요가 있다. 이에 본 연구에서는 DID (Difference-in-Differences) 기법을 사용하여 담뱃값 인상의 장기적 효과를 평가하고자 한다.

### 4.2 데이터 준비

분석에 사용할 데이터는 OECD Statistics 2023에서 제공하는 나라별 매일 흡연자 비율 데이터를 활용하였다. RDD 분석에서 증명되었듯이, 현재 흡연율과 매일 흡연율 지표는 큰 상관관계를 보이므로 두 나라의 데이터가 모두 존재하는 매일 흡연율 데이터를 대체 데이터로 활용하였다.

분석 대상 데이터는 담뱃값 인상 처치 시점인 2015년 1월 1일을 기준으로 이전 4년(2011~2014)와 이후 4년(2015~2018)로 설정하였다. 또한, 2010년과 2018년에 일본에서 담배 세율에 변

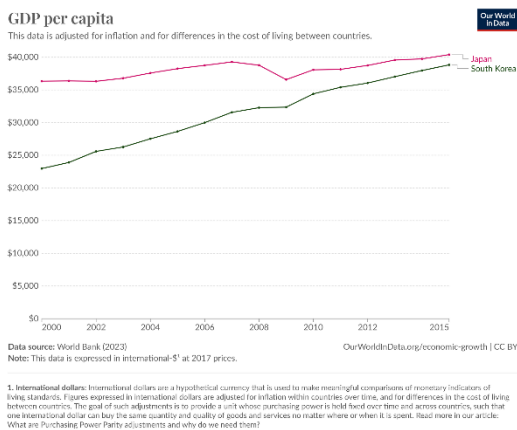
<sup>1</sup> <https://www.kipf.re.kr/kor/Publication/All/kiPublish/CC/view.do?serialNo=522984>

화가 있었기 때문에, 이로 인한 흡연율 변화를 배제하여 순수한 처치 효과를 파악하고자 한다.

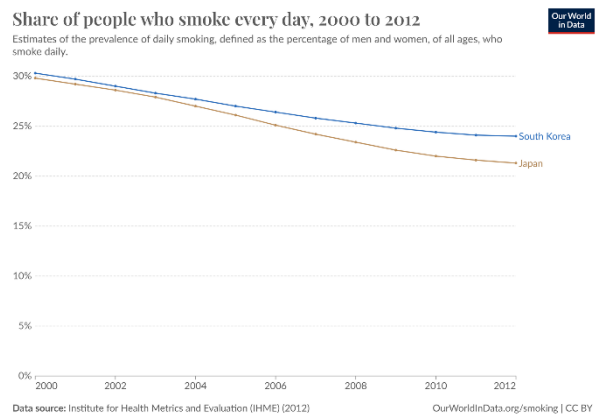
### 4.3 병행추세 가정 검증

처치군을 한국의 연간 매일 흡연율로 설정하였고, 통제군은 일본의 연간 매일 흡연율로 설정하였다. 일본과 한국은 지리적, 문화적으로 유사성을 보이며, 거의 비슷한 시기에 국제담배규제협약에 가입하여 담배 규제 정도에서도 비슷한 양상을 보인다.

처치 이전의 결과 변수들의 변화 추이를 확인한 결과, 두 나라는 병행추세를 보인다고 가정할 수 있었다. 다음은 이러한 가정을 뒷받침하는 그래프들이다.



<2000~2015년의 GDP 변화 추이>



<2000~2012년의 매일 흡연율 변화 추이>

### 4.4 분석 및 결과 해석

Variables	Coefficient	Std.Error	P-value
Intercept	749.3	743.1	0.335
국가	0.9563	1.462	0.526
연도	-0.3595	0.3804	0.365
처치 후	-1.487	1.037	0.179
GDP	-0.0001541	0.0006039	0.803

분석 결과, 국가, 연도, 처치 후, GDP 변수의 p-value가 모두 0.05보다 크므로, 이 변수들은 개별적으로 통계적으로 유의하지 않은 것을 확인할 수 있다. 이는 분석 진행 전에 예상한 대로 국가나 연도에 따른 흡연율 차이가 유의미하게 벌어지지 않았음을 의미한다.

중요한 것은 처치 후 변수로, DID Estimator가 음수 값을 가지지만 p-value가 유의하지 않았으므로 해당 처치가 흡연율 감소를 일으켰을 수는 있으나 '선행' 230회 영상에서 제시된 주장대로 장기적 처치 효과는 통계적으로 유의하지 않았음을 확인할 수 있다.

따라서, 차후 유의미한 흡연율 감소를 발생시키려면 최대 세수보다는 조금 더 담뱃값을 증가시켜 흡연자 수가 원상 복구되지 않을 정도의 인상을 일으켜야 한다는 결론을 도출할 수 있다.

단, 위 분석에도 한계점이 존재하는데, 분석에 사용된 데이터 표본 수 자체가 매우 적은 편이

고, 이로 인해 처치 효과를 유의미하게 포착하지 못했을 수 있다는 점이다.

## 5. 담배 판매량 변화 예측

### 5.1 배경

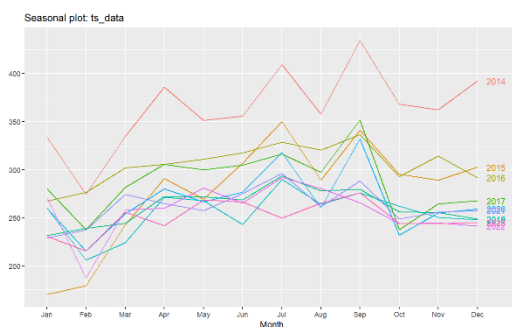
2024년 1월, 총선 후 기존 한 갑에 4500원이던 담뱃값이 8000원, 1만원으로 인상될 수도 있다는 내용의 기사가 게재되었다. 총선이 지난 현재 기사에서 언급되었던 담뱃값 인상 정책은 시행되지 않았지만, 우리 조는 이러한 시나리오를 가정하여 만약 실제로 총선 후 담뱃값이 8000원으로 인상되었다면, 담배 판매량이 어떻게 변화했을지를 예측하고자 한다.

RDD 분석 결과와 DID 분석 결과에서 도출된 단기적 및 장기적 처치 효과를 기반으로, 담뱃값 인상이 흡연율과 담배 판매량에 미치는 영향을 평가하였다. 이러한 분석 결과를 토대로 시계열 예측 기법을 적용하여 향후 담배 판매량 변화를 예측함으로써, 담뱃값 인상이 소비자 행동과 시장에 미치는 잠재적 영향을 구체적으로 파악하고자 한다.

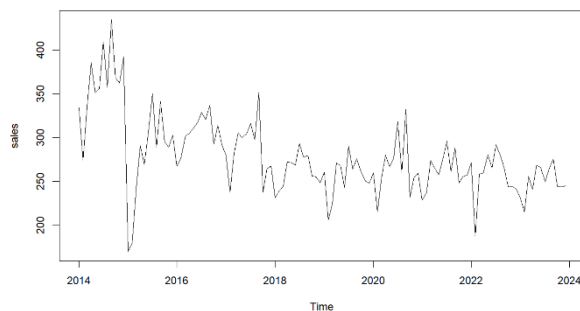
### 5.2 데이터 준비

시계열 분석에 사용할 데이터는 기획재정부 보도자료인 '2023년 담배시장 동향'에서 제공된 2014년 1월부터 2023년 12월까지의 담배 판매량 데이터를 활용하였다.

데이터를 시각화한 결과는 다음과 같다.



<월에 따른 연도별 담배 판매량>



<연도별 담배 판매량 그래프>

데이터를 살펴보면 다음과 같은 점을 알 수 있다:

- ✓ 신년 다짐으로 인해 연초에 담배 판매량이 하락하는 경향이 있다.
- ✓ 2015년 담뱃값 인상 당시, 담배 판매량이 급락한 것을 확인할 수 있다.

