# 범죄와 여러요인 상관관계 분석

3조 유리창 : 문동민, 김수진, 신효섭, 이승훈

# 목차 (CONTENTS)

# 1. 프로젝트 선정이유

## CDA(Confirmatory Data Analysis)

우선 **CDA(Confirmatory Data Analysis)**는 확증적 데이터 분석인데 가설을 세운 후 이를 데이터를 통해 검증하는 방식이다.

- 목적을 가지고 데이터를 확보하여 분석하는 방법
- 관측된 형태나 효과의 재현성 평가, 유의성 검정, 신뢰구간 추정 등 통계적 추론을 하는 단계
- 가설검정, 보통은 설문조사, 논문에 대한 내용을 입증하는데 많이 사용

## EDA(Exploratory Data Analysis)

**EDA(Exploratory Data Analysis)**는 탐색적 데이터 분석인데 데이터를 먼저 살펴본 후 인사이트를 도출하는 과정이다.

- 쌓여있는 데이터를 기반으로 가설을 세워 데이터를 분석하는 방법
- 데이터의 구조와 특징을 파악하며 여기서 얻은 정보를 바탕으로 통계모형을 만드는 단계
- 빅데이터 분석에 사용됨

# 1. 프로젝트 선정이유


"깨진 유리창 이론"으로 불법투기 근절에 앞장서!

박철우 기자 | 입력 2021.06.22 16:09 | 댓글 0


통영시, '깨진 유리창 이론'에 도전한다

상습 쓰레기 투기장소에 벽화 제작 시민의식 전환 시도

서용찬 기자(=통영) | 기사입력 2019.10.22. 14:53:51 최종수정 2019.10.22. 14:54:00


HOME > 인천

장수서창동'깨진 유리창 이론'을 활용한 환경시범거리 지정

임영화 기자 webmaster@kmaeil.com | 승인 2015.11.13 16:58 | 댓글 0

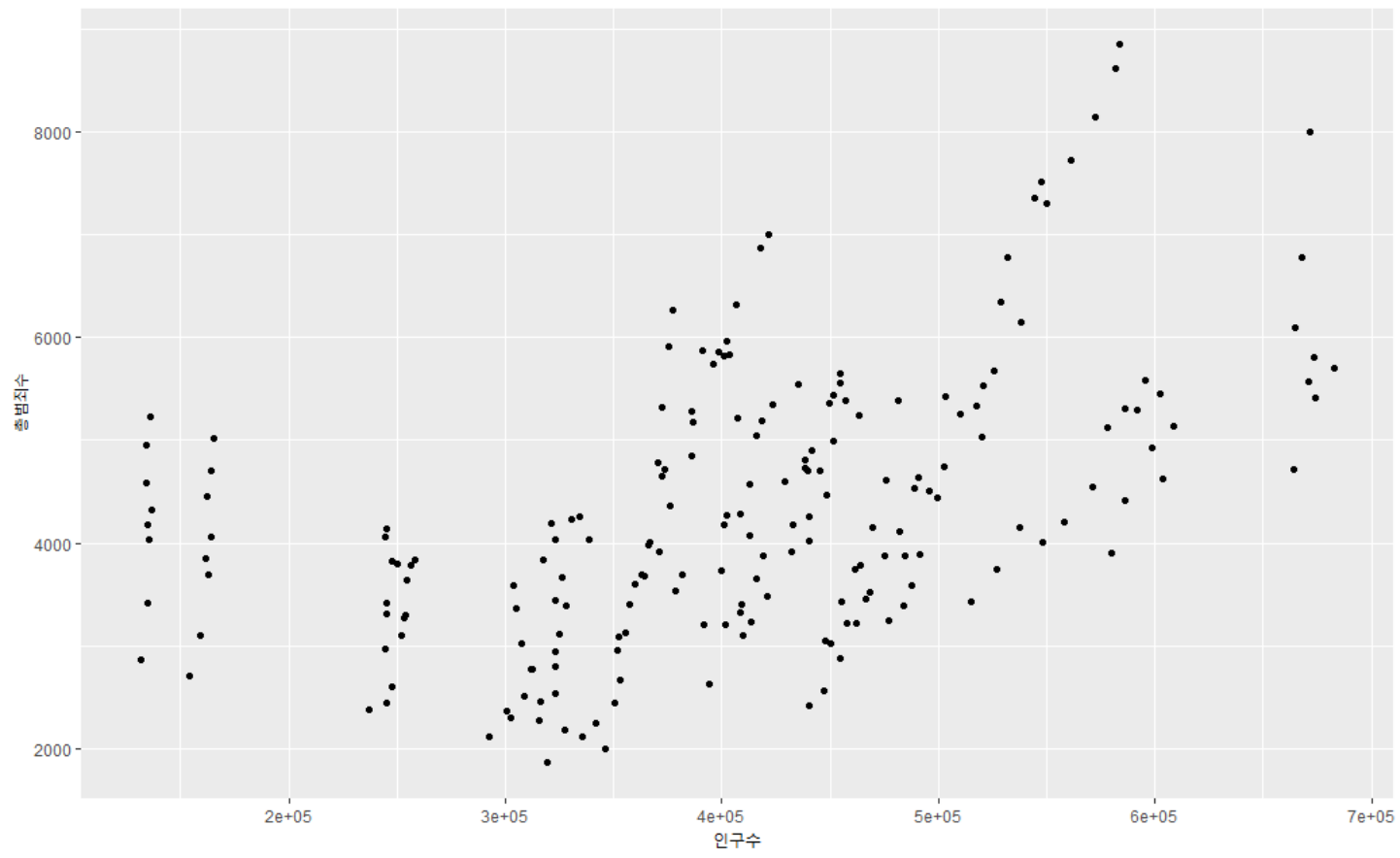
## 깨진 유리창 이론

깨진 유리창 하나를 방치해 두면 그 지점을 중심으로 범죄가 확산되기 시작한다는 이론으로, 사소한 무질서를 방치했다가 나중엔 지역 전체로 확산될 가능성이 높다는 의미를 담는다.

우리조는 주위에 있는 사소한 것과 범죄와의 비교분석을 통하여 상관관계를 알아보고자 한다.

- 주위에서 발생하는 사건
- 서울시 구별
- 데이터,범죄정보,쓰레기배출량,가로등개수,CCTV개수,인구수,
- 인구밀도등의 데이터를 받아서 데이터 전처리 진행

# 2. 데이터분석 – 데이터 정규화




년도마다의 증가치가 분석의 신뢰도를 떨어뜨림

# 2. 데이터분석 – 피어슨상관계수 히트맵 결과

- 회귀분석은 다음과 같은 조건을 만족해야함
  - **선형성**
  - **등분산성**
  - **독립성**
  - **정규성**

### ★ 귀무가설

변수1과 변수2와 관련이 없다.

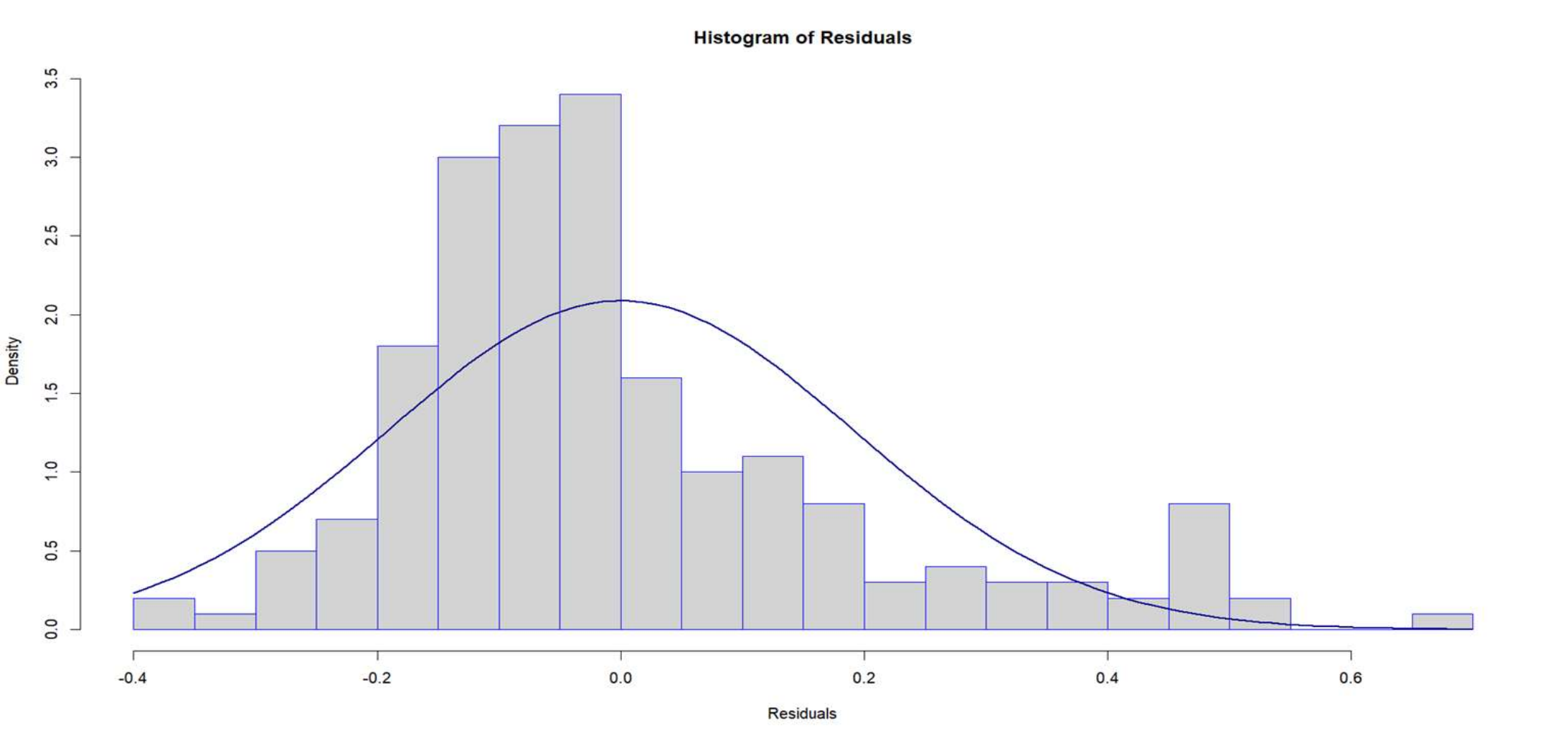
- 기존에 보편적인 사실인 것처럼 알려져 있는 것
- P-value 값이 0.05보다 크다.
- 귀무가설이 참이면 대립가설은 거짓

### ★ 대립가설

변수1과 변수2와 관련이 있다.

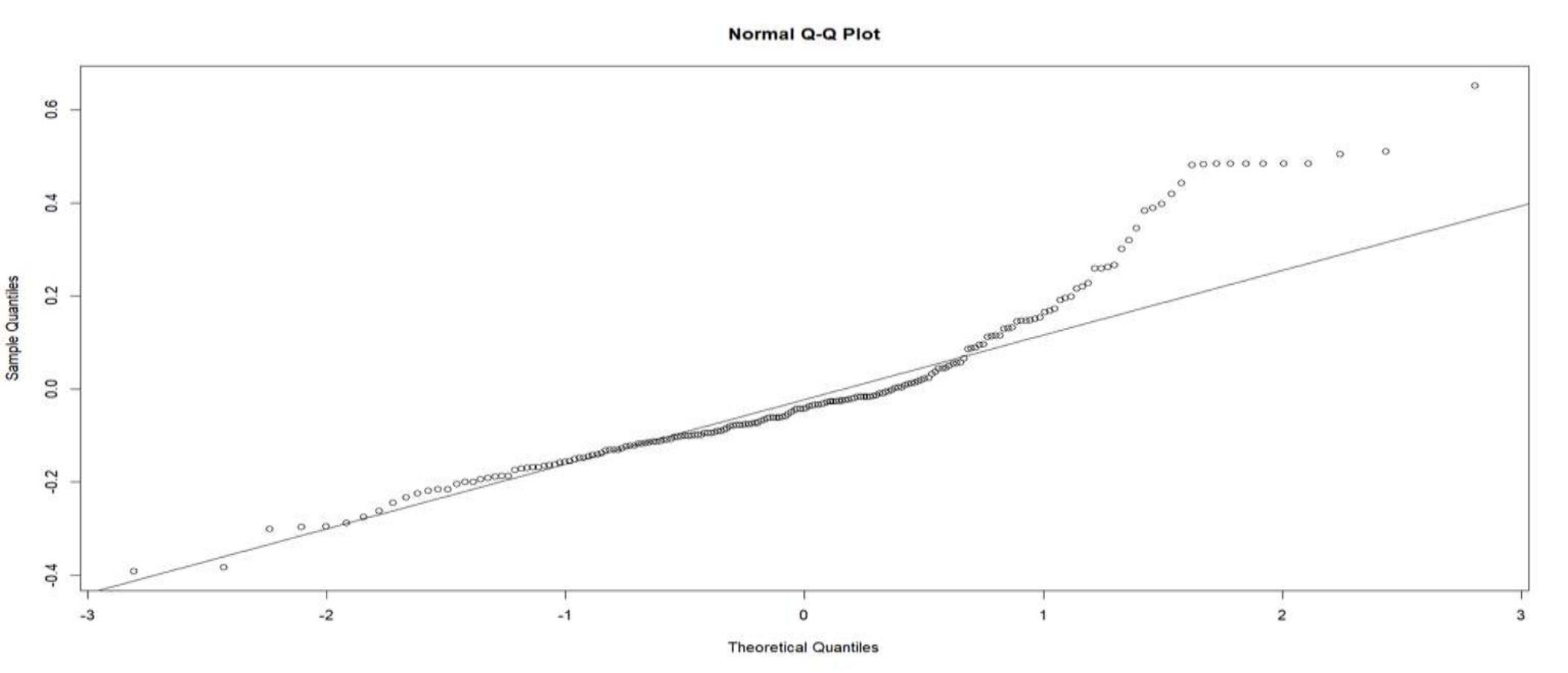
- 사실처럼 알려져 있는 것이 상식과 다름을 증명하려는 것
- P-value 값이 0.05보다 작다.
- 귀무가설이 거짓이면 대립가설은 참

