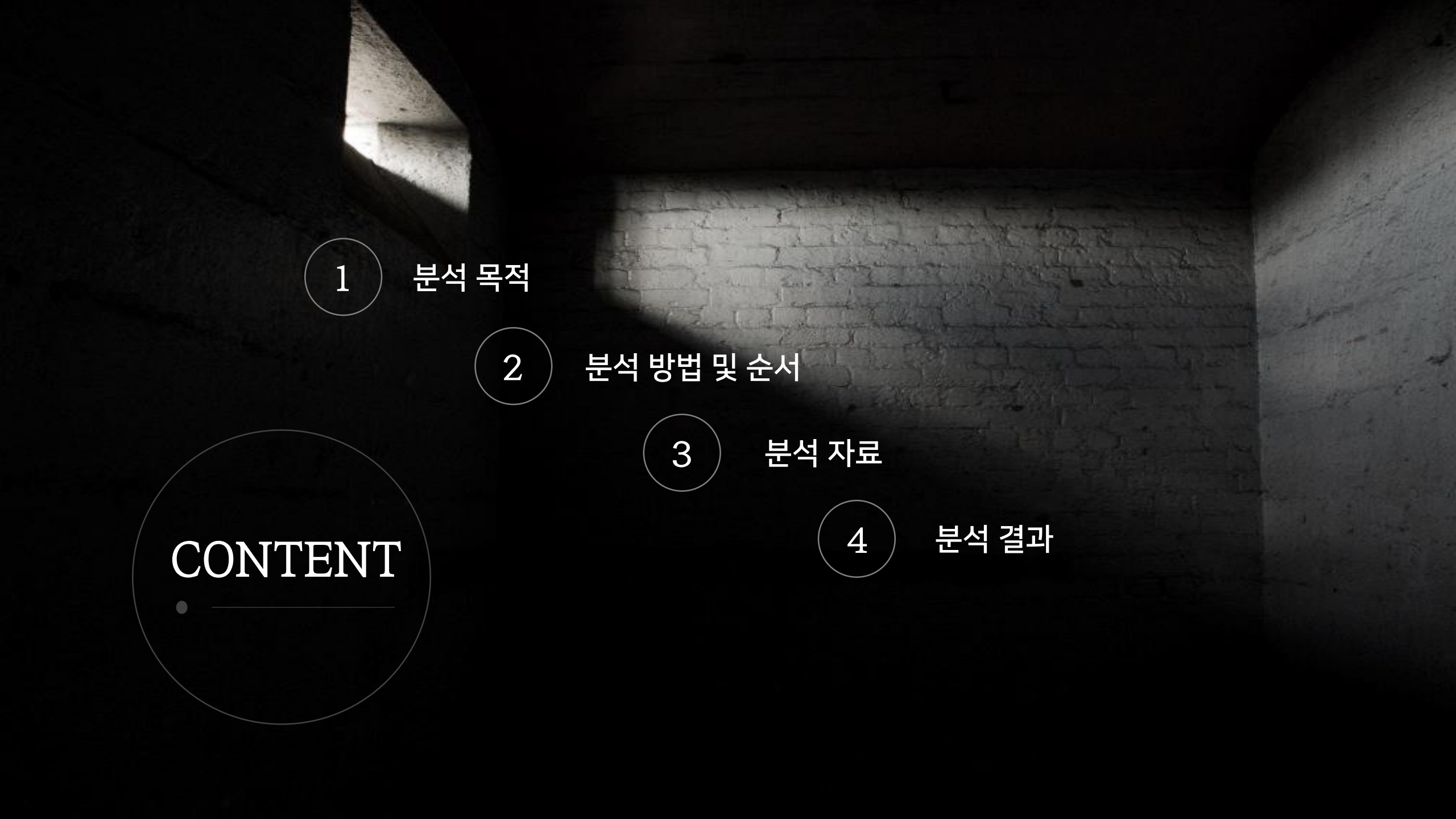
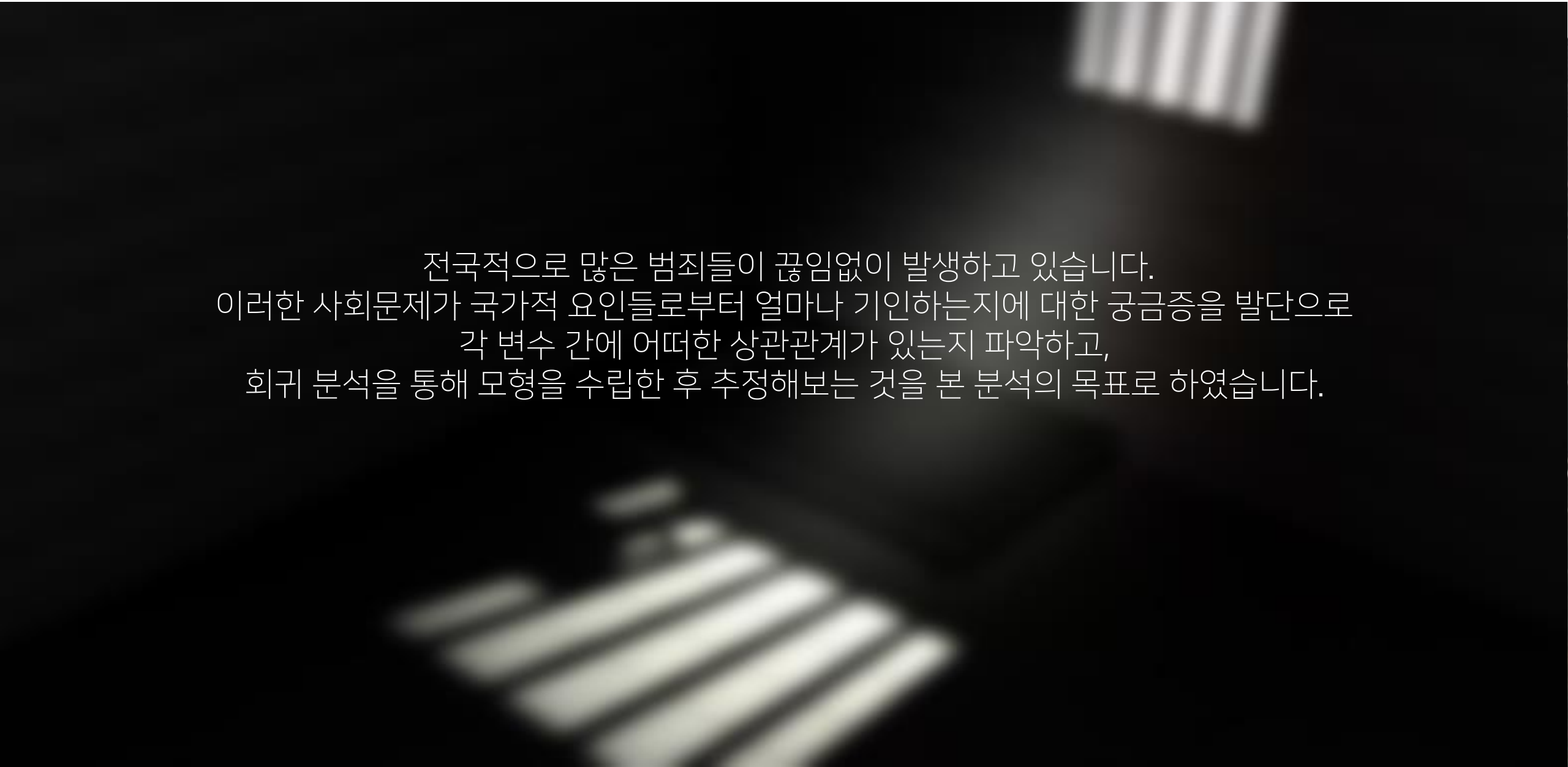
범죄 발생 건수에 대한 회귀분석 및 추정