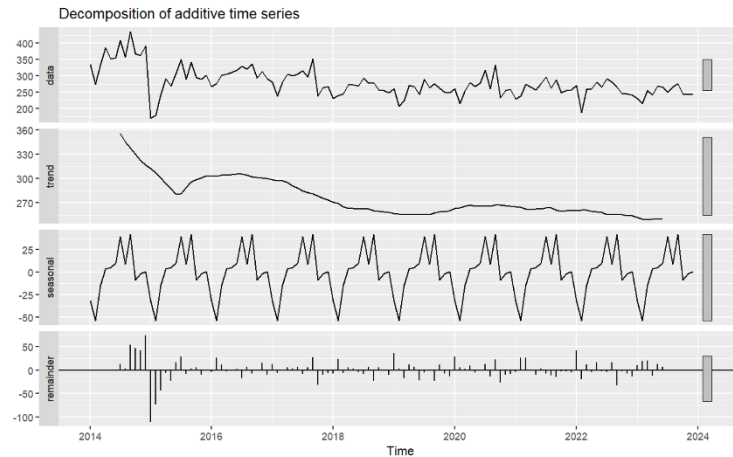
### 5.3 시계열 성분 분해

담배 판매량 데이터의 계절성, 추세, 불규칙성을 분리함으로써 각 요인의 영향을 독립적으로 분석하여 담뱃값 인상이 흡연율과 판매량에 미친 영향을 명확히 파악하기 위해 시계열 성분 분해를 진행하였다.

우선적으로 모형 선택을 진행하였다. 담배 판매량 데이터를 시각화한 결과, 담배 판매량이 비교

적 일정한 변동성 수준을 보이기에, 가법 모형이 적합하다고 판단하여 가법 모형을 선택하였다.

가법 모형 선택 후, 시계열 성분 분해 단계에 맞게 계절성분, 추세성분, 나머지 불규칙 변동성분 순서로 시계열 성분 분해를 진행하였다. 분해된 시계열 성분을 시각화한 결과는 다음과 같다.



시계열 성분 분해 결과 다음과 같은 점을 알 수 있다:

- ✓ 1년 주기의 뚜렷한 계절성을 확인할 수 있다.
- ✓ 2015년 담뱃값 4500원 인상 당시 급격한 감소 이후, 약간의 증가가 나타났다.
- ✓ 전반적으로 감소하는 추세를 확인할 수 있었다.

## 5.4 단위근 검정

### 5.4.1 DF 검정

시계열 분석에서 데이터가 정상성을 갖는지 여부는 예측 모델의 정확성에 있어 매우 중요하다. 따라서 데이터의 정상성을 검토하기 위해, DF(Dickey-Fuller) 검정을 진행하였다. 이는 시계열 데이터에 단위근이 존재하는지 여부를 확인하여, 데이터가 안정적인지 판단하기 위한 방법이다.

시계열 성분 분석을 통해 전체적으로 감소하는 결정적인 추세를 중심으로 서서히 변한다고 판단하였기에 Random walk with trend (and drift) (상수항 0, 추세 0) 모형을 선택하여 단위근 검정을 수행하였다.

단위근 검정 결과는 다음과 같다:

1) DF 통계량  $r = -6.8618$

- ✓ 절대값 기준 1% 유의 수준 임계값보다 큼
- ✓ 단위근이 존재하지 않음 → 정상 시계열

2) drift 통계량 = 15.7175

- ✓ 절대값 기준 1% 유의 수준 임계값보다 큼
- ✓ 단위근이 존재하지 않음 → 정상 시계열

3) trend = 23.5472

- ✓ 절대값 기준 1% 유의 수준 임계값보다 큼
- ✓ 단위근이 존재하지 않음 → 정상 시계열

추세와 상수항을 가정했을 때, DF 검정 결과 월별 담배 판매량 데이터는 정상 시계열 데이터임을 알 수 있었다. 이는 시계열 모델이 적합하게 예측을 수행할 수 있음을 의미한다.

#### 5.4.2 ADF 검정

앞서 진행한 DF 검정은 오차항 간 상관관계가 없다고 가정하고 있으나, 시계열 데이터의 특성상 오차항 간 상관관계가 존재할 수 있다. 이에 대응하기 위해 ADF (Augmented Dickey-Fuller) 검정을 진행하였다. ADF 검정은 DF 검정의 확장된 형태로, 오차항 간의 상관관계를 고려하여 더 신뢰성 있는 단위근 검정을 제공한다.

단위근 검정 결과는 다음과 같다:

1) DF 통계량  $r = -4.9117$

- ✓ 절대값 기준 1% 유의 수준 임계값보다 큼
- ✓ 단위근 존재하지 않음 → 정상 시계열

2) drift 통계량 = 8.069

- ✓ 절대값 기준 1% 유의 수준 임계값보다 큼
- ✓ 단위근 존재하지 않음 → 정상 시계열

3) trend = 12.0725

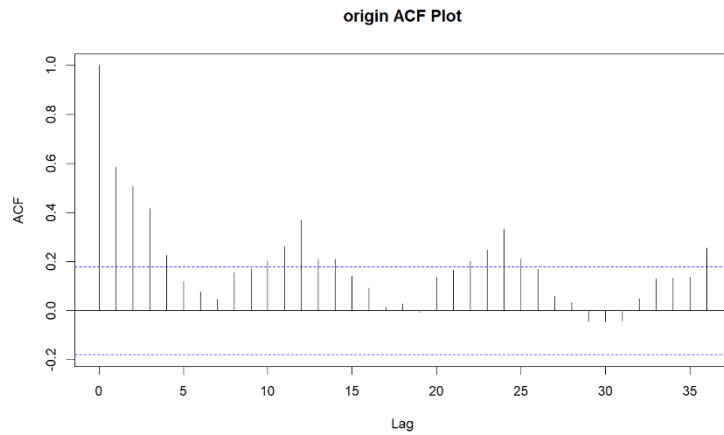
- ✓ 절대값 기준 1% 유의 수준 임계값보다 큼
- ✓ 단위근 존재하지 않음 → 정상 시계열

추세와 상수항을 가정했을 때, ADF 검정 결과 월별 담배 판매량 데이터는 정상 시계열 데이터임을 알 수 있었다. 이는 시계열 모델이 적합하게 예측을 수행할 수 있음을 의미한다. ADF 검정을 통해 오차항 간의 상관관계를 고려한 결과, 데이터가 정상성을 유지하고 있음을 확인하였다.

### 5.5 ACF, PACF를 통한 AR, MA 모수 추정

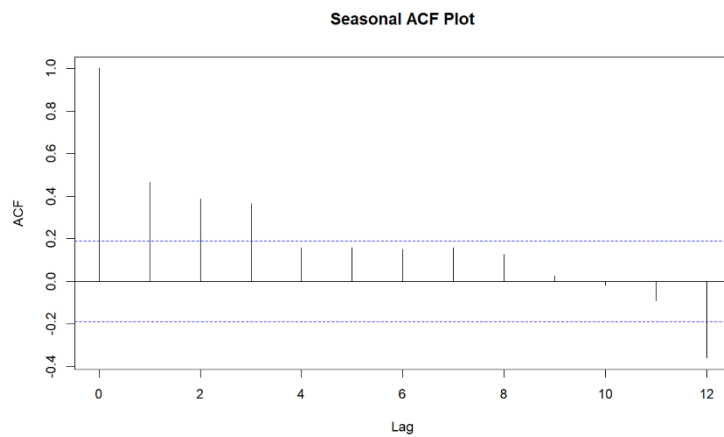
ACF(자기상관함수)와 PACF(부분자기상관함수)는 시계열 분석에서 중요한 도구로, 각각 시계열 데이터의 자기상관성을 측정하는 데 사용된다. ACF는 시차에 따른 데이터 간의 상관관계를 나타내며, PACF는 중간에 다른 시차의 영향을 제거한 후 특정 시차 간의 상관관계를 나타낸다. 이들은 AR 및 MA 등 시계열 모형의 모델링 과정에서 몇 시점 전까지의 관측치 혹은 오차를 모형에 포함할지를 결정할 때 도움이 된다.