## 2. 가로등과 쓰레기무단투기단속건수



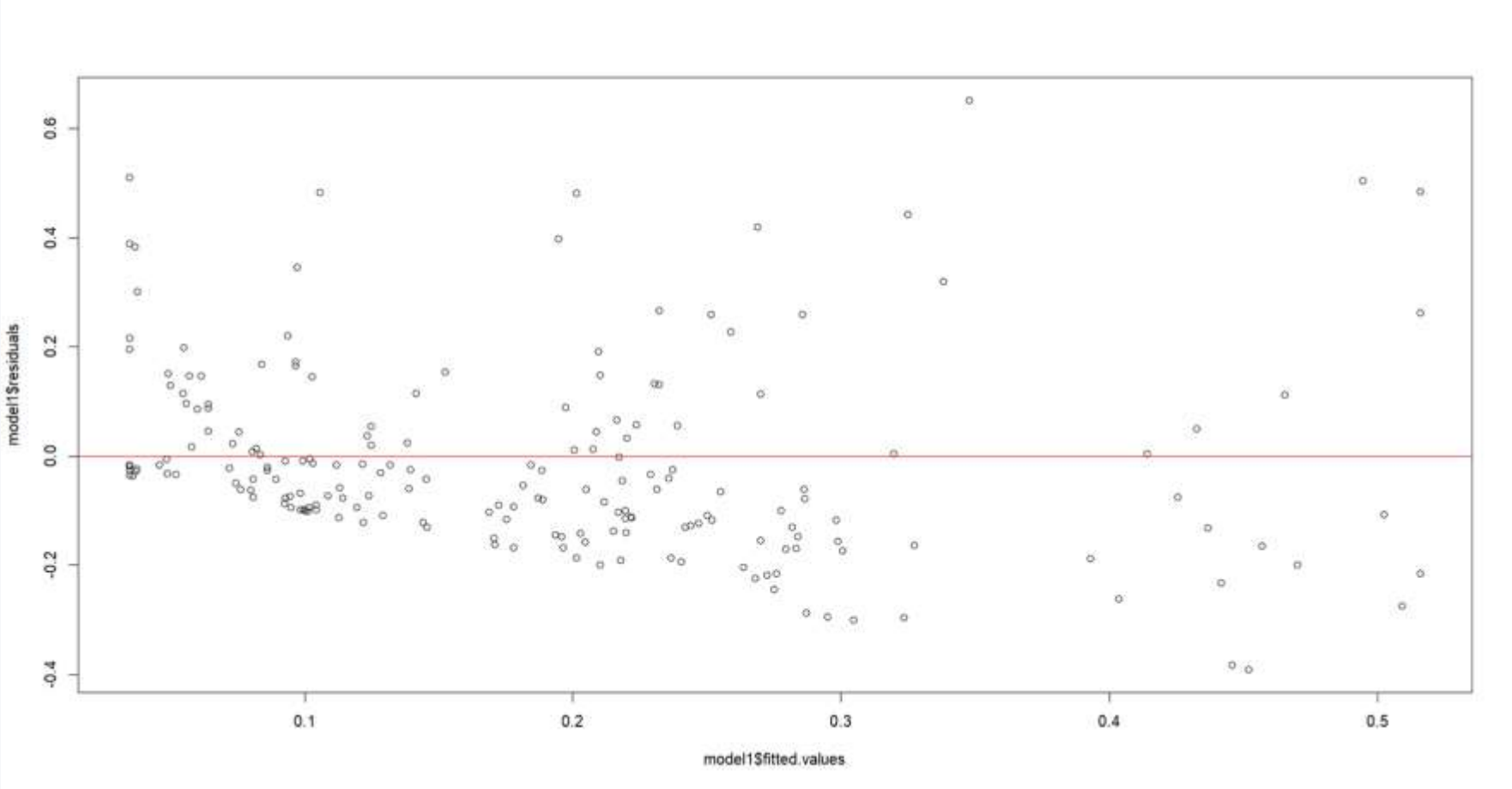

곡선이 한쪽에 치우쳐저서 정규성이 부족한것으로 확인된다.

## 2. 가로등과 쓰레기무단투기단속건수



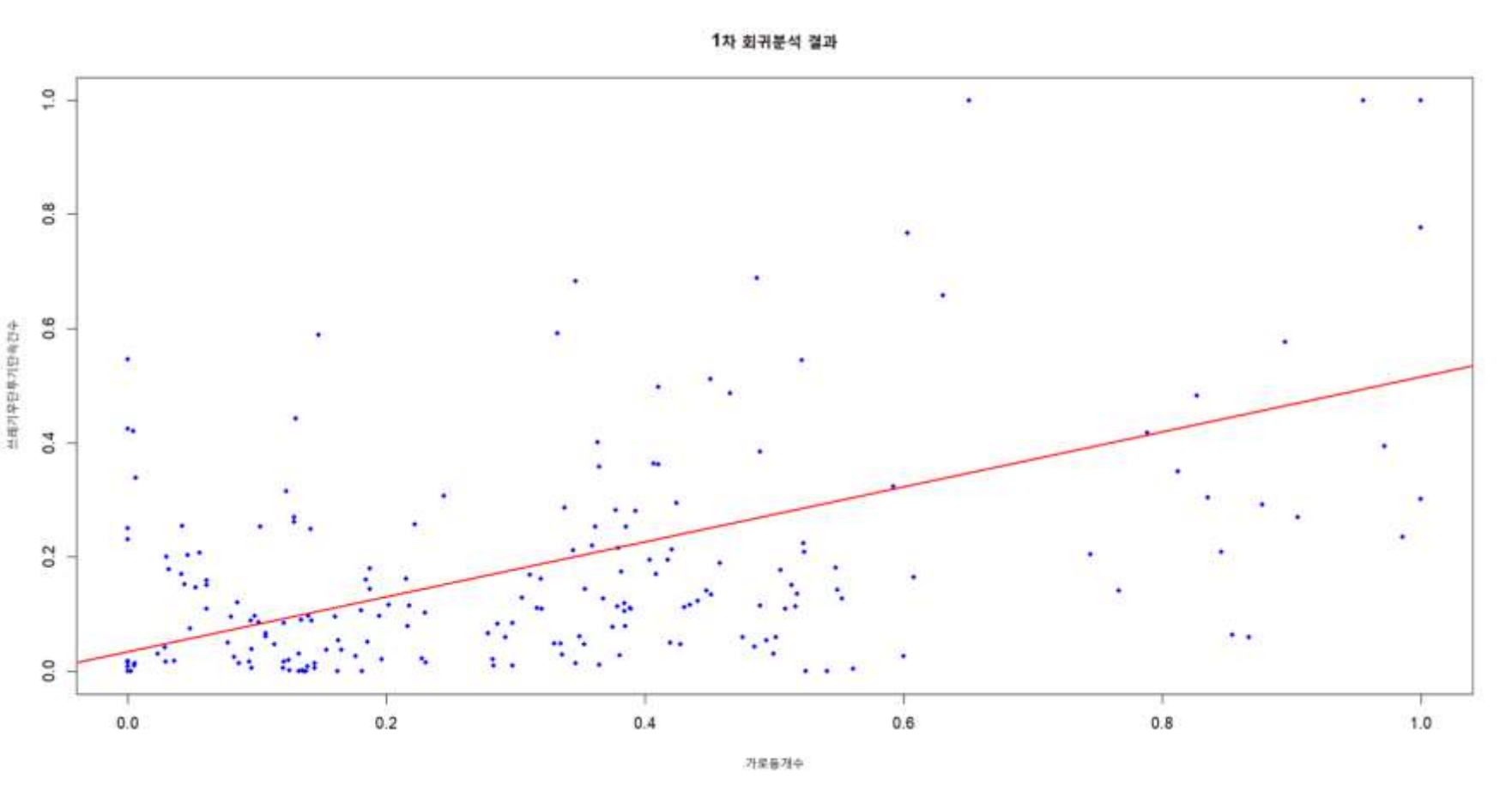

Q-Q plot 확인 결과 직선형이 아니다.

# 2. 가로등과 쓰레기무단투기단속건수




등분산성이 좋지않다.

# 2. 데이터분석 – 가로등개수과 쓰레기무단투기단속건수 회귀분석 summary



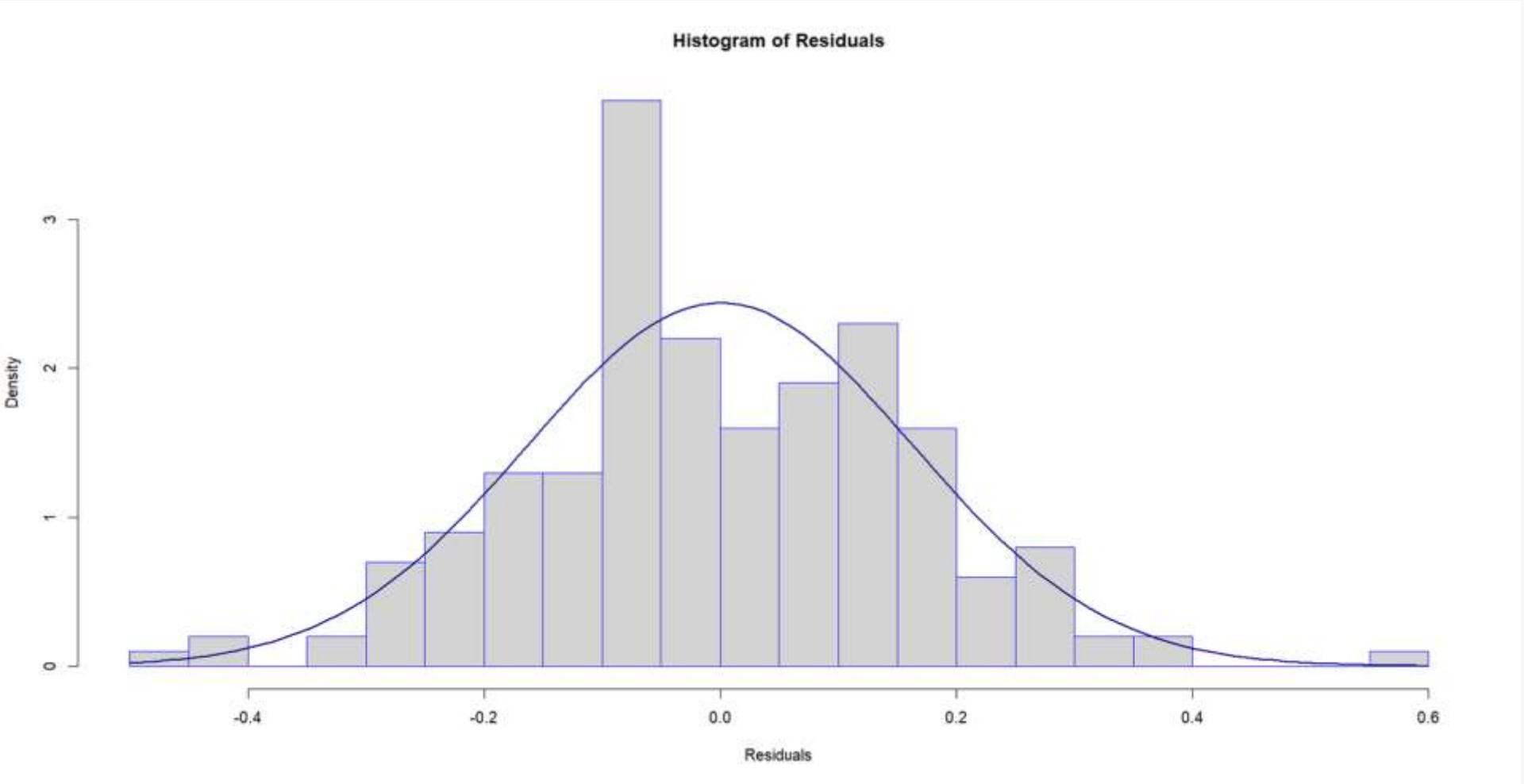
# 2. 데이터분석 – 가로등개수과 쓰레기무단투기단속건수 회귀분석 summary

```
lm(formula = 쓰레기무단투기단속건수 ~ 가로등개수,
    data = data_9)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39166 -0.11633 -0.04139  0.07073  0.65240

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03487    0.02186   1.595    0.112
가로등개수    0.48107    0.05075   9.480   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1915 on 198 degrees of freedom
Multiple R-squared:  0.3122,    Adjusted R-squared:  0.3087
F-statistic: 89.87 on 1 and 198 DF,  p-value: < 2.2e-16
```