CONTENT

분석 목적

분석 목적



전국적으로 많은 범죄들이 끊임없이 발생하고 있습니다.
이러한 사회문제가 국가적 요인들로부터 얼마나 기인하는지에 대한 궁금증을 발단으로
각 변수 간에 어떠한 상관관계가 있는지 파악하고,
회귀 분석을 통해 모형을 수립한 후 추정해보는 것을 본 분석의 목표로 하였습니다.

데이터 이해

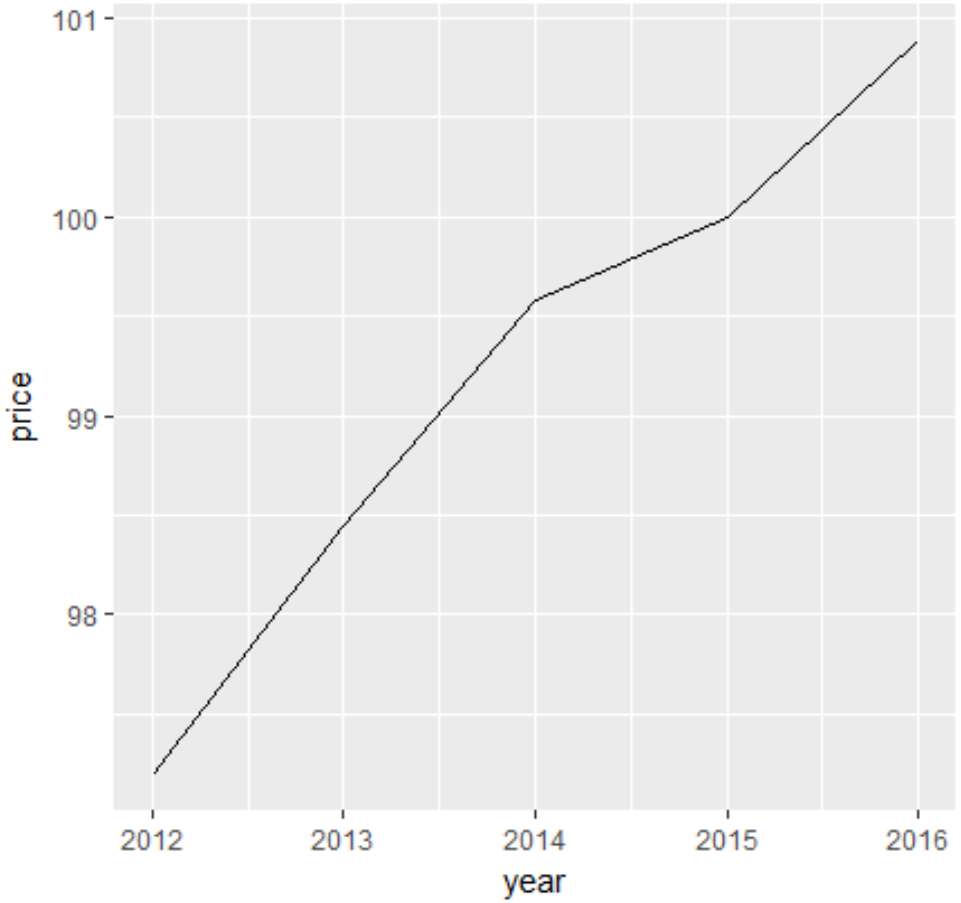
데이터 이해

```
> summary(data[,3:7])
     price          unemployment     population         stress          crime       
 Min.   : 95.91   Min.   :1.600   Min.   :   90.0   Min.   :20.50   Min.   : 29045  
 1st Qu.: 98.14   1st Qu.:2.500   1st Qu.:  222.5   1st Qu.:25.68   1st Qu.: 41095  
 Median : 99.63   Median :3.050   Median :  698.0   Median :27.65   Median : 57837  
 Mean   : 99.22   Mean   :3.076   Mean   : 2270.3   Mean   :27.84   Mean   :100990  
 3rd Qu.:100.00   3rd Qu.:3.725   3rd Qu.: 2816.8   3rd Qu.:29.73   3rd Qu.: 98833  
 Max.   :101.29   Max.   :5.100   Max.   :16583.0   Max.   :33.20   Max.   :466970  
```