#### 5.5.1 원본 시계열 데이터 ACF 확인



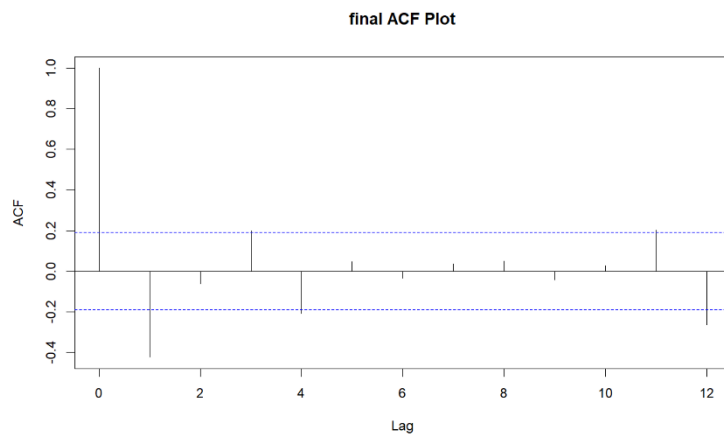
→ 원본 시계열 데이터 ACF 확인 결과, 뚜렷한 자기 상관 계절성이 확인되어 계절성 차분(12)을 수행하였다.

### 5.5.2 계절성 차분 이후, ACF 관찰



→ 원본 시계열 데이터에 대해 계절성 차분 수행 후, ACF 해석 결과 천천히 줄어드는 패턴인 것을 확인되었다. 이를 근거로 비정상성 데이터로 판단, 1차 차분 수행하였다.

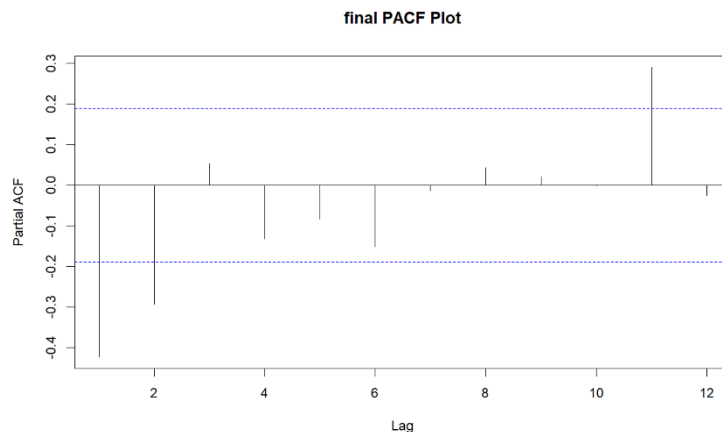
### 5.5.3 계절성 차분 + 일반 차분 이후 ACF, PACF 관찰



→ ACF plot에서 lag 1 이후 ACF 값이 급격히 감소되는 것이 관찰되었다. 이를 근거로 데이터의



정상성을 확보했으며, 자기회귀 차수 값( $p=0$ )을 구했다.



→ PACF plot에서 lag 2 이후 PACF 값이 급격히 감소되는 것이 관찰되었다. 이를 근거로 데이터의 정상성을 확보했으며, 이동평균 차수 값( $q=1$ )을 구했다.

## 5.6 SARIMA 모형 설계

### 5.6.1 사용자 정의 모델 설계

AR, MA 모수 추정 과정에서 구한 값들을 토대로 사용자 정의 모델을 구성하였다. 비계절성 모수는 앞에서 구한 ( $p=0, d=1, q=1$ ) 값으로 설정했고, 계절성 모수는 차분만 설정했다 ( $D=1$ ).

구성한 모델의 적합도를 평가하는 지표로 AIC(Akaike Information Criterion)를 사용하였다. AIC는 모델의 적합도를 평가하는 지표로, 값이 작을수록 더 좋은 모델을 의미한다. SARIMA 모형 구성 결과, AIC 값은 1047.98이 나왔다.

모델 검증과정에서 Ljung-Box 검정을 통해 현재 구성된 SARIMA 모형의 잔차가 백색잡음인지(잔차간 상관관계가 없는지)를 검정할 것이다. Ljung-Box 검정은 시계열 데이터의 잔차가 독립적인지를 확인하는 통계적 방법이다.

Ljung-Box 검정 결과 p-value 값이 0.2034로 0.05보다 크다. 따라서 모델 잔차의 백색 잡음을 검증하였다. 이는 잔차가 시계열의 독립성을 만족하여 모델이 적합함을 나타낸다.

### 5.6.2 auto.arima를 활용한 자동 모델 설계 (Seasonal=TRUE)

모델을 auto.arima를 활용하여 모델 학습을 진행하였다. auto.arima는 R의 forecast 패키지에 포함되어 있는 함수로, 주어진 시계열 데이터에 대해 가장 적절한 형태의 ARIMA 모형을 찾아서 모수를 추정해준다. 모델 학습 결과, 사용자 정의 모델 설계와 달리, 계절성 부분에  $Q=1$ 이 추가되었으며, AIC 값이 1028.31로 사용자 정의 모델보다 성능이 우수함을 보여준다.

다만, auto.arima는 AIC 등 특정 모형 성능지표에 기반하여 최적값을 찾아줄 뿐이기 때문에, 과도한 차분이나 AR, MA 수준을 결정할 위험이 있다. 따라서 Ljung Box 검증을 재차 진행하여 모델 잔차의 백색잡음을 검증하고자 한다.

Ljung-Box 검정 결과 p-value 값이 0.06089로 0.05보다 크다. 따라서 자동 학습 ARIMA 또한 모델의 잔차가 백색잡음이므로, 모형이 데이터의 자기상관에 휘둘리지 않고, 잘 설명하고 있다는 것을 알 수 있다.

## 5.7 2024년 담배 판매량 예측

구성된 모델을 사용하여 2024년 담배 판매량을 예측하고자 한다. 모델은 대부분의 성능 지표, 특히 AIC 성능 지표가 더 높은 auto.arima로 학습된 모델을 예측에 활용하였다.

2014년~2023년까지의 담배 판매량 데이터를 기반으로 예측된 2024년 6월 담배 판매량은 약 2.58억 갑이다.

## 6. 종합 결론

본 프로젝트는 2015년 담뱃값 인상이 흡연율과 담배 판매량에 미친 영향을 단기적 및 장기적 관점에서 분석하고, 이를 바탕으로 2024년 담배 판매량 변화를 예측하였다.

먼저, RDD 분석을 통해 2015년 담뱃값 인상 직후의 단기적 효과를 평가하였다. 6개월, 1년, 1년 6개월의 대역폭을 설정하고 다양한 RDD 회귀모형을 적용한 결과, 특히 1년 6개월의 대역폭과 3차항 비선형 RDD 회귀모형이 가장 적합한 것으로 나타났다. 이 분석에서 추정된 LATE 값은 -2.0929로, 담뱃값 인상이 단기적 관점에서 담배 판매량을 유의미하게 감소시키는 효과가 있었음을 확인하였다.

다음으로, DID 분석을 통해 장기적 효과를 평가하였다. 한국과 일본을 비교 대상으로 설정하여 분석한 결과, 국가, 연도, 처치 후, GDP 변수의 p-value가 모두 0.05보다 커 개별적으로 통계적으로 유의하지 않았다. 이는 국가나 연도에 따른 흡연율 차이가 유의미하게 나타나지 않았음을 의미한다. 특히, DID Estimator의 p-value가 유의하지 않아 장기적 처치 효과는 통계적으로 유의미하지 않음을 확인하였다. 이는 담뱃값 인상이 장기적으로 흡연율 감소에 유의미한 영향을 미치지 않았음을 시사한다. 따라서 유의미한 흡연율 감소를 위해서는 세수 증대보다 더 큰 폭의 가격 인상이 필요할 수 있다.

마지막으로, 시계열 분석을 통해 2024년 담배 판매량을 예측하였다. auto.arima로 학습된 모델을 활용하였으며, 2014년~2023년까지의 담배 판매량 데이터를 기반으로 예측된 2024년 6월 담배 판매량은 약 2.58억으로 나타났다. 2024년 총선 이후, 2024년 6월에 8000원 또는 1만원으로 담뱃값 인상 정책이 시행되었다 가정하면, 이전 RDD, DID 분석 결과를 반영하여 예측한 2024년 6월 담배 판매량은 기존 예측 값인 약 2.58억 갑보다 대폭 감소할 것이다.