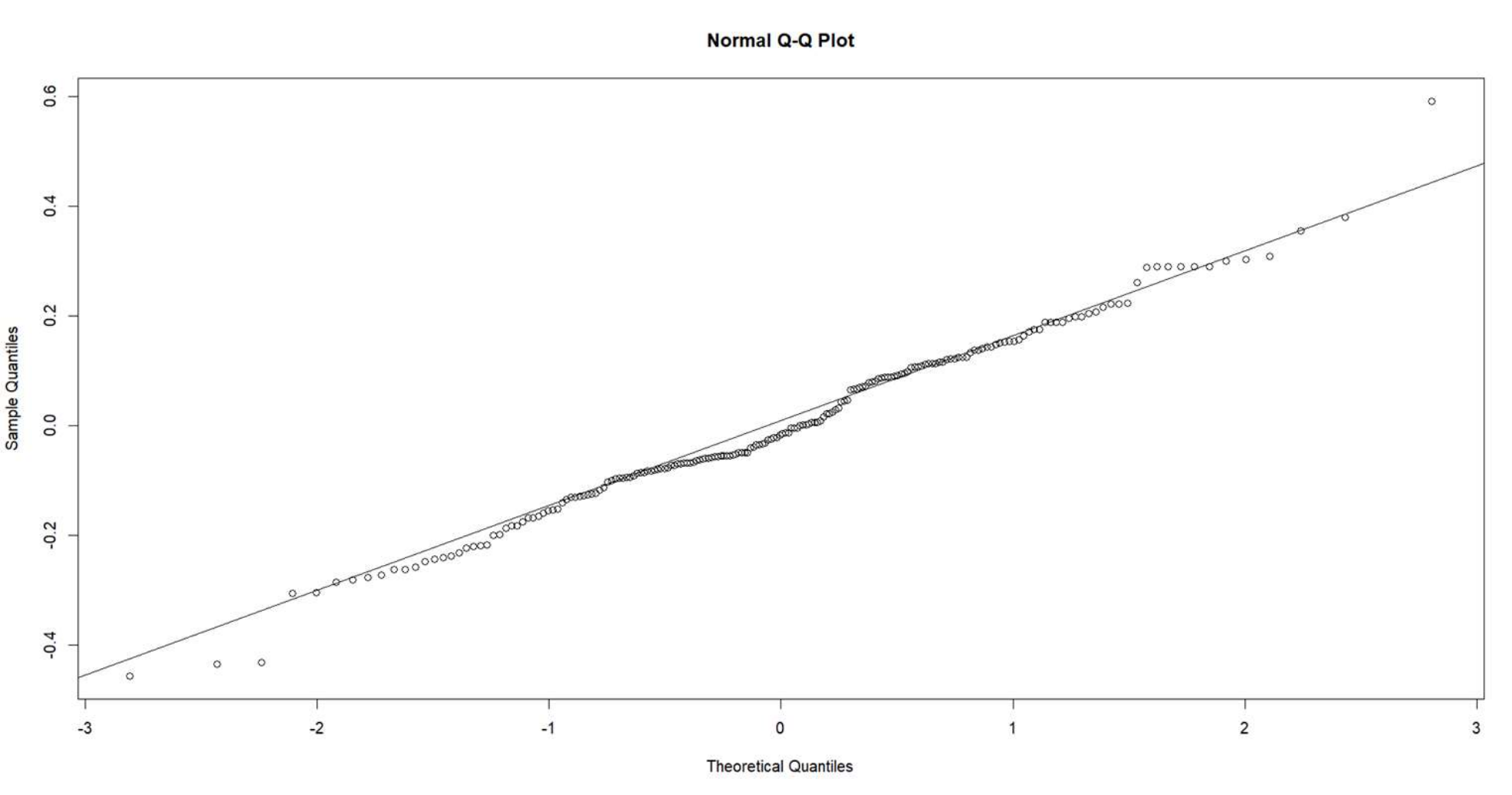
쓰레기무단투기단속건수와
가로등개수를 분석해본 결과, p-value
값이 2.2e-16 이므로 0.05보다 매우
작다. 이것은 **통계적으로 유의하다고**
해석할 수 있다.