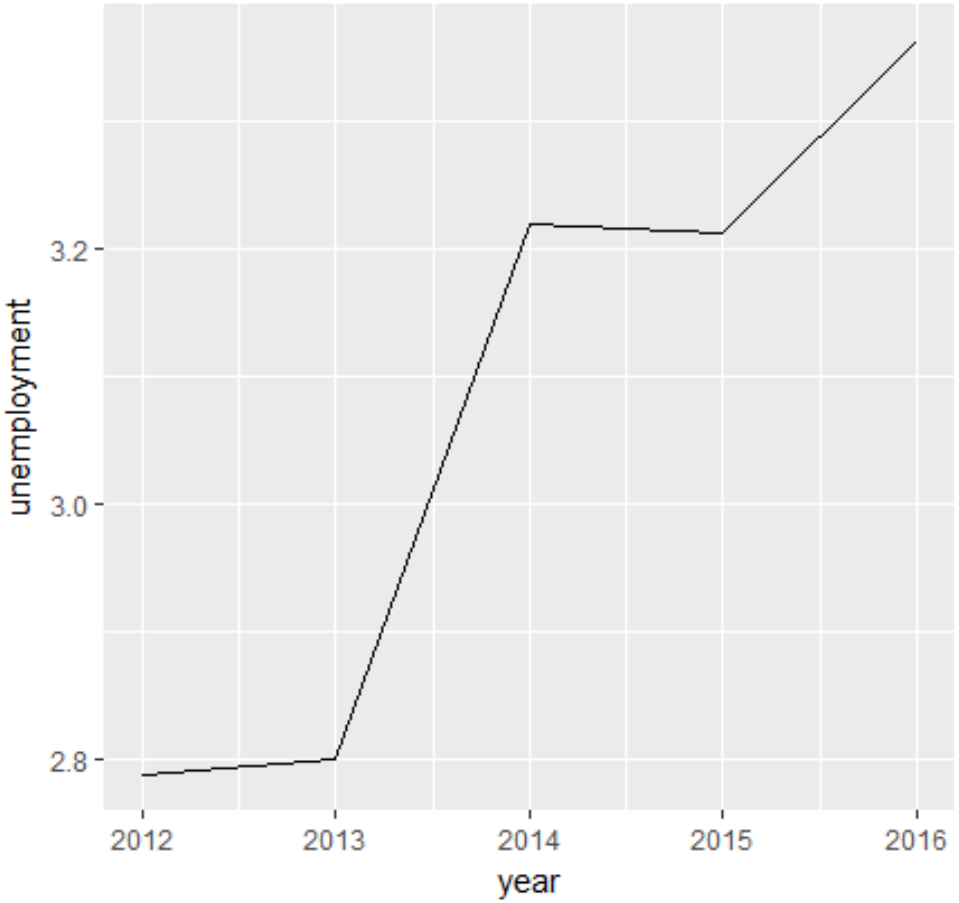
물가지수	실직률	인구	스트레스 인지율	범죄 발생 수
2015년도 기준 100	%	인구밀도	%	건 수