본 프로젝트는 담뱃값 인상의 효과를 다각도로 분석함으로써, 정책 결정에 중요한 시사점을 제공한다. 다만, 데이터 표본 수의 한계로 인해 일부 결과의 통계적 유의성을 확보하지 못했음을 감안할 필요가 있다. 향후 연구에서는 더 많은 데이터와 다양한 변수들을 고려하여 보다 정밀한 분석이 필요할 것이다.

## 7. 참고문헌

- 윤용호 외 1명, 『 담뱃세 인상정책의 흡연억제 효과 : 시계열자료를 통한 실증분석 』, 규제연구, 2007. 12, 22page
- 민경문 외 2명, 『 담배가격 인상에 따른 청소년 흡연을 변화 』, 한국산학기술학회 논문지, 2018, 11page
- 김환희, 담배가격 인상에 따른 청소년의 흡연 및 금연 시도 관련요인 분석 : 제10,11,12,13차 청소년건강행태온라인조사(2014-2017년) 분석, YUHSpace, 2017, 63page
- 이상학, ""총선 후 담뱃값 오른다?"...한 갑 4500원→8000원·1만원 '인상설 솔솔'", 뉴스1, 2024.01.16, <https://www.news1.kr/articles/?5290870>

## 8. 부록

### • RDD 분석 데이터

월년	담배판매량	월년	담배판매량
Jan-12	3.4	Jul-14	4.1
Feb-12	3.2	Aug-14	3.6
Mar-12	3.4	Sep-14	4.3
Apr-12	3.6	Oct-14	3.7
May-12	3.9	Nov-14	3.6
Jun-12	3.7	Dec-14	3.9
Jul-12	4	Jan-15	1.7
Aug-12	3.7	Feb-15	1.8
Sep-12	4.1	Mar-15	2.4
Oct-12	3.4	Apr-15	2.9
Nov-12	3.7	May-15	2.7
Dec-12	3.5	Jun-15	3.1
Jan-13	3.3	Jul-15	3.5
Feb-13	3.2	Aug-15	2.9
Mar-13	3.5	Sep-15	3.4
Apr-13	3.5	Oct-15	2.9
May-13	3.9	Nov-15	2.9
Jun-13	3.4	Dec-15	3
Jul-13	4.1	Jan-16	2.7
Aug-13	3.7	Feb-16	2.8
Sep-13	3.8	Mar-16	3
Oct-13	3.7	Apr-16	3.1
Nov-13	3.5	May-16	3.1
Dec-13	3.6	Jun-16	3.2
Jan-14	3.3	Jul-16	3.3
Feb-14	2.7	Aug-16	3.2
Mar-14	3.3	Sep-16	3.4
Apr-14	3.9	Oct-16	2.9
May-14	3.5	Nov-16	3.1
Jun-14	3.6	Dec-16	2.9

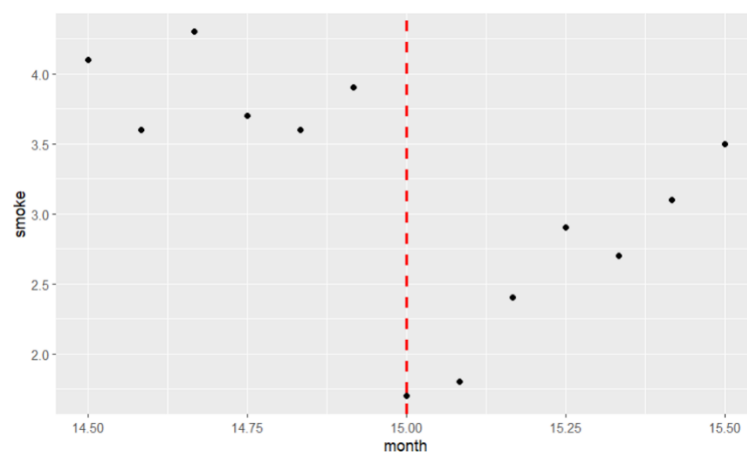
단위: 억 갑

- 현재, 매일 흡연을 – 담배 판매량 상관계수 데이터

연도	현재 흡연율	담배 판매량	매일 흡연율
2012	25.80%	43.6	23.60%
2013	24.10%	43.2	21.40%
2014	24.20%	43.5	21.70%
2015	22.60%	33.2	18.70%
2016	23.90%	36.7	20.20%

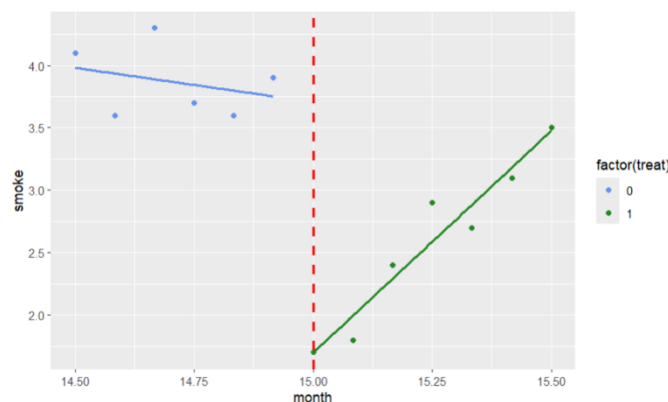
단위: 억 갑

- 대역폭 6개월 설정 시각화



- 대역폭: 6개월 & 모형: RDD 회귀모형

- ✓ RDD 회귀 모형 결과 시각화



✓ RDD 회귀 모형 추정 결과

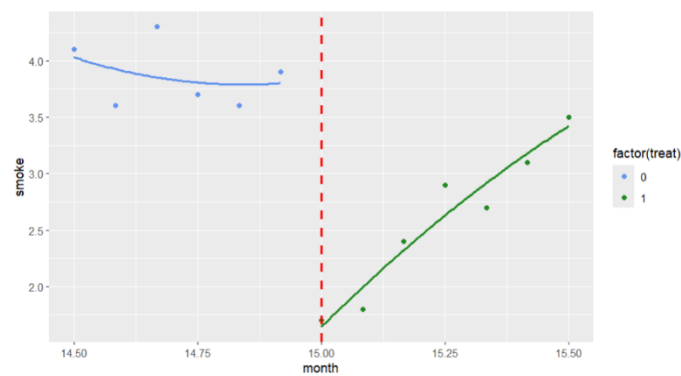
```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.7067    0.2332   15.891 6.83e-08 ***
treat          -2.0102    0.2891   -6.955 6.65e-05 ***
month_d        -0.5486    0.7187   -0.763  0.46484
treat:month_d    4.1057    0.9162    4.481  0.00153 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

- 대역폭: 6개월 & 모형: 2차항 비선형 RDD 회귀모형

✓ 비선형 RDD 회귀 모형 결과 시각화 (2차항)



✓ 비선형 RDD 회귀 모형 추정 (2차항)

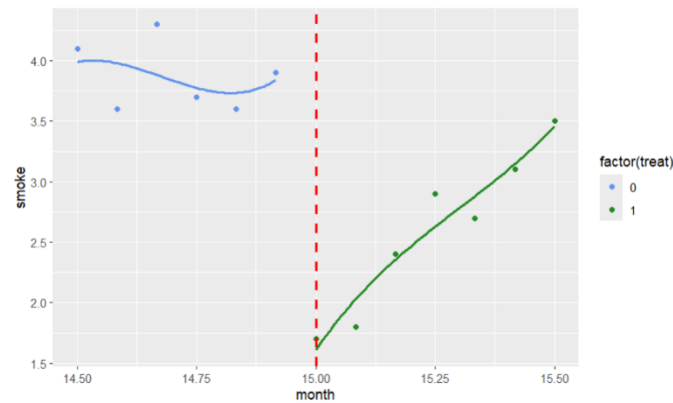
```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.8400    0.5004    7.674 0.000119 ***
treat          -2.1971    0.5568   -3.946 0.005559 **
month_d         0.6514    3.9284    0.166 0.872981
I(month_d^2)    2.0571    6.5924    0.312 0.764093
treat:month_d    3.6771    4.5457    0.809 0.445176
treat:I(month_d^2) -3.6000    7.9231   -0.454 0.663317
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

- 대역폭: 6개월 & 모형: 3차항 비선형 RDD 회귀모형

- ✓ 비선형 RDD 회귀 모형 결과 시각화 (3차항)