R-squared 즉 결정계수는 30.87%만큼
영향을 끼친다는 뜻

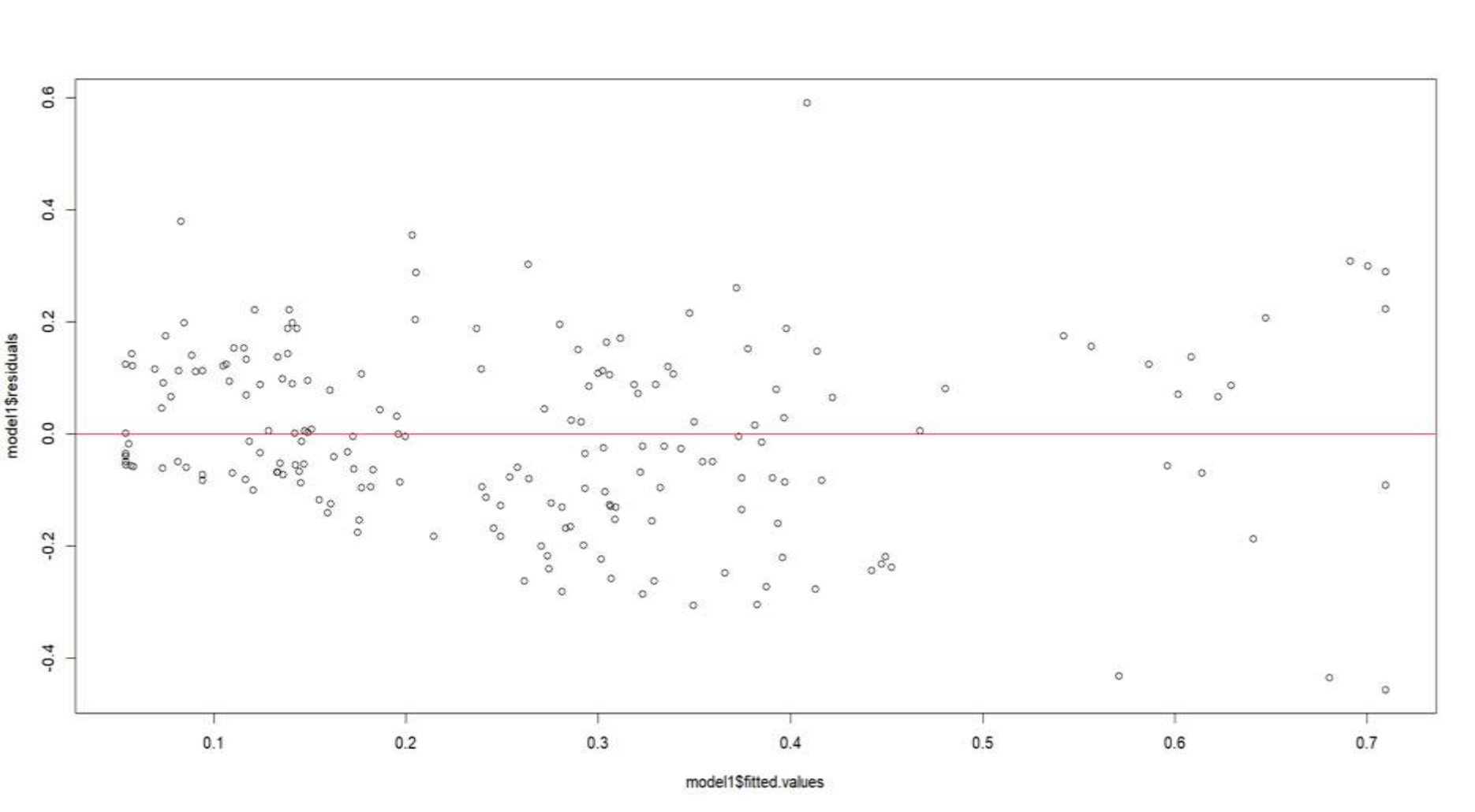
# 2. 가로등과 일일쓰레기총배출량



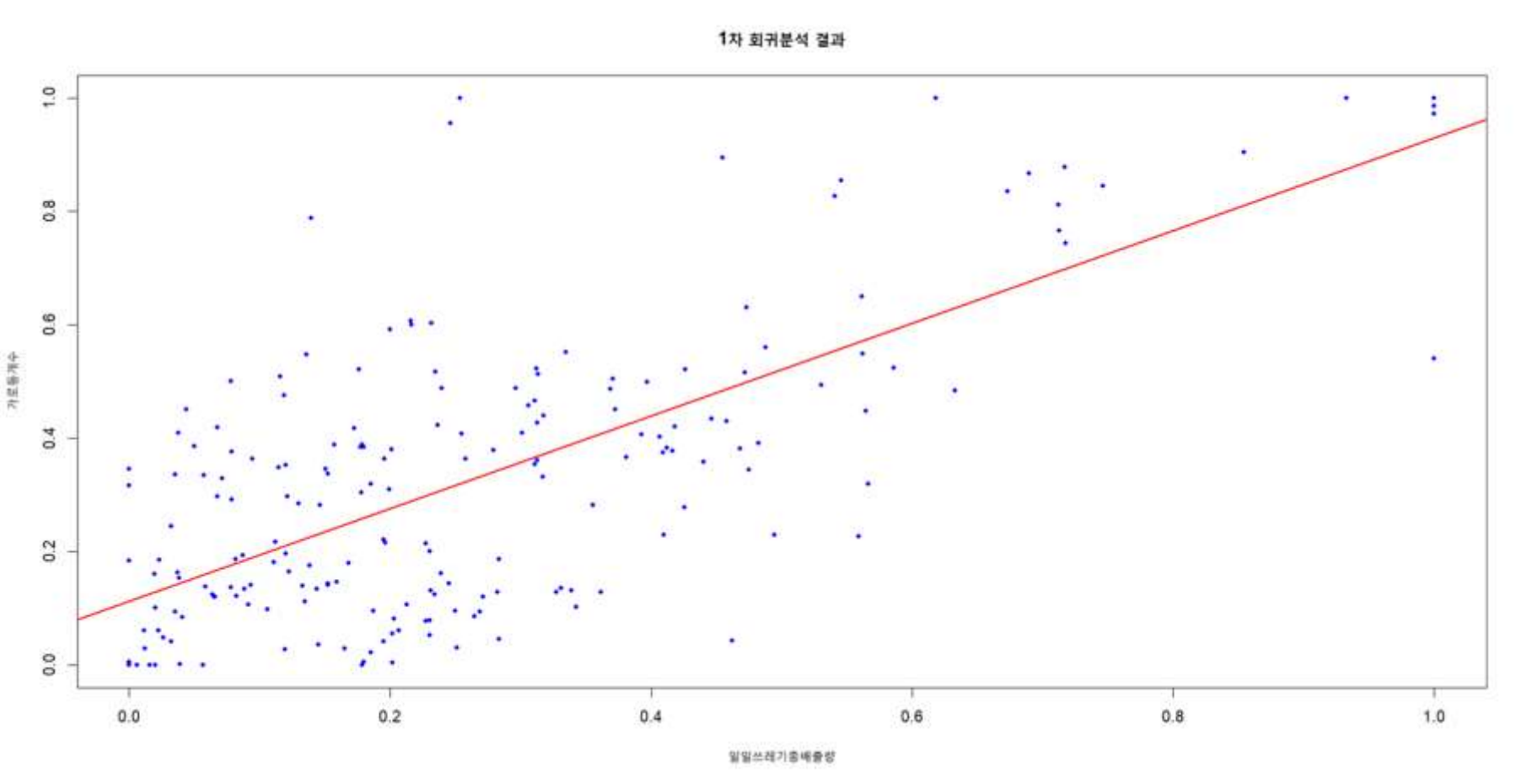
# 2. 가로등과 일일쓰레기총배출량



# 2. 가로등과 일일쓰레기총배출량



# 2. 데이터분석 – 가로등개수과 일일쓰레기총배출량 회귀분석 summary



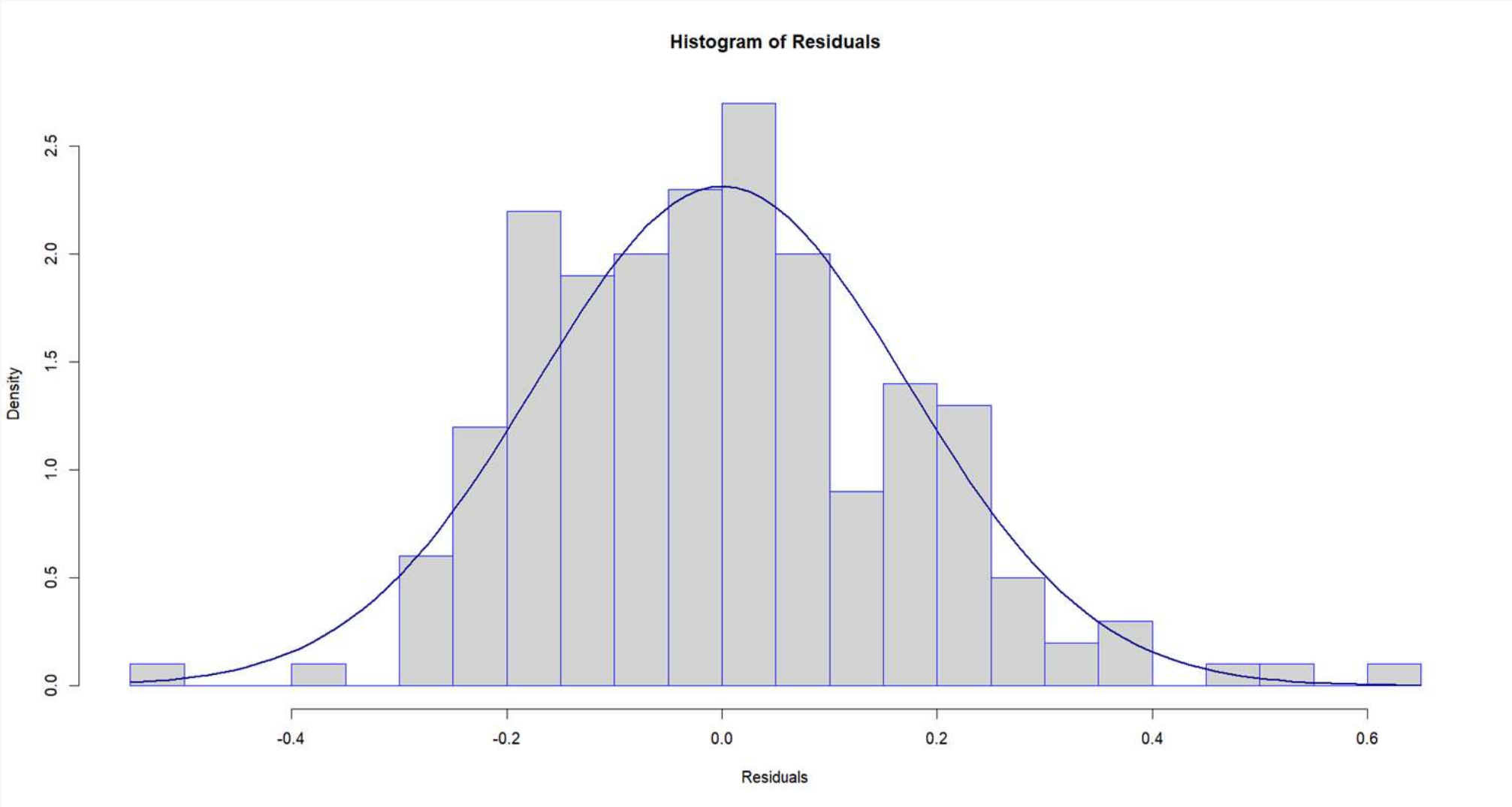
## 2. 데이터분석 – 가로등개수과 일일쓰레기총배출량 회귀분석 summary

```
lm(formula = 가로등개수 ~ 일일쓰레기총배출량, data = data_9)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44655 -0.11395 -0.02281  0.11135  0.68001

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.11302    0.01974   5.724 3.81e-08 ***
일일쓰레기총배출량  0.81644    0.05408  15.097  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1828 on 198 degrees of freedom
Multiple R-squared:  0.5351,    Adjusted R-squared:  0.5328
F-statistic: 227.9 on 1 and 198 DF,  p-value: < 2.2e-16
```

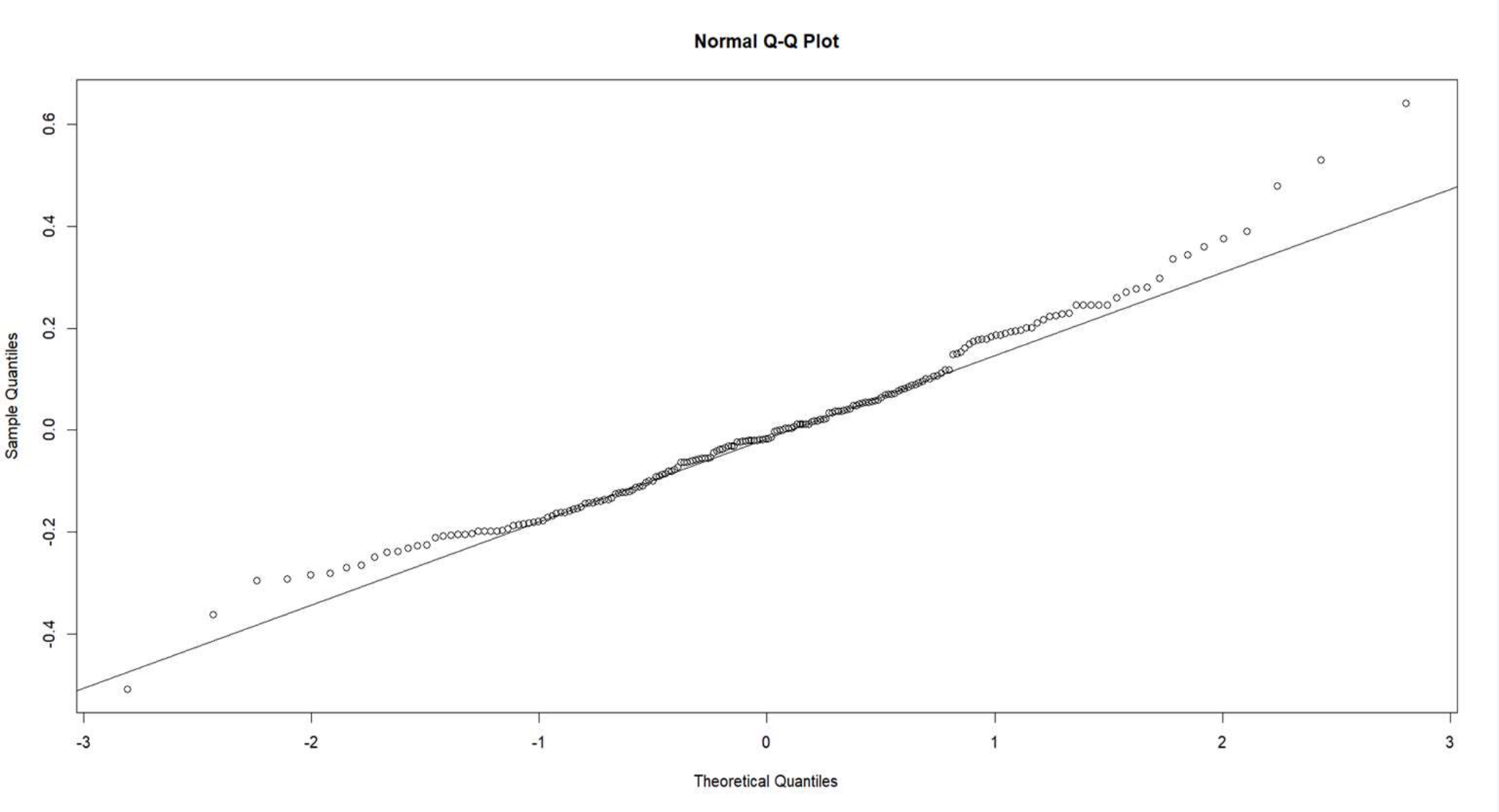
일일쓰레기배출량과 가로등개수를 분석해본 결과, p-value 값이 2.2e-16 이므로 0.05보다 매우 작다. 이것은 **통계적으로 유의하다고** 해석할 수 있다.