연도별 그래프

■ 물가지수

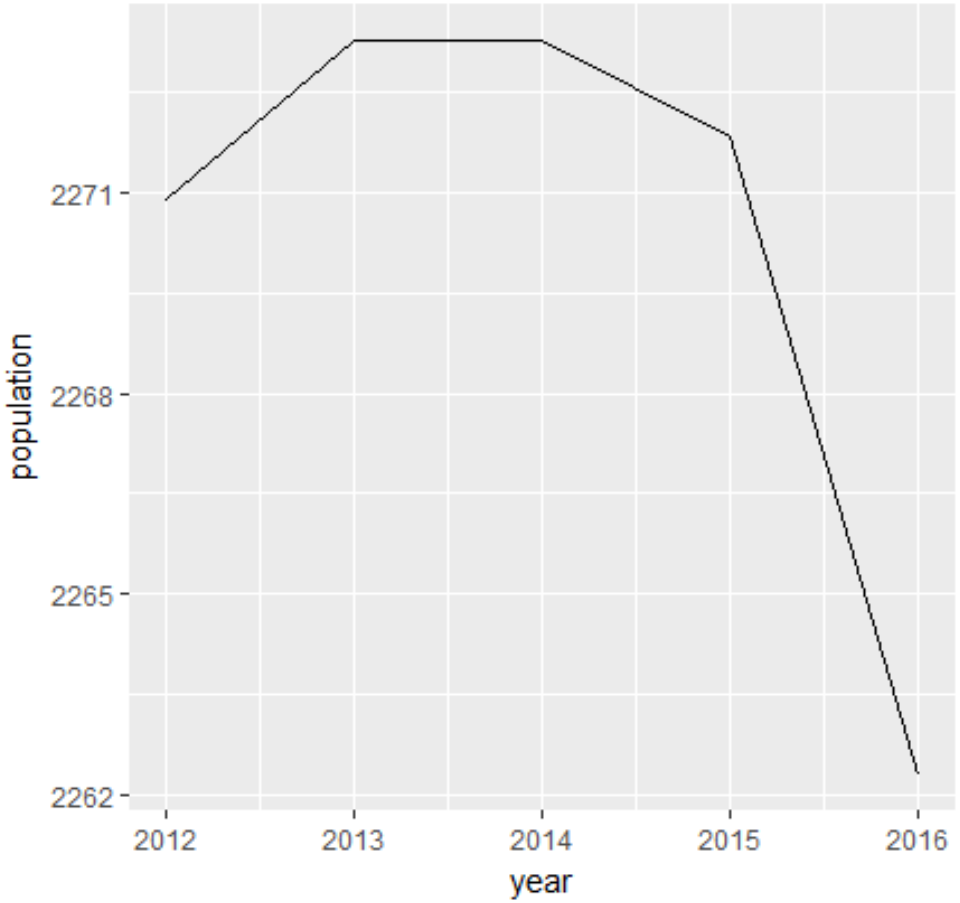


■ 실직률

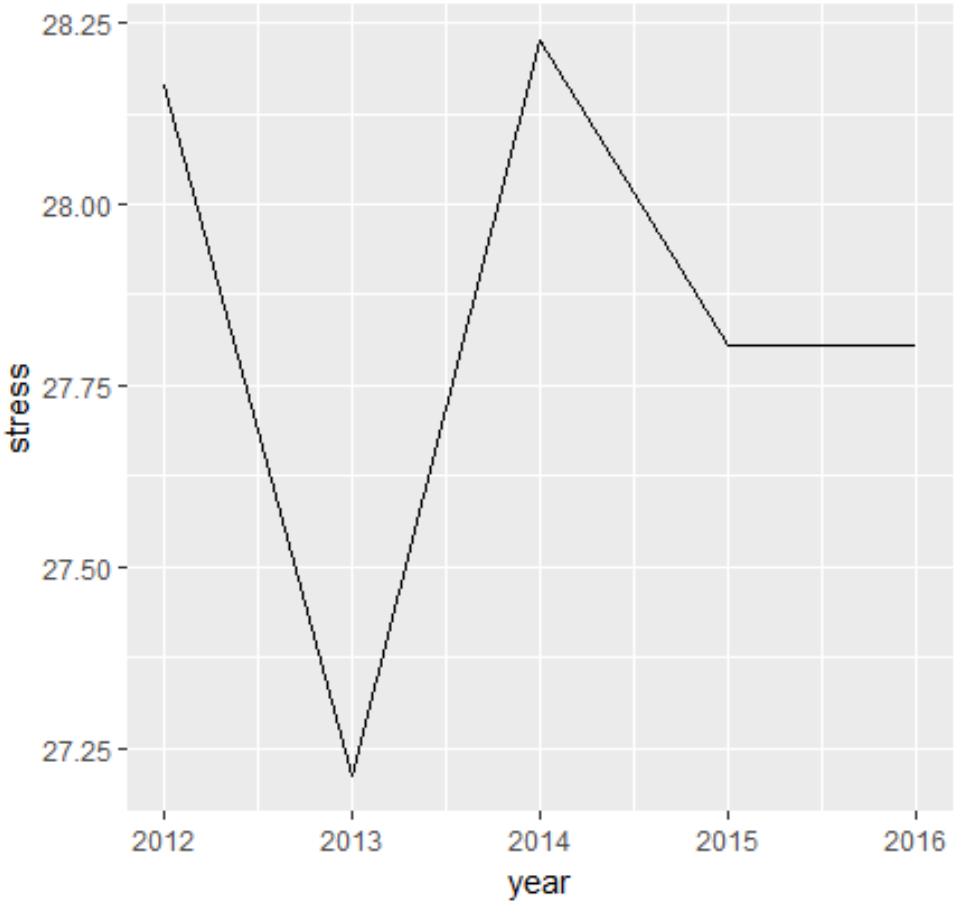


연도별 그래프

■ 인구

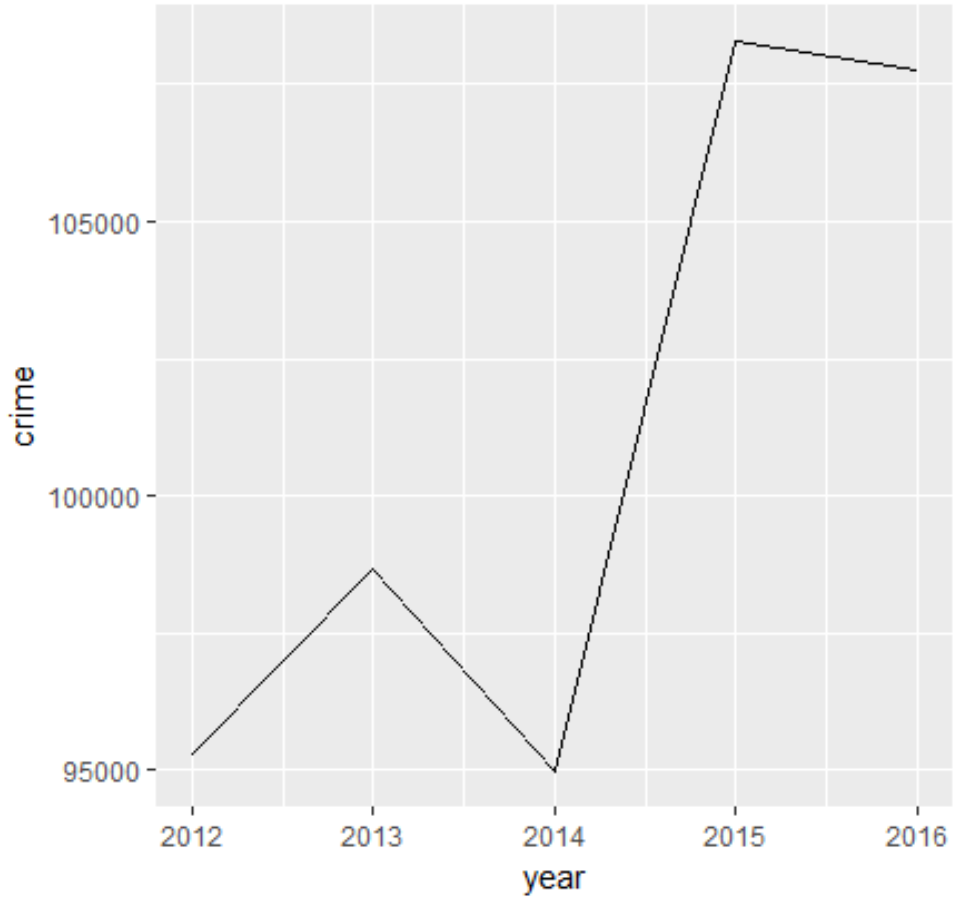


■ 스트레스 인지율



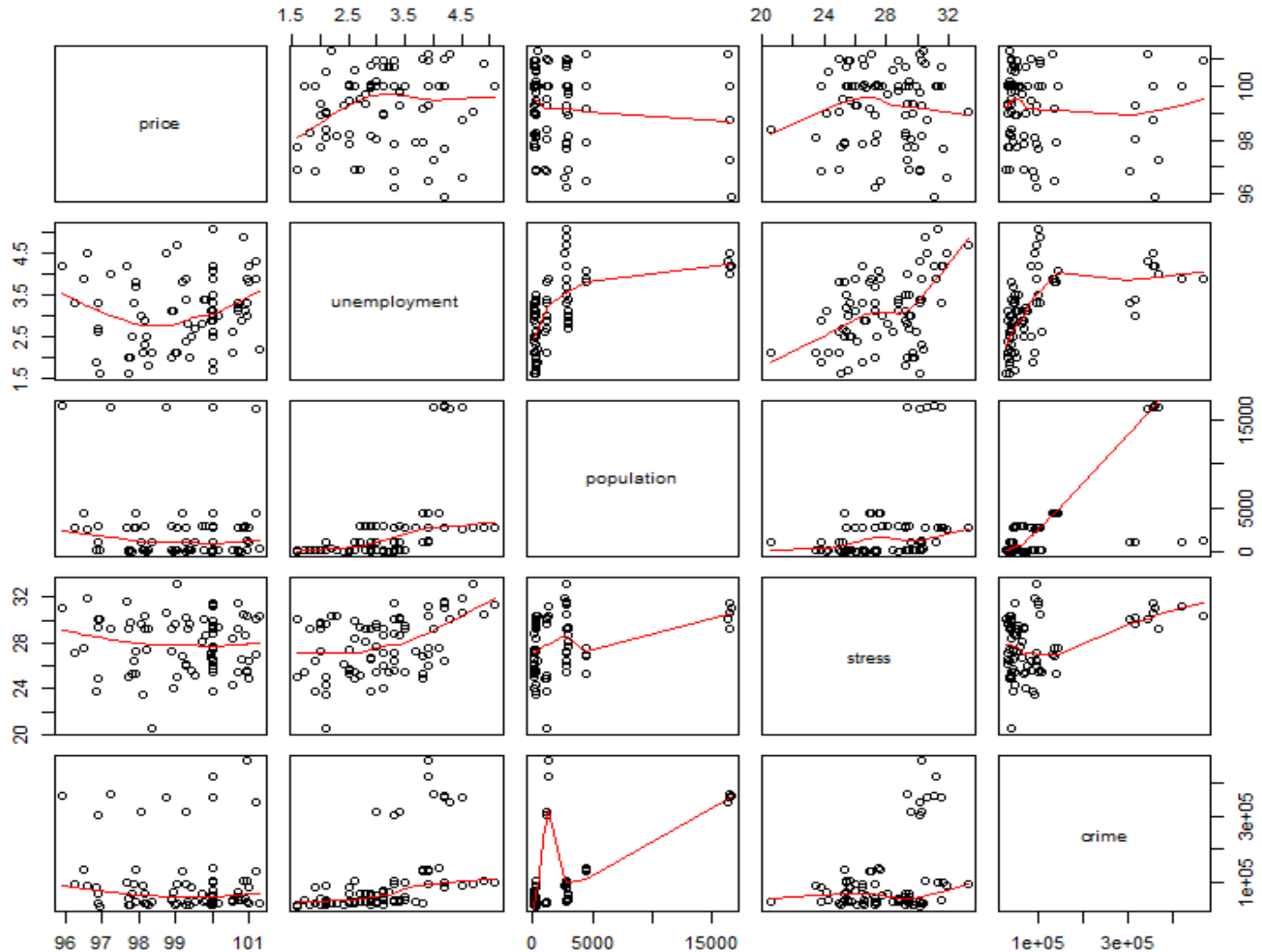
연도별 그래프

- 범죄 발생 수



상관분석

각 변수들과 범죄 발생건수 간의 산점도 행렬



각 변수들과 범죄 발생건수 간 상관관계

- ✓ 물가 지수와의 연관성이 떨어짐
- ✓ 인구 수와 실직률은 범죄 발생수와 관련성이 있음
- ✓ 스트레스 인지율은 위의 요인들 보다는 비교적 낮은 연관성

각 변수들과 범죄 발생건수 간 상관관계 검정

```
> cor.test(crime, price)

        Pearson's product-moment correlation

data:  crime and price
t = -0.89331, df = 78, p-value = 0.3744
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3134209  0.1217755
sample estimates:
      cor
-0.100634
```

```
> cor.test(crime, population)

        Pearson's product-moment correlation

data:  crime and population
t = 7.7276, df = 78, p-value = 3.165e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5130027 0.7672114