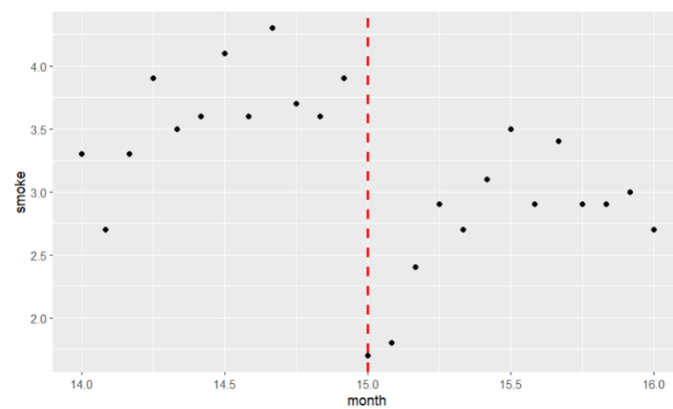
- ✓ 비선형 RDD 회귀 모형 추정 (3차항)

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.16667    1.17406   3.549  0.0164 *
treat         -2.55714    1.21527  -2.104  0.0893 .
month_d        5.58254   16.05275   0.348  0.7422
I(month_d^2)  21.65714   61.64165   0.351  0.7397
I(month_d^3)  22.40000   69.89952   0.320  0.7616
treat:month_d   0.07937   17.11885   0.005  0.9965
treat:I(month_d^2) -30.40000  68.19362  -0.446  0.6744
treat:I(month_d^3) -12.80000  79.69771  -0.161  0.8787
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

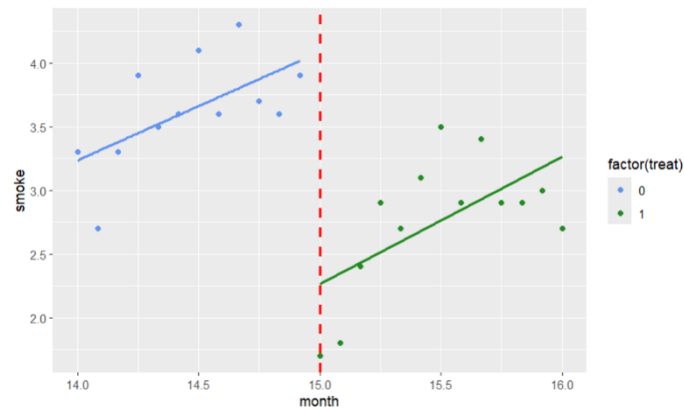
```

- 대역폭 1년 설정 시각화



- 대역폭: 1년 & 모형: RDD 회귀모형

- ✓ RDD 회귀 모형 결과 시각화



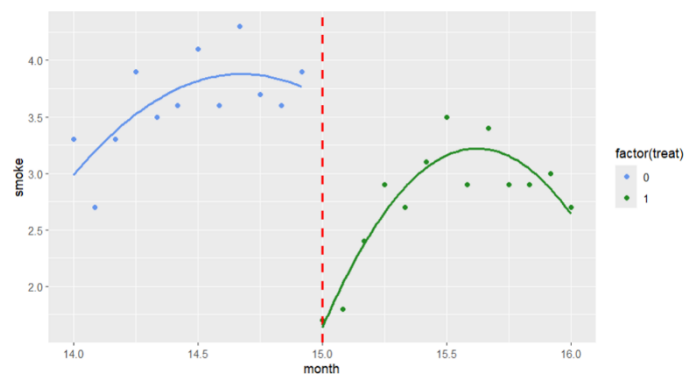
- ✓ RDD 회귀 모형 추정 결과

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0864     0.2448  16.690 1.35e-13 ***
treat         -1.8259     0.3216  -5.678 1.23e-05 ***
month_d        0.8517     0.3992   2.134  0.0448 *
treat:month_d   0.1504     0.5335   0.282  0.7807
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  
```

- 대역폭: 1년 & 모형: 2차항 비선형 RDD 회귀모형

- ✓ 비선형 RDD 회귀 모형 결과 시각화 (2차항)





✓ 비선형 RDD 회귀 모형 추정 (2차항)

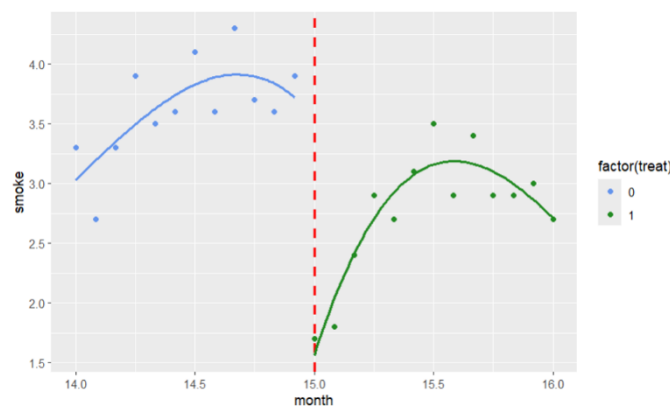
```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6750    0.2871  12.799 8.65e-11 ***
treat          -2.0387    0.3497   -5.830 1.29e-05 ***
month_d        -1.2638    1.2186   -1.037 0.312700
I(month_d^2)    -1.9528    1.0950   -1.783 0.090509 .
treat:month_d    6.3515    1.5315    4.147 0.000547 ***
treat:I(month_d^2) -2.1327    1.4137   -1.509 0.147853
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

• 대역폭: 1년 & 모형: 3차항 비선형 RDD 회귀모형

✓ 비선형 RDD 회귀 모형 결과 시각화 (3차항)



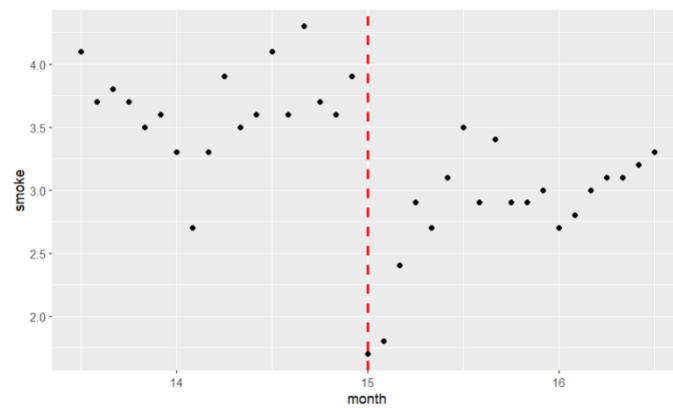
✓ 비선형 RDD 회귀 모형 추정 (3차항)

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.5566    0.4747    7.492 8.81e-07 ***
treat          -1.9953    0.5354   -3.727 0.00168 **
month_d        -2.3623    3.6425   -0.649 0.52531
I(month_d^2)    -4.3892    7.6559   -0.573 0.57394
I(month_d^3)    -1.4993    4.6585   -0.322 0.75149
treat:month_d    8.5818    4.2734    2.008 0.06078 .
treat:I(month_d^2) -2.6418    9.3247   -0.283 0.78036
treat:I(month_d^3)  3.4629    5.8231    0.595 0.55988
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

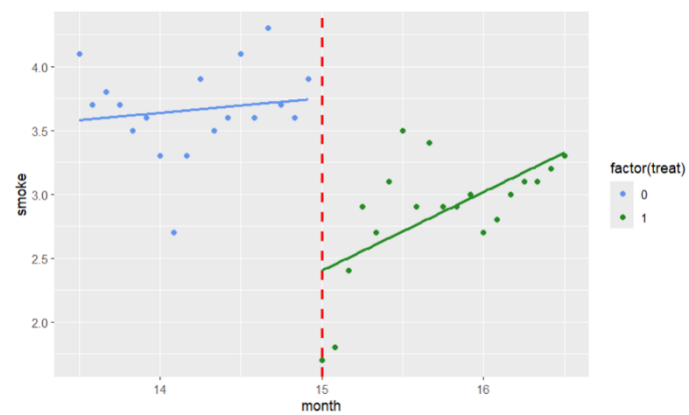
```