R-squared 즉 결정계수는 53.28%만큼 영향을 끼친다는 뜻

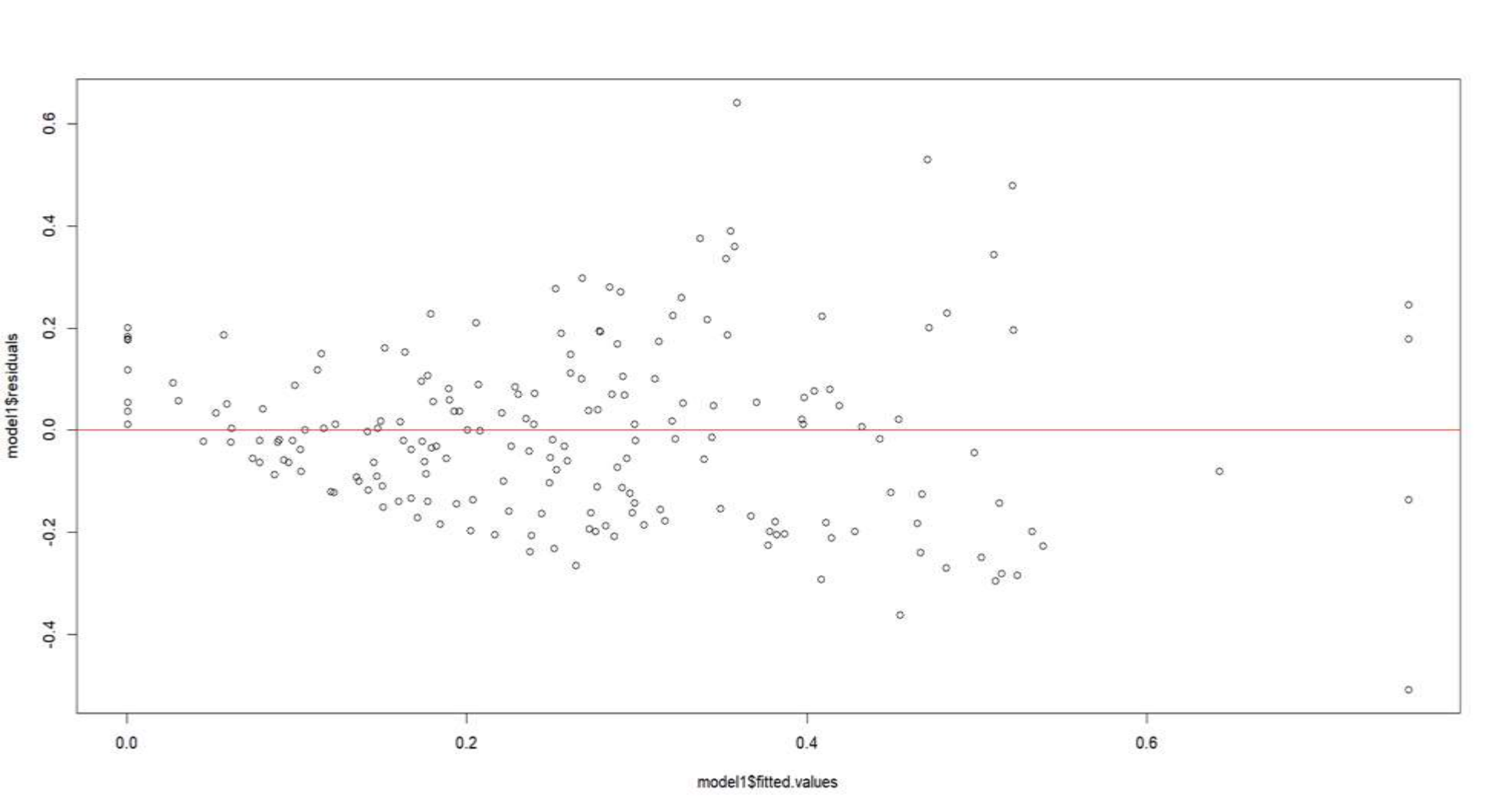
# 2. 일일쓰레기배출량과 총범죄수



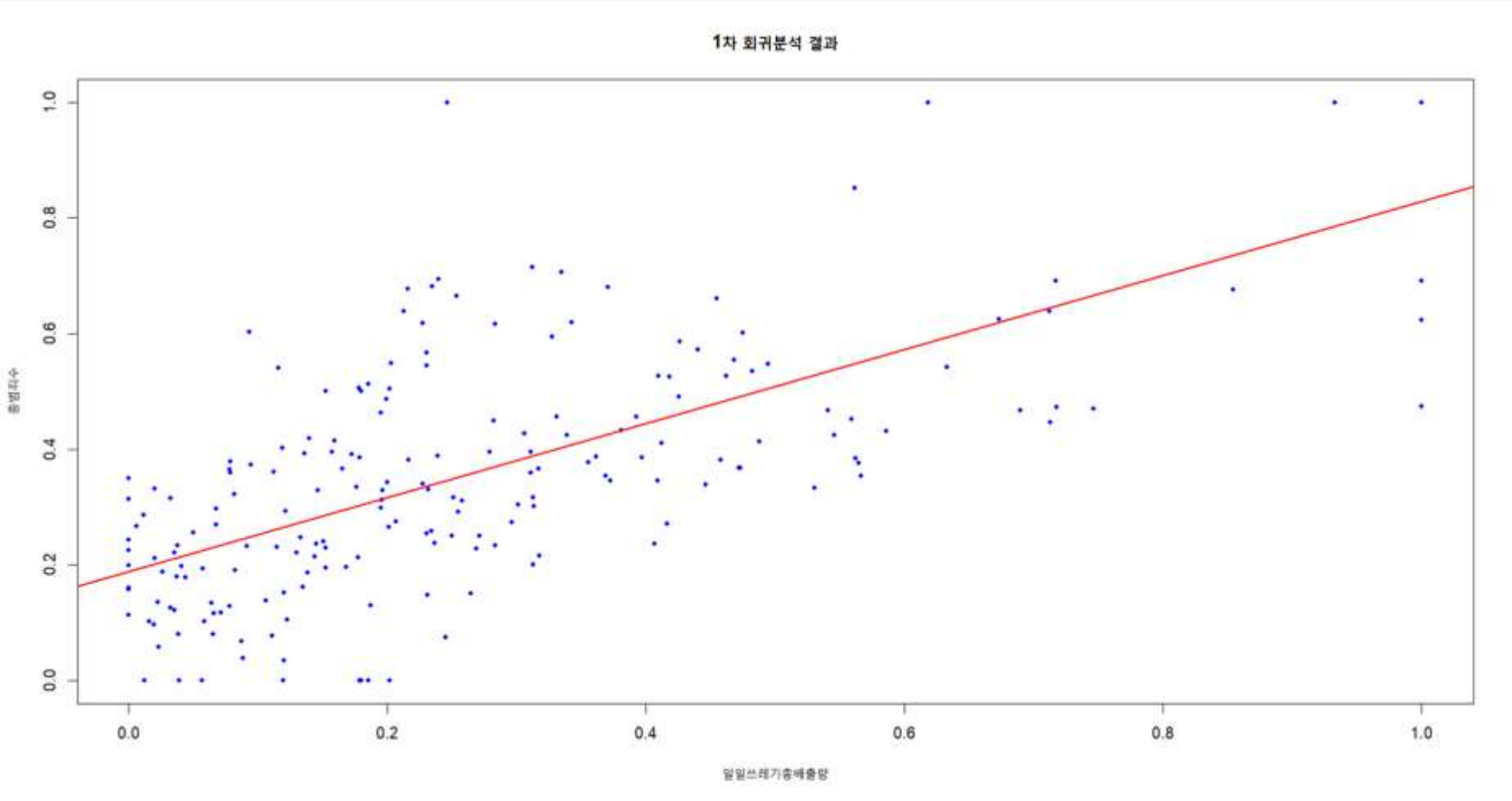
# 2. 일일쓰레기배출량과 총범죄수



# 2. 일일쓰레기배출량과 총범죄수



# 2. 데이터분석 – 일일쓰레기배출량과 총범죄수 회귀분석 summary



# 2. 데이터분석 – 일일쓰레기배출량과 총범죄수 회귀분석 summary

```
lm(formula = 총범죄수 ~ 일일쓰레기총배출량, data = data_9)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35380 -0.10633 -0.02662  0.10124  0.65390

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.18882    0.01721   10.97   <2e-16 ***
일일쓰레기총배출량  0.64002    0.04715   13.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1594 on 198 degrees of freedom
Multiple R-squared:  0.482,     Adjusted R-squared:  0.4794
F-statistic: 184.3 on 1 and 198 DF,  p-value: < 2.2e-16
```

일일쓰레기배출량과 총범죄수를 분석해본 결과, p-value 값이 2.2e-16 이므로 0.05보다 매우 작다. 이것은 **통계적으로 유의하다고** 해석할 수 있다.

R-squared 즉 결정계수는 47.94%만큼 영향을 끼친다는 뜻

## 2. 데이터분석결과

- 전체적인 분석으로 비교해 본 결과, p-value 값은 의미있는 결과 였지만 전체적으로 상관관계가 있다고 볼 수 없었다. 그래서 인과관계를 포함하여 요소가 무엇이 있는지 고민해보고 분석해보기로 했다.

# 2. 다른 요인 조사

미주중앙일보 2021.12.17.

**범죄 예방** 위해 **가로등** 1만개 추가 설치

애틀랜타 시가 **범죄** 발생을 줄이기 위해 **가로등** 1만개를 추가로 설치한다. 키이샤 랜스 바텀스 애틀랜타 시장은 16일 크리스 워맥 조지아 파워 회장, 조시 로완 교통담당 커미셔너 등이 참석한 가운데 "하나의 애틀랜타...

TV조선 PICK 2020.01.08. 네이버뉴스

**'가로등'만 설치해도 범죄율** 감소...공동현관 잠금장치도 효과

경찰은 "골목길 등은 **가로등**이나 보안등 같은 조명과 폐쇄회로(CC)TV 설치가 **범죄율**을 줄였고, 공동 주택의 경우 정문 출입 통제 장치가 **범죄율**을 줄였다"고 분석...

**가로등**만 있어도 **범죄율** 16% 줄어...비상벨은 효... 라이센스뉴스 2020.01.08.

아시아경제 2019.03.19. 네이버뉴스

**"CCTV 설치 지역 범죄율** 16% 감소...경찰활동 병행 시 더욱 효과...

국제사례 비교를 통해 폐쇄회로(CC)TV 설치 지역이 미설치 지역보다 실제 **범죄율**이 줄었다는 연구 결과가 나왔다. CCTV가 절도 등 재산범죄에는 효과적이나 폭...


세계일보 2019.08.27. 네이버뉴스

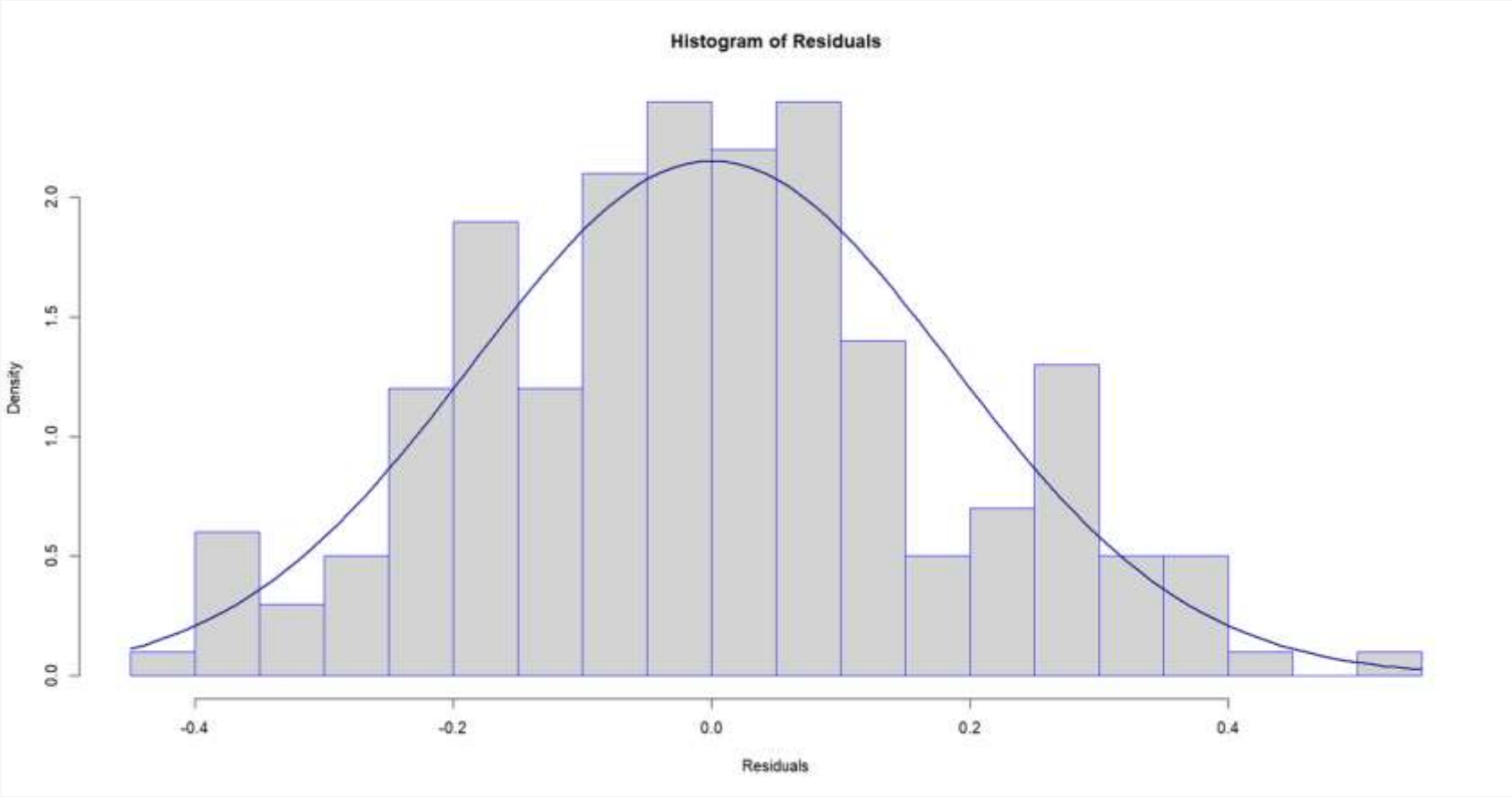
**설치만으로도 범죄율** 뚝... **CCTV** 사상 첫 100만대 넘어서

27일 행정안전부에 따르면 공공기관이 설치한 CCTV만 총 103만대로 전년대비 8.2%(7만8618대) 증가한 것으로... 더불어 CCTV의 효율적 운영과 사건사고에 신속...

총 범죄수와 관련있는 다른 요인을 조사하여 관련이 있는 요소와 상관분석을 진행했다.