sample estimates:
      cor
0.6584969
```

```
> cor.test(crime, unemployment)

        Pearson's product-moment correlation

data:  crime and unemployment
t = 5.2944, df = 78, p-value = 1.07e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3319458 0.6593884
sample estimates:
      cor
0.5141627
```

```
> cor.test(crime, stress)

        Pearson's product-moment correlation

data:  crime and stress
t = 3.4876, df = 78, p-value = 0.000804
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1605290 0.5431739
sample estimates:
      cor
0.3672912
```

✓ 네 변수 모두 p값이 작으므로 유의한 상관관계

회귀분석

모형 선택

```
> anova(m_before)
Analysis of Variance Table

Response: crime
             Df     Sum Sq    Mean Sq F value    Pr(>F)
price         1 8.6727e+09 8.6727e+09  1.4607    0.2306
unemployment  1 2.4196e+11 2.4196e+11 40.7528 1.295e-08 ***
population    1 1.5079e+11 1.5079e+11 25.3977 3.135e-06 ***
stress        1 9.6553e+09 9.6553e+09  1.6262    0.2062
Residuals    75 4.4530e+11 5.9373e+09
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

✓ 분산분석표에서 price와 stress가 유의하지 않으므로 두 변수를 제거한 모형을 만듦

모형 선택

```
> anova(m_after, m_before)
Analysis of Variance Table

Model 1: crime ~ unemployment + population
Model 2: crime ~ price + unemployment + population + stress
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1     77 4.5707e+11
2     75 4.4530e+11  2 1.1776e+10 0.9917 0.3758
```

✓ 변수의 중요도를 평가하기 위해 F-test 실시
✓ P값이 0.3758이므로 두 변수를 제거하지 않음

```
> AIC(m_after, m_before)
          df      AIC
m_after    4 2032.317
m_before   6 2034.228
```

✓ AIC값 또한 원래의 모형이 더 좋음을 나타냄
✓ 따라서 변수를 제거하기 전 모형 선택

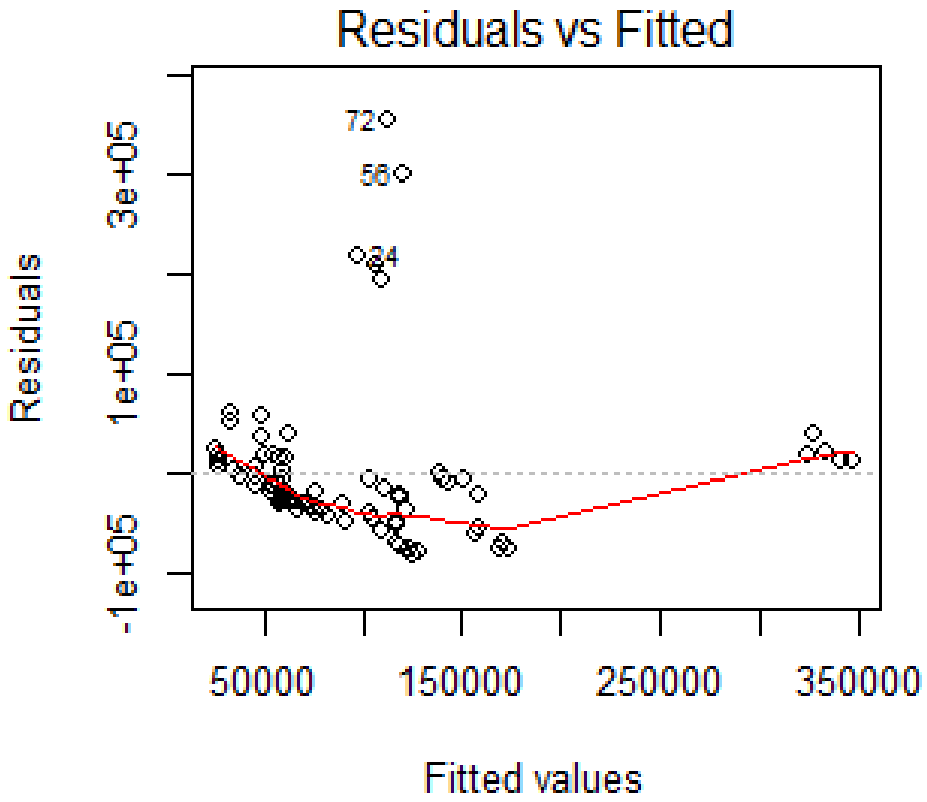
모형 선택