- 대역폭 1년 6개월 설정 시각화



- 대역폭: 1년 6개월 & 모형: RDD 회귀모형

- ✓ RDD 회귀 모형 결과 시각화



- ✓ RDD 회귀 모형 추정 결과

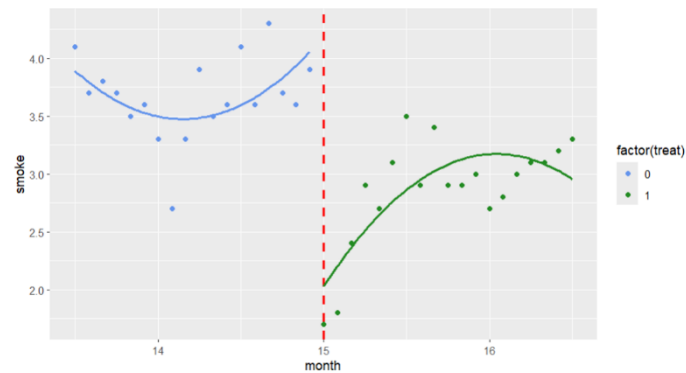
```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7523    0.1835   20.445 < 2e-16 ***
treat         -1.3518    0.2466  -5.482 4.45e-06 ***
month_d        0.1152    0.2035    0.566  0.575
treat:month_d  0.5017    0.2767    1.813  0.079 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

- 대역폭: 1년 6개월 & 모형: 2차항 비선형 RDD 회귀모형

✓ 비선형 RDD 회귀 모형 결과 시각화 (2차항)



✓ 비선형 RDD 회귀 모형 추정 (2차항)

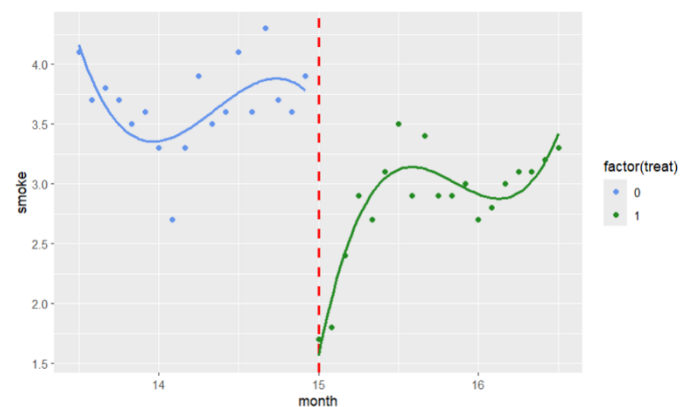
```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.1853    0.2625   15.942 < 2e-16 ***
treat          -2.1566    0.3335   -6.467 3.28e-07 ***
month_d         1.6740    0.7634    2.193 0.03596 *
I(month_d^2)    0.9845    0.4685    2.101 0.04384 *
treat:month_d   0.5175    0.9934    0.521 0.60607
treat:I(month_d^2) -2.0343    0.6219   -3.271 0.00263 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

- 대역폭: 1년 6개월 & 모형: 3차항 비선형 RDD 회귀모형

✓ 비선형 RDD 회귀 모형 결과 시각화 (3차항)



## ✓ 비선형 RDD 회귀 모형 추정 (3차항)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6534    0.3016   12.113 7.21e-13 ***
treat          -2.0929    0.3592   -5.827 2.56e-06 ***
month_d        -1.8851    1.6052   -1.174 0.249796
I(month_d^2)    -4.4861    2.3228   -1.931 0.063271 .
I(month_d^3)    -2.3034    0.9661   -2.384 0.023871 *
treat:month_d    8.4192    1.9790    4.254 0.000200 ***
treat:I(month_d^2) -4.0003    2.9524   -1.355 0.185902
treat:I(month_d^3) 5.6086    1.2528    4.477 0.000108 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## • RDD 분석 코드

```
1 #데이터 활동 패키지 및 시각화 패키지 설치 및 불러오기
2 install.packages("dplyr")
3 install.packages("ggplot2")
4 library(dplyr)
5 library(ggplot2)
6
7 #데이터 불러오기
8 mydata <- read.csv("C:\\Users\\yo218218\\Desktop\\smoke_rdd2.csv", sep=',')
9
10 #데이터 둘러보기
11 head(mydata)
12 summary(mydata)
13
14 #시각화 기본
15 mydata %>%
16   ggplot(aes(x = month, y = smoke)) +
17     geom_point() +
18     geom_vline(xintercept = 15, color = "red", size = 1, linetype = "dashed") +
19     labs(y = "smoke", x = "month")
20
21 #대역폭 설정
22 filtered_data <- mydata %>%
23   filter(month >= 13.5 & month <= 16.5)
24
25 #시각화
26 filtered_data %>%
27   ggplot(aes(x = month, y = smoke)) +
28     geom_point() +
29     geom_vline(xintercept = 15, color = "red", size = 1, linetype = "dashed") +
30     labs(y = "smoke", x = "month")
31
32 #RDD 회귀 모형 추정
33 mymodel1 <- lm(smoke ~ treat + month_d + treat:month_d, data = filtered_data)
34
35 summary(mymodel1)
36
37 #RDD 회귀 모형 결과 시각화
38 filtered_data %>%
39   ggplot(aes(x = month, y = smoke, color = factor(treat))) +
40     geom_point() +
41     geom_vline(xintercept = 15, color = "red", size = 1, linetype = "dashed") +
42     geom_smooth(data = subset(filtered_data, month < 15), method = "lm", color = "cornflowerblue", se = FALSE) +
43     geom_smooth(data = subset(filtered_data, month >= 15), method = "lm", color = "forestgreen", se = FALSE) +
44     scale_color_manual(values = c("0" = "cornflowerblue", "1" = "forestgreen")) +
45     labs(y = "smoke", x = "month")
46
47 #RDD 비선형 회귀 모형 추정 (2차항)
48 mymodel2 <- lm(smoke ~ treat + month_d + I(month_d^2) + treat:month_d + treat:I(month_d^2), data = filtered_data)
49
50 summary(mymodel2)
51
52 #RDD 비선형 회귀 모형 결과 시각화 (2차항)
53 filtered_data %>%
54   ggplot(aes(x = month, y = smoke, color = factor(treat))) +
55     geom_point() +
56     geom_vline(xintercept = 15, color = "red", size = 1, linetype = "dashed") +
57     geom_smooth(data = subset(filtered_data, month < 15), method = "lm", formula = y ~ x + I(x^2), color = "cornflowerblue", se = FALSE) +
58     geom_smooth(data = subset(filtered_data, month >= 15), method = "lm", formula = y ~ x + I(x^2), color = "forestgreen", se = FALSE) +
59     scale_color_manual(values = c("0" = "cornflowerblue", "1" = "forestgreen")) +
60     labs(y = "smoke", x = "month")
61
62 #RDD 비선형 회귀 모형 추정 (3차항)
63 mymodel3 <- lm(smoke ~ treat + month_d + I(month_d^2) + I(month_d^3) + treat:month_d + treat:I(month_d^2) + treat:I(month_d^3), data = filtered_data)
64
65 summary(mymodel3)
66
67 #RDD 비선형 회귀 모형 결과 시각화 (3차항)
68 filtered_data %>%
69   ggplot(aes(x = month, y = smoke, color = factor(treat))) +
70     geom_point() +
71     geom_vline(xintercept = 15, color = "red", size = 1, linetype = "dashed") +
72     geom_smooth(data = subset(filtered_data, month < 15), method = "lm", formula = y ~ x + I(x^2) + I(x^3), color = "cornflowerblue", se = FALSE) +
73     geom_smooth(data = subset(filtered_data, month >= 15), method = "lm", formula = y ~ x + I(x^2) + I(x^3), color = "forestgreen", se = FALSE) +
74     scale_color_manual(values = c("0" = "cornflowerblue", "1" = "forestgreen")) +
75     labs(y = "smoke", x = "month")
```