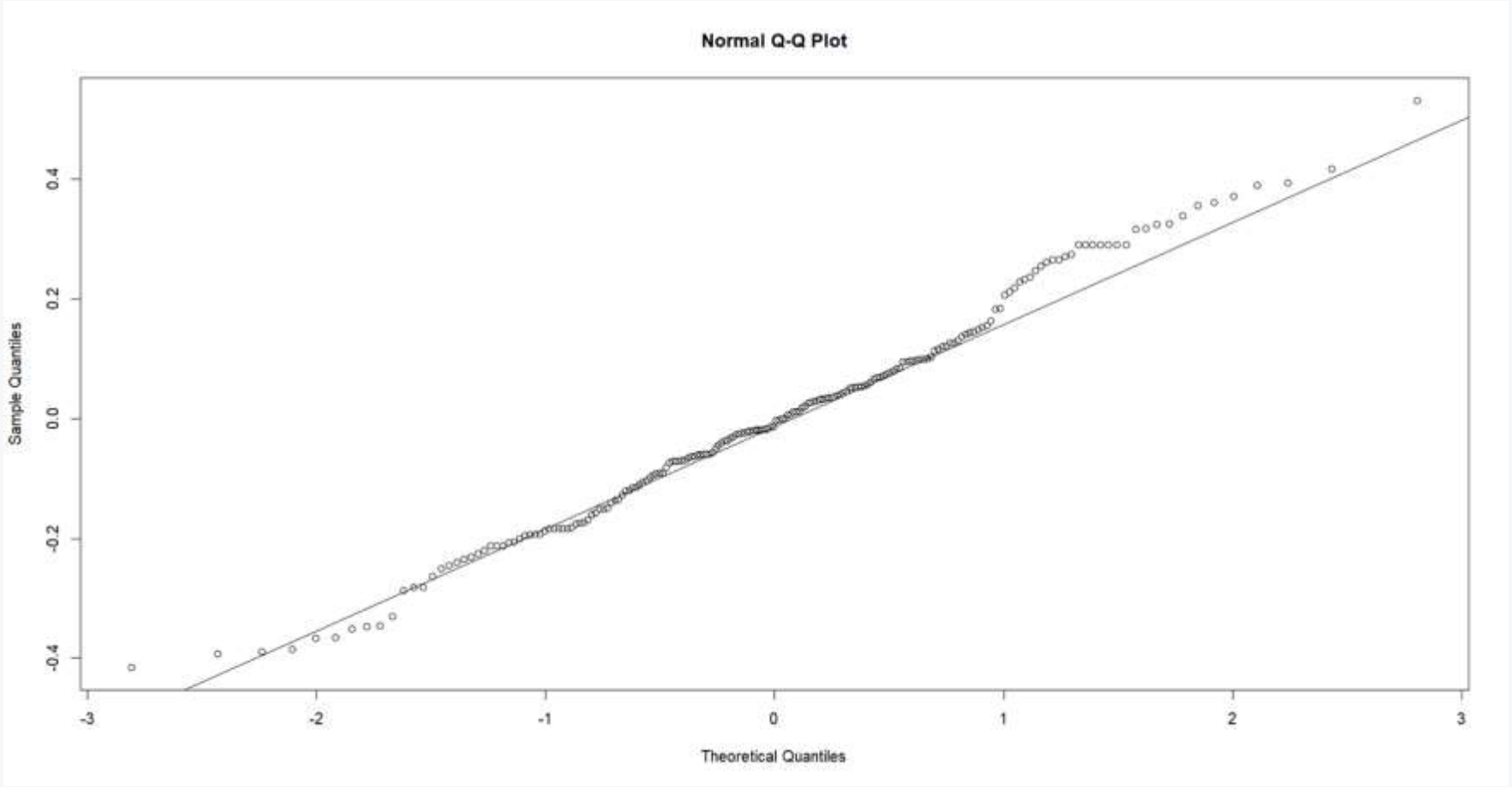
# 3. CCTV 와 범죄 발생 건수 - 상관관계

# 3. CCTV 와 범죄 발생 건수 - Histogram



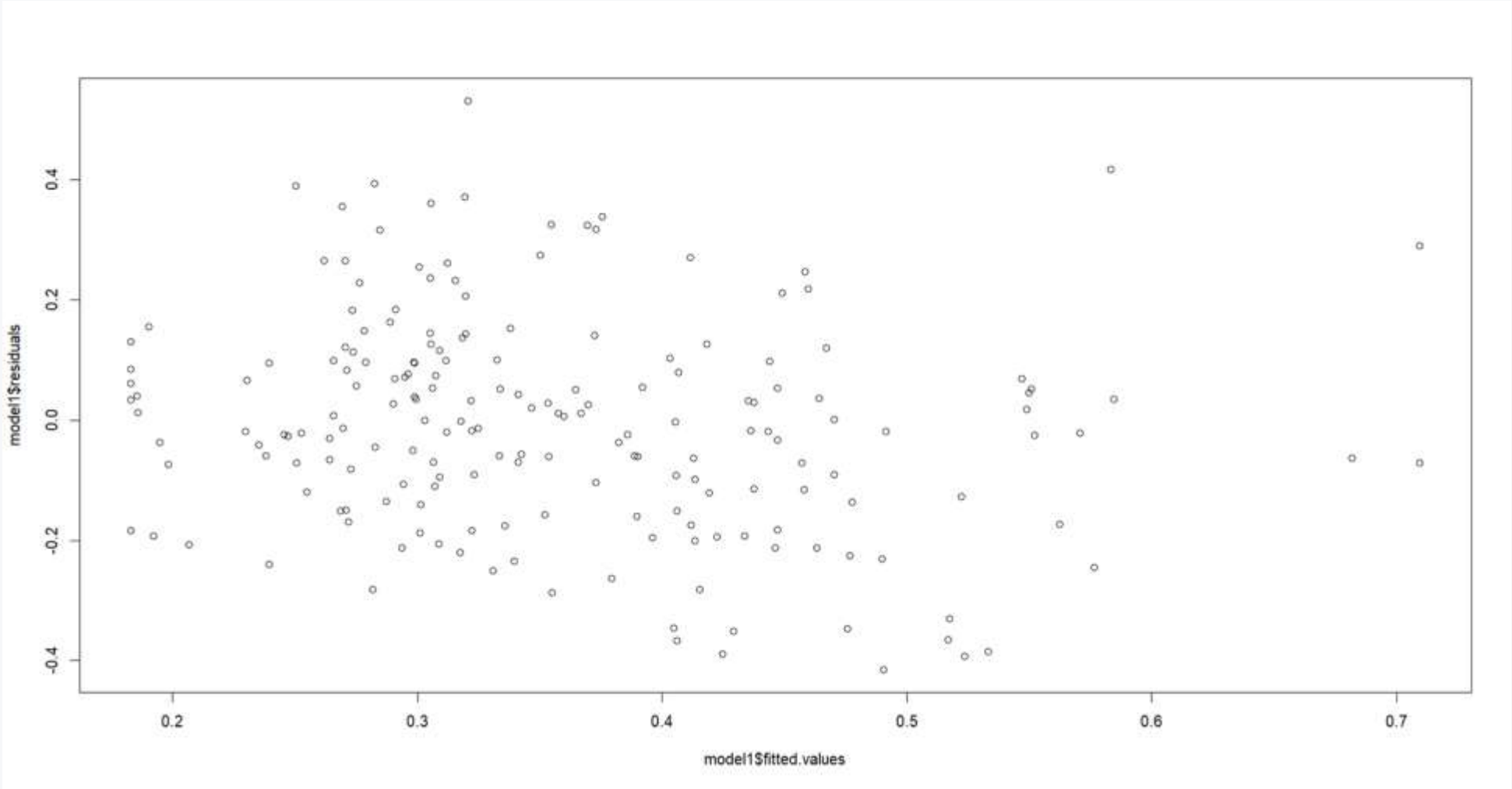

Histogram을 통해 데이터의 분포를 확인한 결과
곡선이 한쪽에 치우치지 않고 중앙에 위치하고 있어 정규성을 띄는 것으로 확인된다.

# 3. CCTV 와 범죄 발생 건수 - Q-Q plot




Q-Q plot 확인 결과 직선형을 띄고 있으며,
선 근처에 데이터가 분포하고 있어 정규분포에 가깝다.

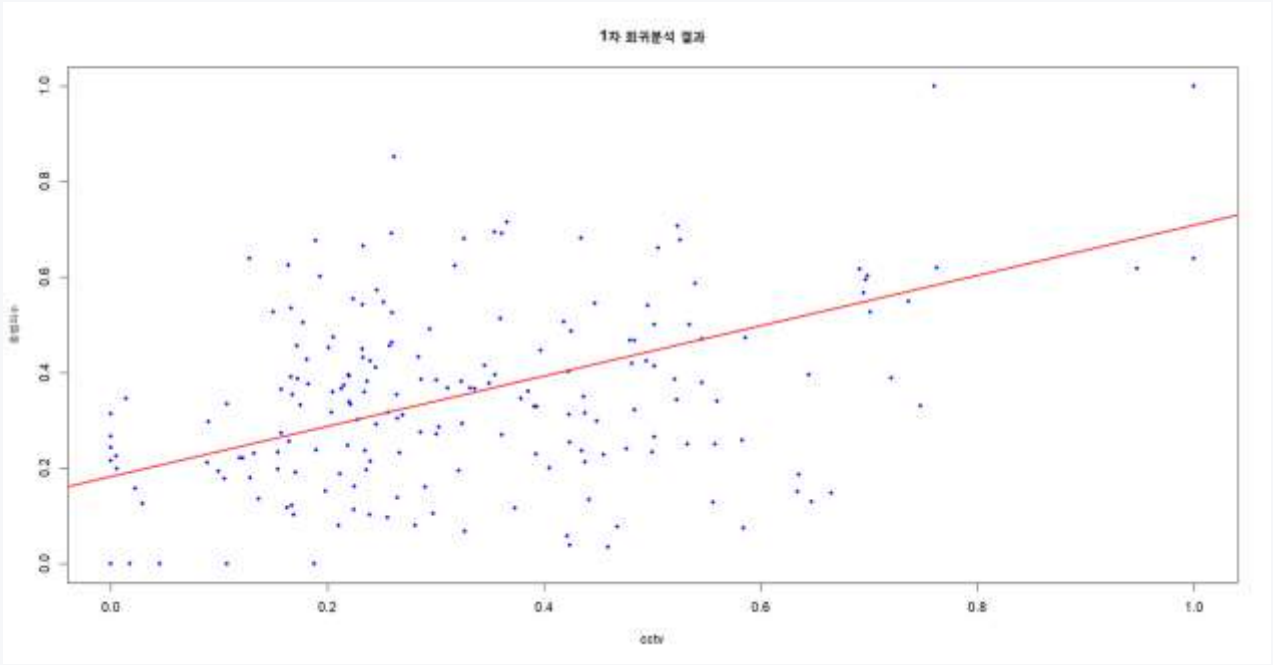
# 3. CCTV 와 범죄 발생 건수 - 산점도




하지만 역시 등분산성이 문제였다.

# 3. CCTV 와 범죄 발생 건수 - 회귀분석

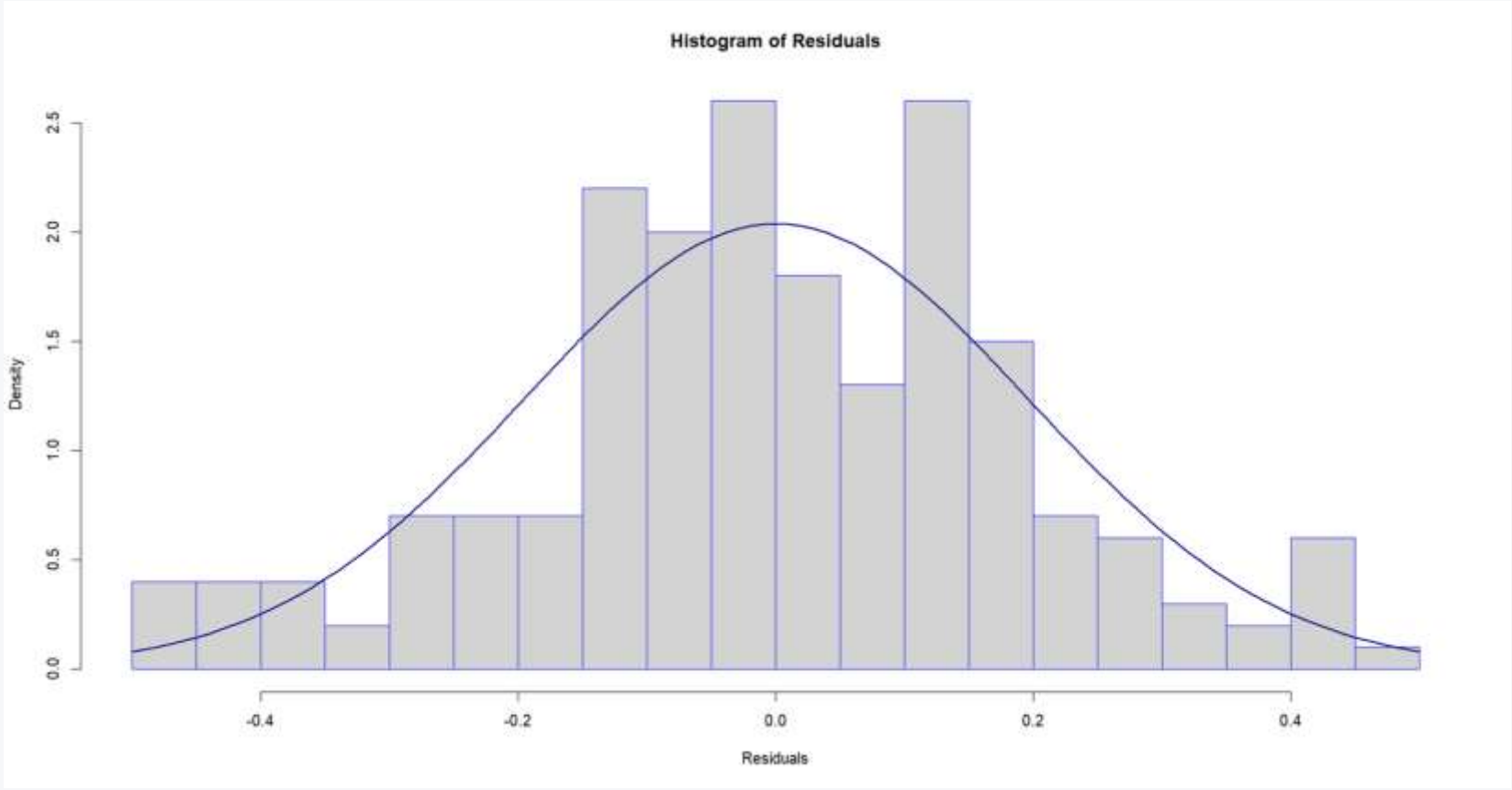
```
lm(formula = 총범죄수 ~ cctv, data = data_9)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41547 -0.12871 -0.00769  0.10135  0.53139

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.18305    0.02389   7.662    8e-13 ***
cctv         0.52646    0.05761   9.139   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1857 on 198 degrees of freedom
Multiple R-squared:  0.2967,    Adjusted R-squared:  0.2931
F-statistic: 83.52 on 1 and 198 DF,  p-value: < 2.2e-16
```




1차 회귀분석 결과 p-value값이 0.05보다 작은값으로 CCTV 수와 총 범죄 수가 상관관계가 있다고 할 수 있다.