```
> summary(m_before)

Call:
lm(formula = crime ~ price + unemployment + population + stress,
    data = data)

Residuals:
   Min     1Q Median     3Q    Max
-79958 -35171 -16466  13577 353440

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  172660.572 667844.672   0.259   0.7967
price         -3179.338   6647.868  -0.478   0.6339
unemployment  23760.088  13688.252   1.736   0.0867 .
population       13.491      2.761   4.887 5.67e-06 ***
stress         5030.559   3944.814   1.275   0.2062
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77050 on 75 degrees of freedom
Multiple R-squared:   0.48,     Adjusted R-squared:  0.4523
F-statistic: 17.31 on 4 and 75 DF,  p-value: 4.248e-10
```

✓ 결정계수 값이 0.4523으로 45.23%의 설명력

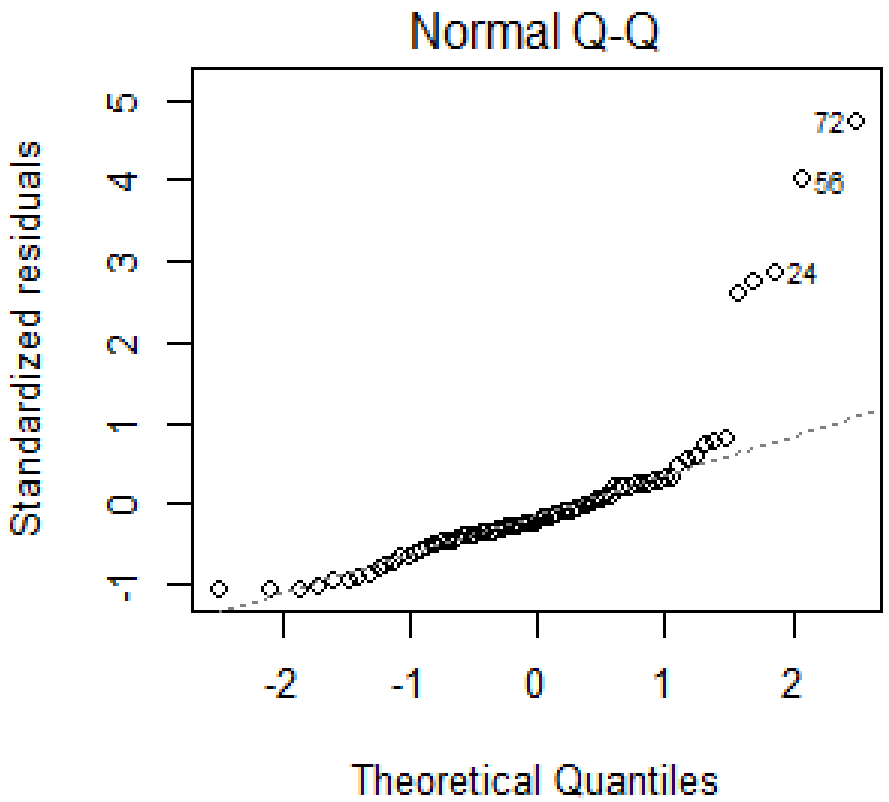
회귀 진단

■ 선형성



✓ 종속변수와 독립변수가 선형관계에 있다면
잔차와 예측치 사이에 관계가 있으면 안됨
✓ 따라서 종속변수와 독립변수가 선형관계

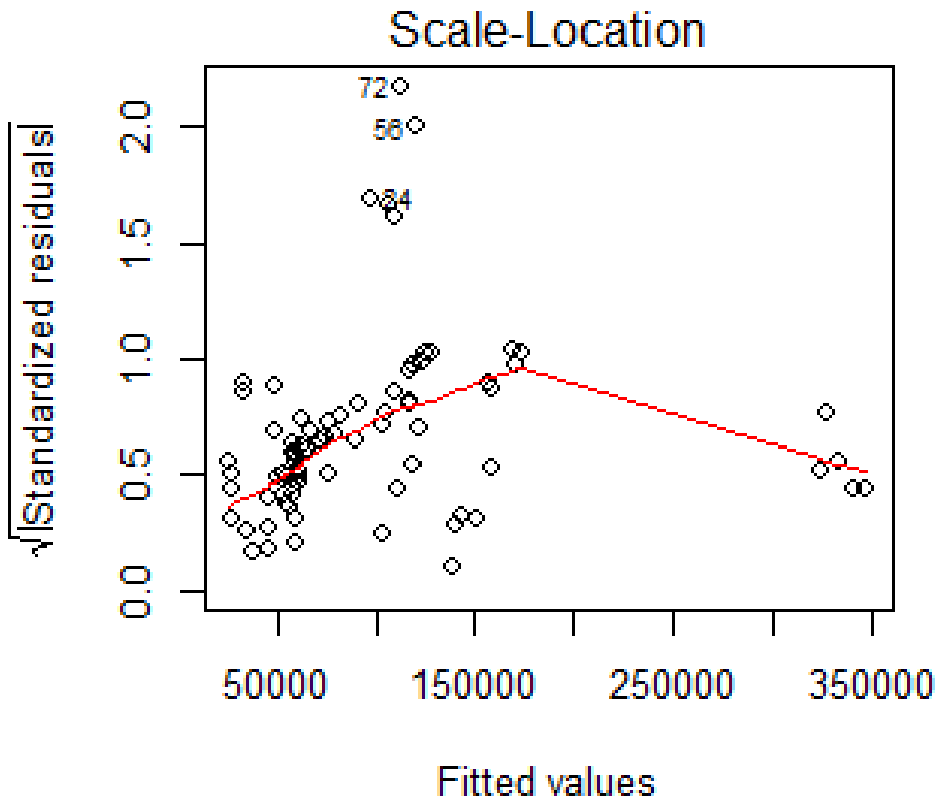
■ 정규성