- DID 분석 사용 데이터

<https://stats.oecd.org/Index.aspx?ThemeTreeId=9> – OECD statistics 2023

- FCTC 가입일자와 국가 목록

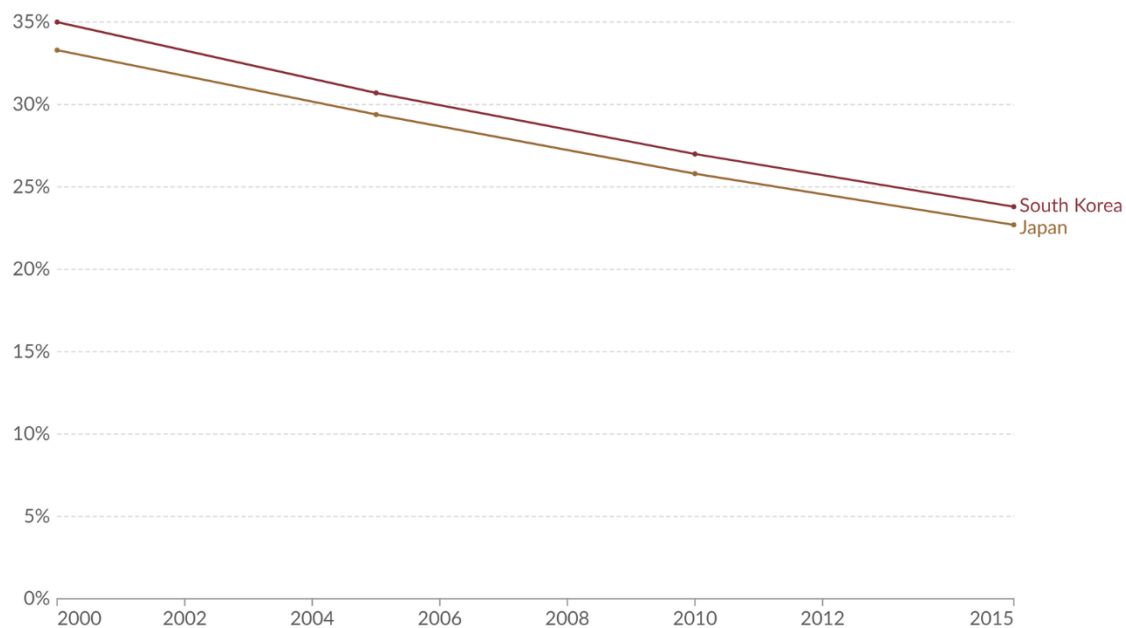
[https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg\\_no=IX-4&chapter=9&clang](https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IX-4&chapter=9&clang)

- 2000~2015년의 성인 흡연율 변화 추이

## Share of adults who smoke, 2000 to 2015

Our World  
in Data

Share of people aged 15 and older who smoke any tobacco product on a daily or non-daily basis. It excludes smokeless tobacco use. Smoking is a risk factor<sup>1</sup> for chronic complications, including cancers, cardiovascular disease<sup>2</sup>, and premature death.



Data source: Multiple sources compiled by World Bank (2024)

OurWorldInData.org/smoking | CC BY

**1. Risk factor:** A risk factor is a condition or behavior that increases the likelihood of developing a given disease or injury, or an outcome such as death. The impact of a risk factor is estimated in different ways. For example, a common approach is to estimate the number of deaths that would occur if the risk factor was absent. Risk factors are not mutually exclusive: people can be exposed to multiple risk factors, which contribute to their disease or death. Because of this, the number of deaths caused by each risk factor is typically estimated separately. [Read more: How do researchers estimate the death toll caused by each risk factor, whether it's smoking, obesity or air pollution?](#) [Read more: Why isn't it possible to sum up the death toll from different risk factors?](#)

**2. Cardiovascular disease:** Cardiovascular diseases cover all diseases of the heart and blood vessels – including heart attacks and strokes, atherosclerosis, ischemic heart disease, hypertensive diseases, cardiomyopathy, rheumatic heart disease, and more. They tend to develop gradually with age, especially when people have risk factors like high blood pressure, smoking, alcohol use, poor diet, and air pollution.

## • DID 분석 코드

```

1 # 필요한 패키지 로드
2 library(dplyr)
3 library(ggplot2)
4
5 # csv 파일 읽기
6 data <- read.csv("연간 흡연율.csv", encoding = "utf-8")
7
8 # DID 분석
9 data$처치후 <- as.integer(data$국가 == 1 & data$연도 >= 2015)
10 did_model <- lm(흡연율 ~ 국가 + 연도 + 처치후 + GDP, data = data)
11 summary(did_model)
12
13 # 시각화를 위한 데이터 준비
14 korea_data <- data %>% filter(국가 == 1) %>% select(연도, 흡연율)
15 japan_data <- data %>% filter(국가 == 0) %>% select(연도, 흡연율)
16
17 # 처치 이전 데이터를 사용하여 회귀모형 적합
18 pre_treatment_data <- korea_data %>% filter(연도 < 2015)
19 pre_treatment_model <- lm(흡연율 ~ 연도, data = pre_treatment_data)
20
21 # 처치군의 반사실 데이터 생성
22 counterfactual_data <- data.frame(연도 = korea_data$연도)
23 counterfactual_data$흡연율 <- ifelse(counterfactual_data$연도 < 2015,
24                                   korea_data$흡연율[counterfactual_data$연도 < 2015],
25                                   predict(pre_treatment_model, newdata = data.frame(연도 = counterfactual_data$연도[counterfactual_data$연도 >= 2015])))
26
27 # 시각화
28 ggplot() +
29   geom_line(data = korea_data, aes(x = 연도, y = 흡연율), color = "red", size = 1) +
30   geom_point(data = korea_data, aes(x = 연도, y = 흡연율), color = "red", size = 3) +
31   geom_line(data = japan_data, aes(x = 연도, y = 흡연율), color = "blue", size = 1) +
32   geom_point(data = japan_data, aes(x = 연도, y = 흡연율), color = "blue", size = 3) +
33   geom_line(data = counterfactual_data, aes(x = 연도, y = 흡연율), color = "red", size = 1, linetype = "dashed") +
34   geom_point(data = counterfactual_data, aes(x = 연도, y = 흡연율), color = "red", size = 3, shape = 1) +
35   geom_vline(xintercept = 2015, linetype = "dashed", color = "gray", size = 1) +
36   labs(x = "연도", y = "매일 흡연율") +
37   scale_x_continuous(breaks = seq(2011, 2018, 1)) +
38   theme_minimal() +
39   theme(text = element_text(size = 12)) +
40   annotate("text", x = 2013, y = max(data$흡연율) * 0.92, label = "처치군(한국)", color = "red", size = 5) +
41   annotate("text", x = 2013, y = max(data$흡연율) * 0.8, label = "통제군(일본)", color = "blue", size = 5) +
42   annotate("text", x = 2015.5, y = max(data$흡연율) * 0.9, label = "처치시점", color = "gray", size = 5) +
43   annotate("text", x = 2017, y = max(data$흡연율) * 0.67, label = "반사실", color = "red", size = 5)

```

## • DID 분석 결과

```
> summary(did_model)
```

Call:

```
lm(formula = 흡연율 ~ 국가 + 연도 + 처치후 + GDP, data = data)
```

Residuals:

```

      Min       1Q   Median       3Q      Max
-1.14727 -0.42148  0.03716  0.41866  1.29816

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.493e+02  7.431e+02   1.008   0.335
국가          9.563e-01  1.462e+00   0.654   0.526
연도         -3.595e-01  3.804e-01  -0.945   0.365
처치후       -1.487e+00  1.037e+00  -1.434   0.179
GDP          -1.541e-04  6.039e-04  -0.255   0.803

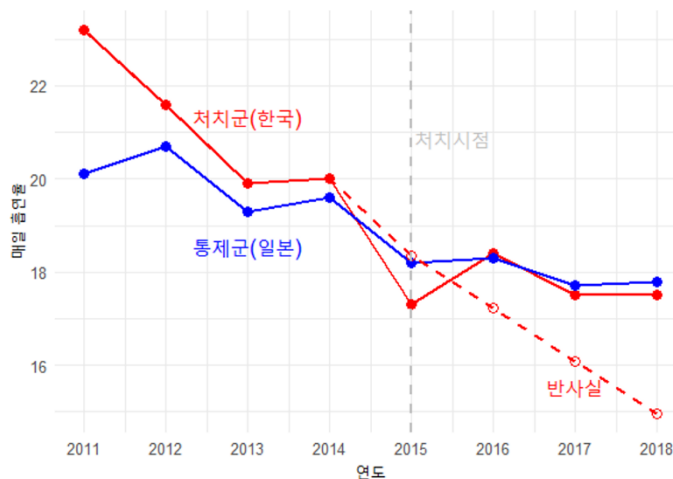
```

```

Residual standard error: 0.7454 on 11 degrees of freedom
Multiple R-squared:  0.8551,    Adjusted R-squared:  0.8024
F-statistic: 16.22 on 4 and 11 DF,  p-value: 0.0001389