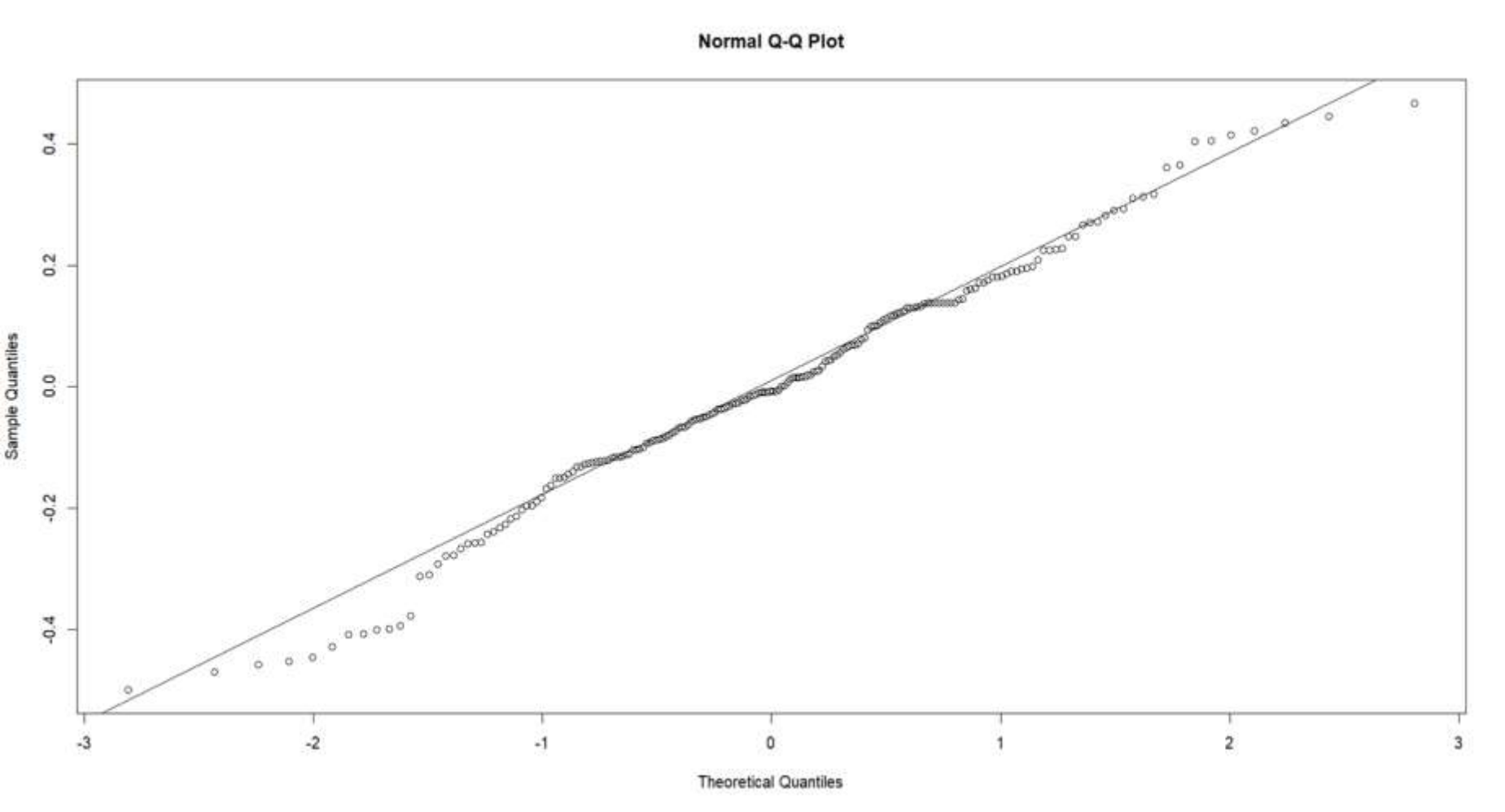
범죄 발생 건수와 CCTV 수의 관계 그래프에 회귀분석 결과 시각화(양의관계)

# 3.가로등개수와 총범죄수의 Histogram



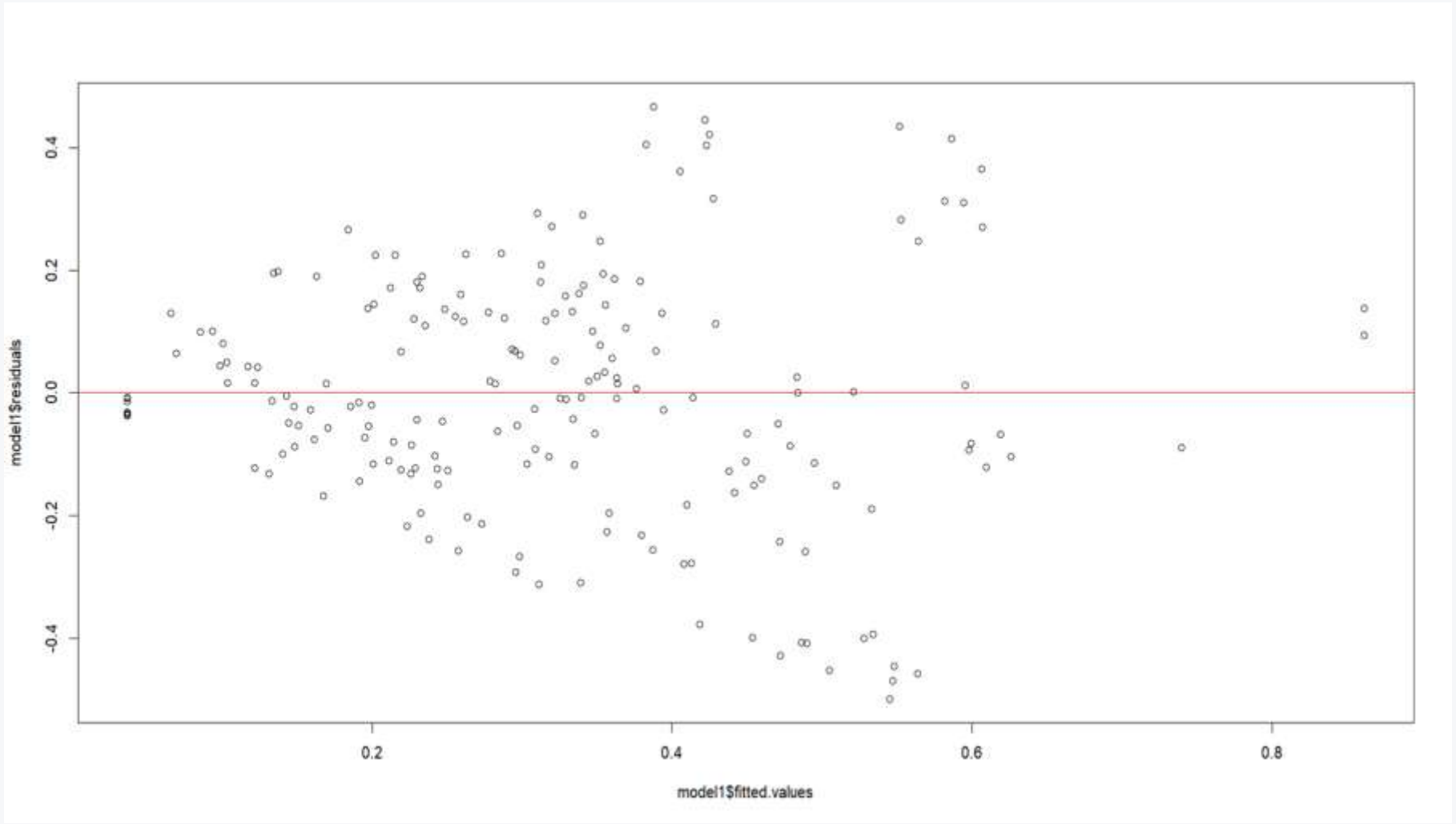

Histogram이 정규성을 따른다면 직선의 곡선의 형태가 중심에 가까워야 한다.

## 3. 가로등개수와 총범죄수의 Q-Q plot



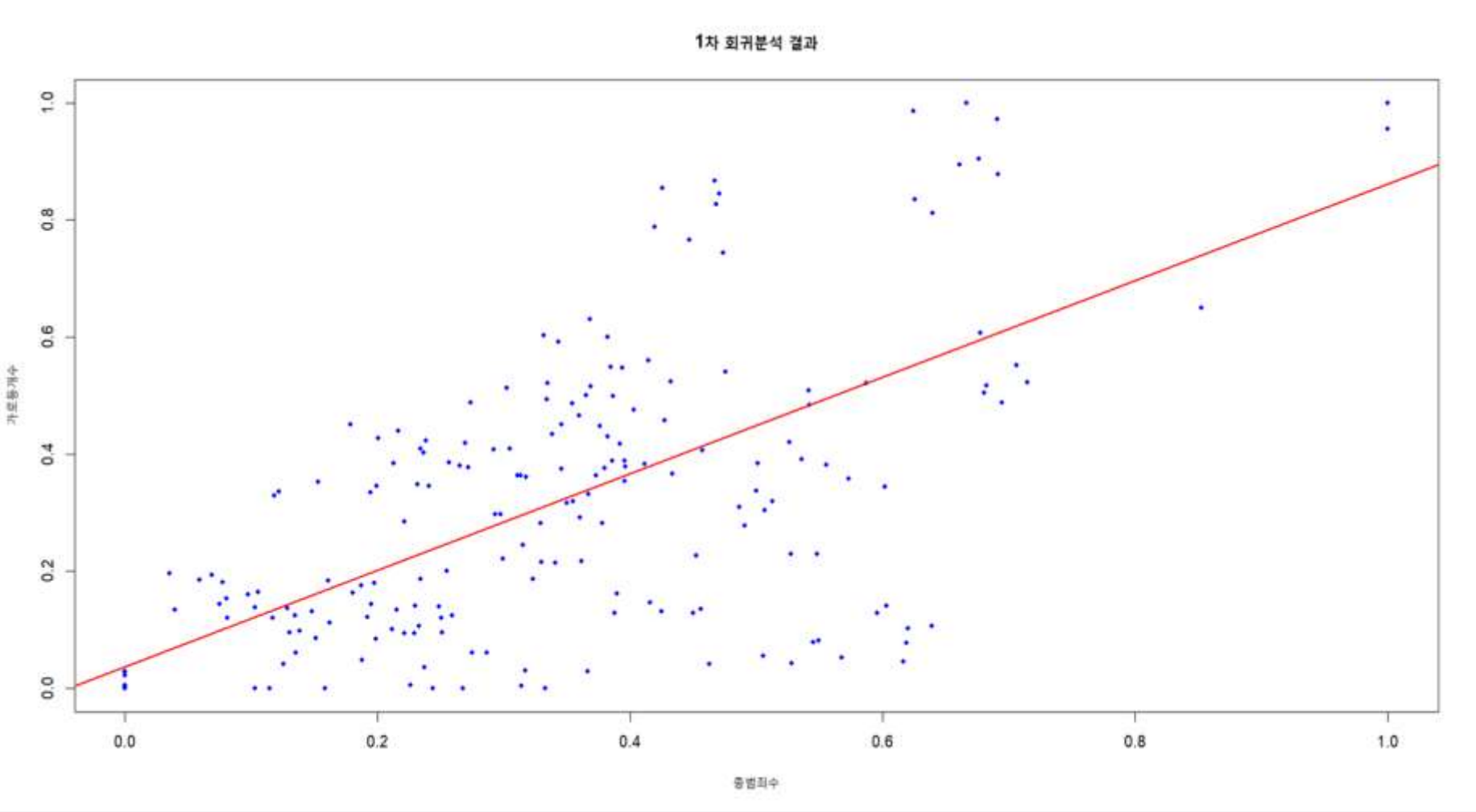

Q-Q plot이 정규성을 따른다면 직선의 형태에 점들이 가까워야한다.