```

## • DID 분석 결과 시각화



- 시계열 분석 사용 데이터

<https://www.korea.kr/briefing/pressReleaseView.do?newsId=156613275>

– 2014.01~2023.12 담배 판매량 데이터

- DF 검정 결과

```
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####

Test regression trend

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min       1Q   Median       3Q      Max
-178.913  -14.315    1.737   16.566   89.277

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  155.13610    32.15650   4.824 4.39e-06 ***
z.lag.1       -0.48312     0.09836  -4.912 3.04e-06 ***
tt            -0.34602     0.11294  -3.064 0.00273 **
z.diff.lag    -0.15479     0.09155  -1.691 0.09361 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.41 on 114 degrees of freedom
Multiple R-squared:  0.3068,    Adjusted R-squared:  0.2886
F-statistic: 16.82 on 3 and 114 DF,  p-value: 4.11e-09

Value of test-statistic is: -4.9117 8.059 12.0725

Critical values for test statistics:
      1pct  5pct 10pct
tau3  -3.99 -3.43 -3.13
phi2   6.22  4.75  4.07
phi3   8.43  6.49  5.47
```

- ADF 검정 결과

```
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####

Test regression trend

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
    Min       1Q   Median       3Q      Max
-178.913  -14.315    1.737   16.566   89.277

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  155.13610    32.15650   4.824 4.39e-06 ***
z.lag.1       -0.48312     0.09836  -4.912 3.04e-06 ***
tt            -0.34602     0.11294  -3.064 0.00273 **
z.diff.lag    -0.15479     0.09155  -1.691 0.09361 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.41 on 114 degrees of freedom
Multiple R-squared:  0.3068,    Adjusted R-squared:  0.2886
F-statistic: 16.82 on 3 and 114 DF,  p-value: 4.11e-09

Value of test-statistic is: -4.9117 8.059 12.0725

Critical values for test statistics:
      1pct  5pct 10pct
tau3  -3.99 -3.43 -3.13
phi2   6.22  4.75  4.07
phi3   8.43  6.49  5.47
```



- 사용자 정의 모델 설계

```
ARIMA(0,1,1)(0,1,0)[12]

Coefficients:
      ma1
    -0.6047
    s.e.   0.0837

sigma^2 = 1017:  log likelihood = -521.99
AIC=1047.98   AICc=1048.1   BIC=1053.33

Training set error measures:
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
Training set 3.051826 29.96557 18.9863 1.03184 7.28222 0.6811909 0.01584903
```

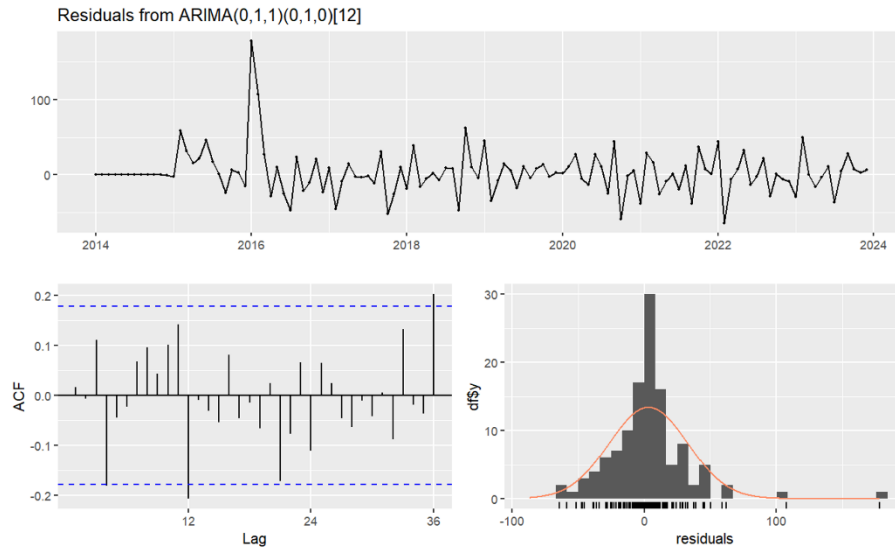
- Ljung-Box 검정 결과

### Ljung-Box test

```
data: Residuals from ARIMA(0,1,1)(0,1,0)[12]
Q* = 28.335, df = 23, p-value = 0.2034

Model df: 1.   Total lags used: 24
```

- Ljung-Box 검정 결과 시각화





- **auto.arima 활용 자동 모델 설계**

```
ARIMA(0,1,1)(0,1,1)[12]

Coefficients:
      ma1      sma1
    -0.6427  -0.6117
s.e.    0.0883   0.1098

sigma^2 = 792.8: log likelihood = -511.04
AIC=1028.07  AICc=1028.31  BIC=1036.09

Training set error measures:
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1
Training set 4.498277 26.33781 16.945 1.662671 6.418097 0.607953 0.06636034
```

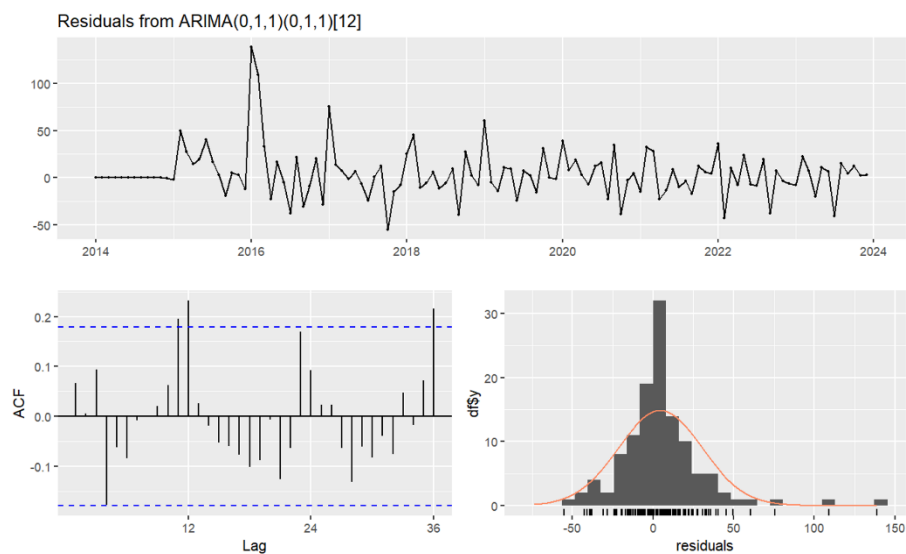
- **Ljung-Box 검정 결과**

Ljung-Box test

data: Residuals from ARIMA(0,1,1)(0,1,1)[12]  
 Q\* = 33.07, df = 22, p-value = 0.06089

Model df: 2. Total lags used: 24

- **Ljung-Box 검정 결과 시각화**



- 2024년 담배 판매량 예측

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan 2024	234.8063	198.7214	270.8913	179.6191	289.9935
Feb 2024	202.3298	164.0108	240.6488	143.7260	260.9337
Mar 2024	247.1007	206.6710	287.5305	185.2688	308.9327
Apr 2024	247.8481	205.4125	290.2838	182.9484	312.7479
May 2024	260.2300	215.8791	304.5809	192.4011	328.0589
Jun 2024	258.3822	212.1954	304.5690	187.7455	329.0189
Jul 2024	269.5676	221.6150	317.5201	196.2305	342.9046
Aug 2024	259.2515	209.5960	308.9069	183.3100	335.1929
Sep 2024	273.3729	222.0710	324.6748	194.9134	351.8323
Oct 2024	235.7042	182.8071	288.6013	154.8051	316.6034
Nov 2024	238.8926	184.4470	293.3382	155.6253	322.1600
Dec 2024	238.9907	183.0395	294.9420	153.4207	324.5608

단위: 백만 갑

- 2024년 담배 판매량 예측 시각화