# 3. 가로등개수와 총범죄수의 산점도




등분산성, 독립성

# 3. 가로등개수와 총범죄수의1차 회귀분석 결과




1차 회귀분석 결과

## 3. 가로등개수과 총범죄수 회귀분석 summary

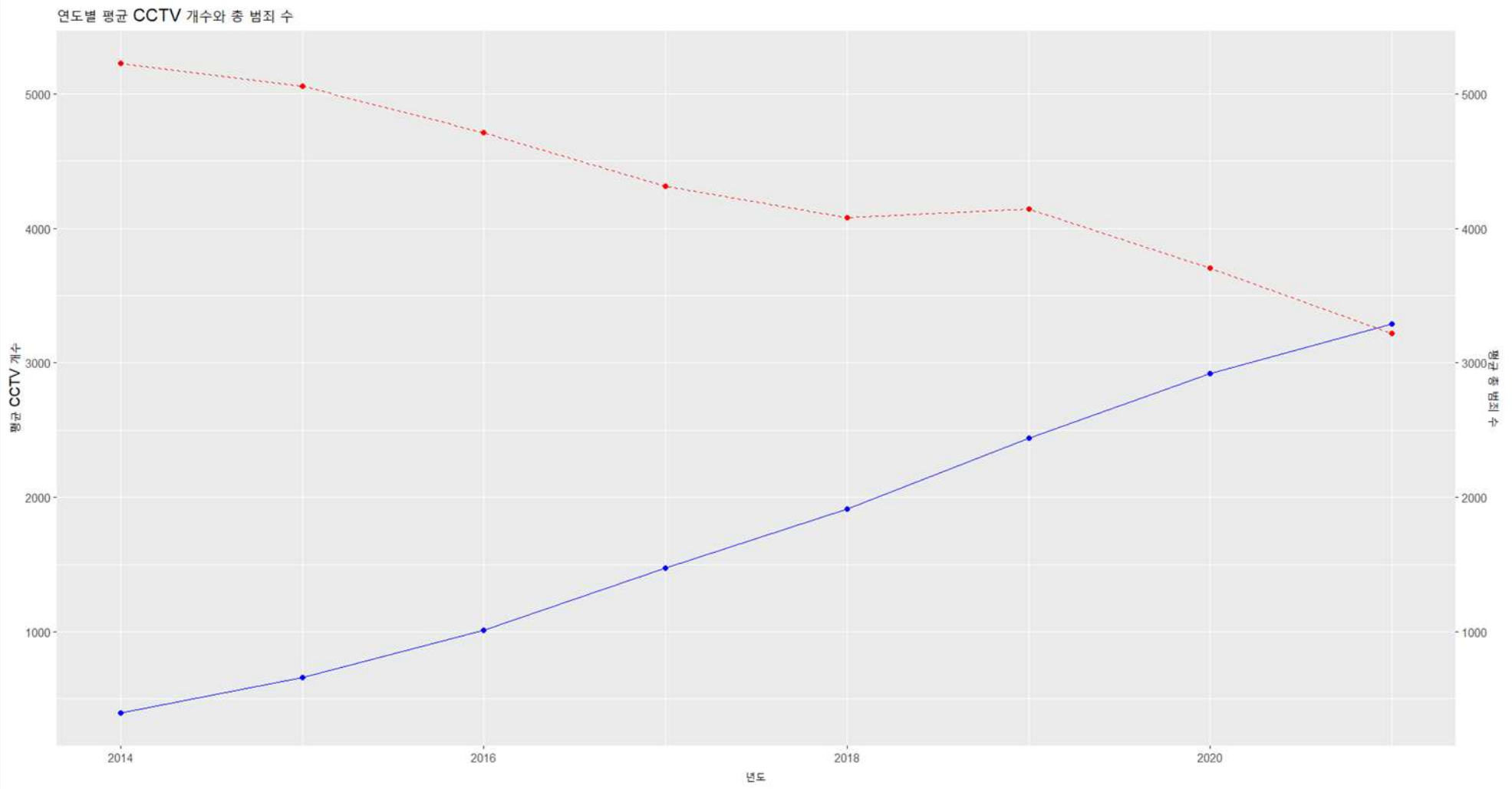
```
lm(formula = 가로등개수 ~ 총범죄수, data = data_9)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49907 -0.11594 -0.00775  0.13700  0.46665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03696    0.02688   1.375    0.171
총범죄수     0.82468    0.06300  13.091   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1963 on 198 degrees of freedom
Multiple R-squared:  0.464,     Adjusted R-squared:  0.4612
F-statistic: 171.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

총범죄수와과 가로등개수를 분석해본 결과, p-value 값이 2.2e-16 이므로 0.05보다 매우 작다. 이것은 **통계적으로 유의하다고** 해석할 수 있다.

R-squared 즉 결정계수는 46.12%만큼 영향을 끼친다는 뜻

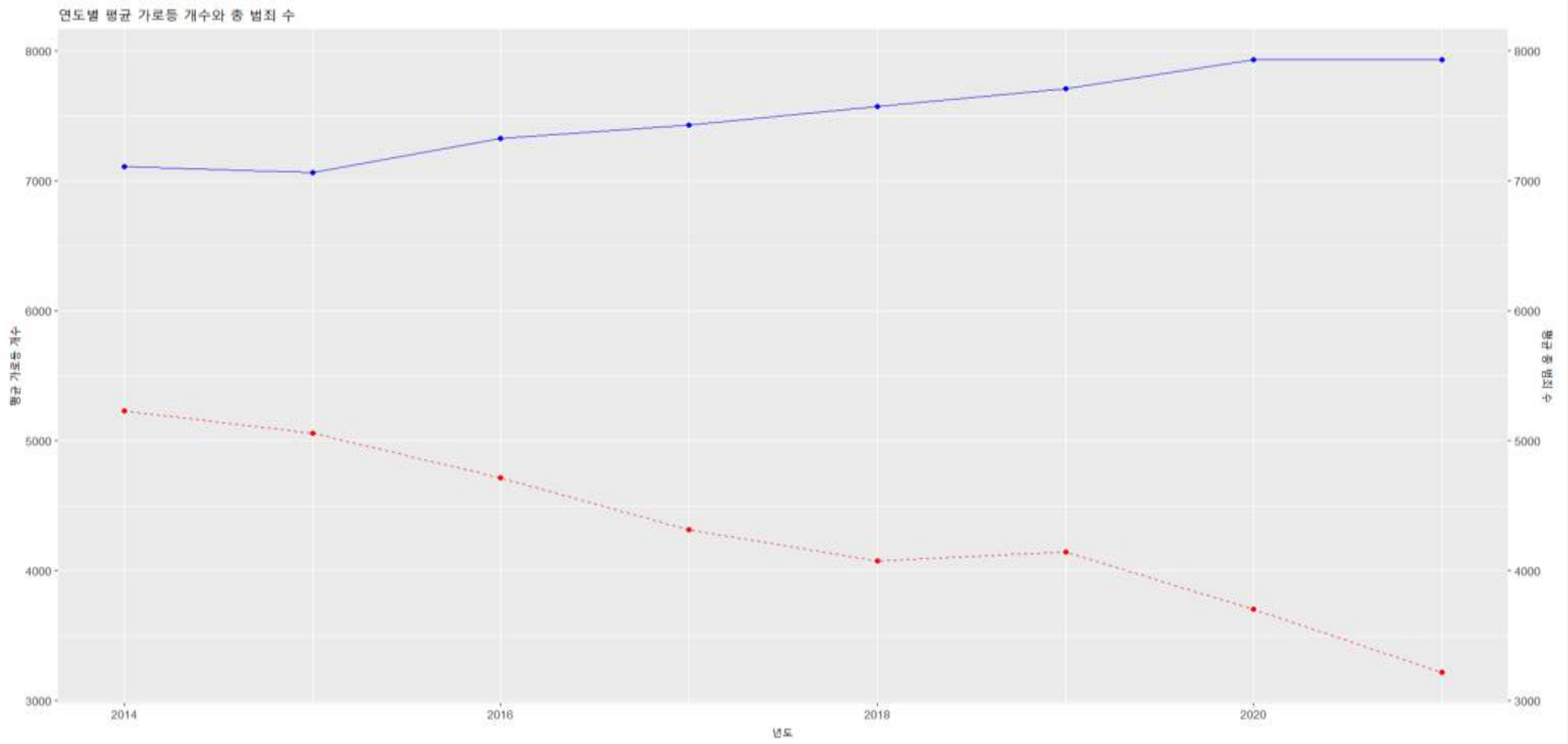
# 3. CCTV 와 범죄 발생 건수 - 연도별 범죄 발생 건수와 CCTV 수 비교

2014년부터 2021년까지 범죄 발생 건수와 CCTV 수 비교



# 4. 가로등과 범죄 발생 건수 - 연도별 범죄 발생 건수와 가로등 수 비교

2014년부터 2021년까지 범죄 발생 건수와 가로등 수 비교

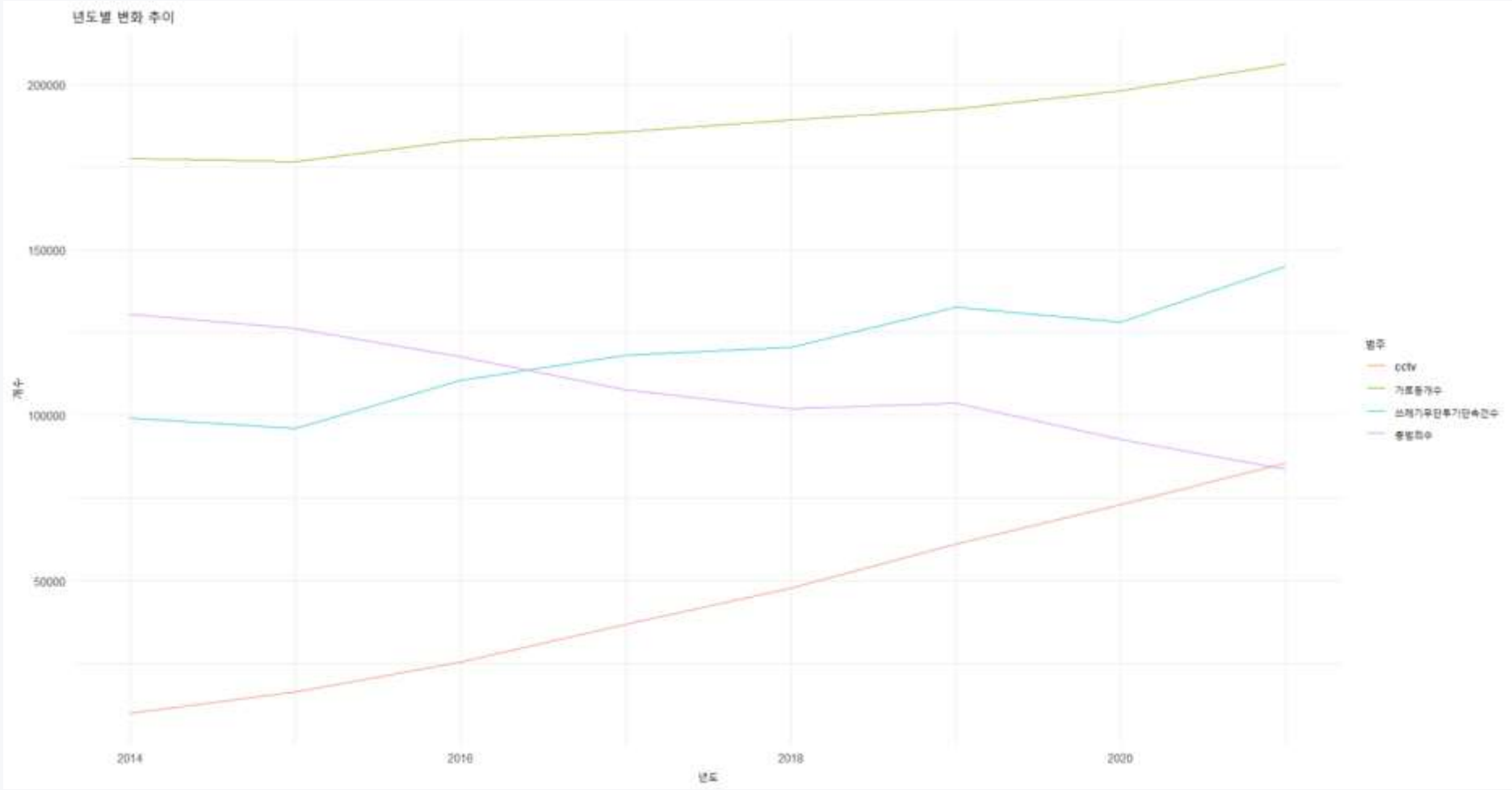


# 5. 비교

2014년부터 2021년까지 범죄 발생 건수와 무단투기 단속 건수 비교

# 5.비교

2014년부터 2021년까지 비교



## 6. 결론

전체적으로 범죄는 감소했지만 인과관계를 신경쓰지않고 데이터분석을 할시에는 cctv가 늘어서 범죄가 늘었다는 등의 잘못된분석을 할수있다.

쓰레기 무단 투기 단속 건수와 범죄수가 반비례관계인것은 경찰인원의 증가나 경찰서의 증가 또는 cctv의 개수증가로 인해 범죄율 감소에 일부 기여했을 수 있음.

회귀분석은 만능이아니며 분석을 하기전에 먼저 선형성, 잔차의 등분산성, 잔차의 정규성을 파악하고 다중공선성을 피해야함 조건을 만족하지 않거나 표본수가 적으면 정통적인 통계방식으로 접근하는것이 좋을수있다.