✓ 정규성 가정을 만족한다면
Q-Q plot의 점들이 45도 직선위에 있어야 함
✓ 따라서 정규성 가정을 위반한 것

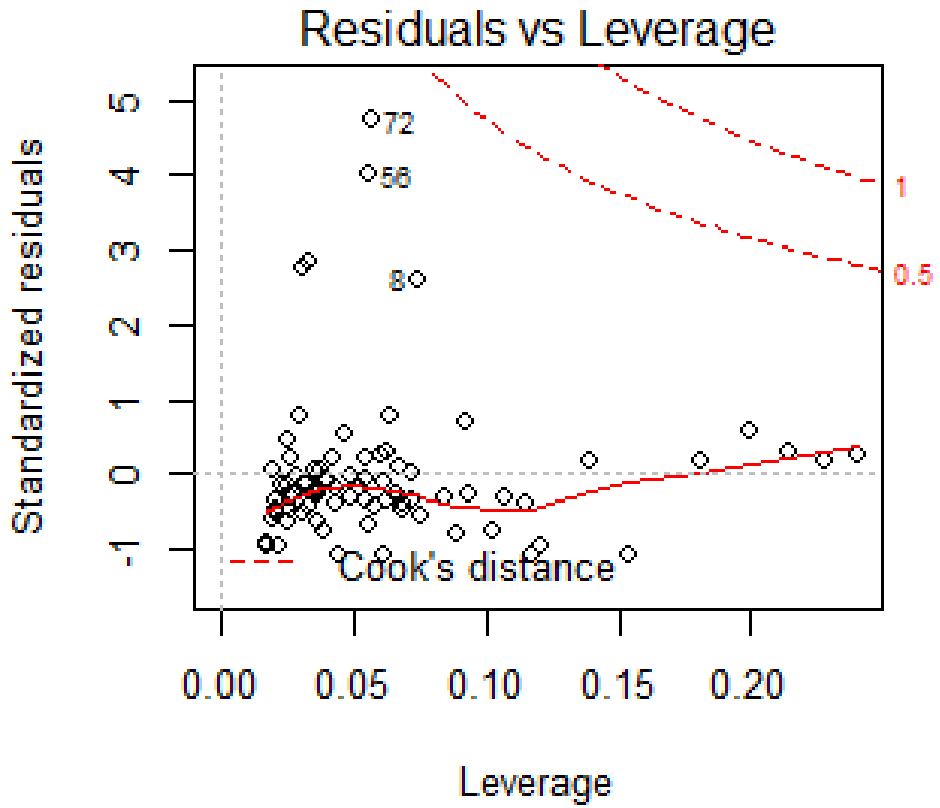
회귀 진단

■ 등분산성



✓ 분산이 일정하다는 가정 만족한다면
그래프의 수평선 주위에 랜덤하게 나타나야 함
✓ 따라서 등분산성을 만족하지 못하는 것으로 보임

■ Residuals vs Leverage plot



✓ 개개의 관찰치에 대한 정보 제공
✓ 72, 56, 8번 관측치가
Cook’s distance가 큰 영향관측치

회귀 진단 – 독립성

```
> durbinWatsonTest(m_before)
 lag Autocorrelation D-W Statistic p-value
   1    -0.003314655       2.005027    0.79
 Alternative hypothesis: rho != 0
```

✓ 독립성 가정 검정 – Durbin-Watson 검정
✓ P값이 0.79로 자기상관은 없다고 할 수 있다

회귀진단 – 다중공선성

```
> vif(m_before)
       price unemployment   population       stress
    1.104637     1.708702     1.564908     1.244174
> sqrt(vif(m_before))>2
       price unemployment   population       stress
       FALSE        FALSE        FALSE        FALSE
```

✓ 다중공선성은 VIF 통계량 사용하여 계산할 수 있음
✓ VIF의 제곱근은 다중공선성의 정도를 나타내며 2 이상인 것은 다중공선성 문제가 있다는 것을 뜻함
✓ 따라서 다중공선성 문제는 없다는 것을 확인

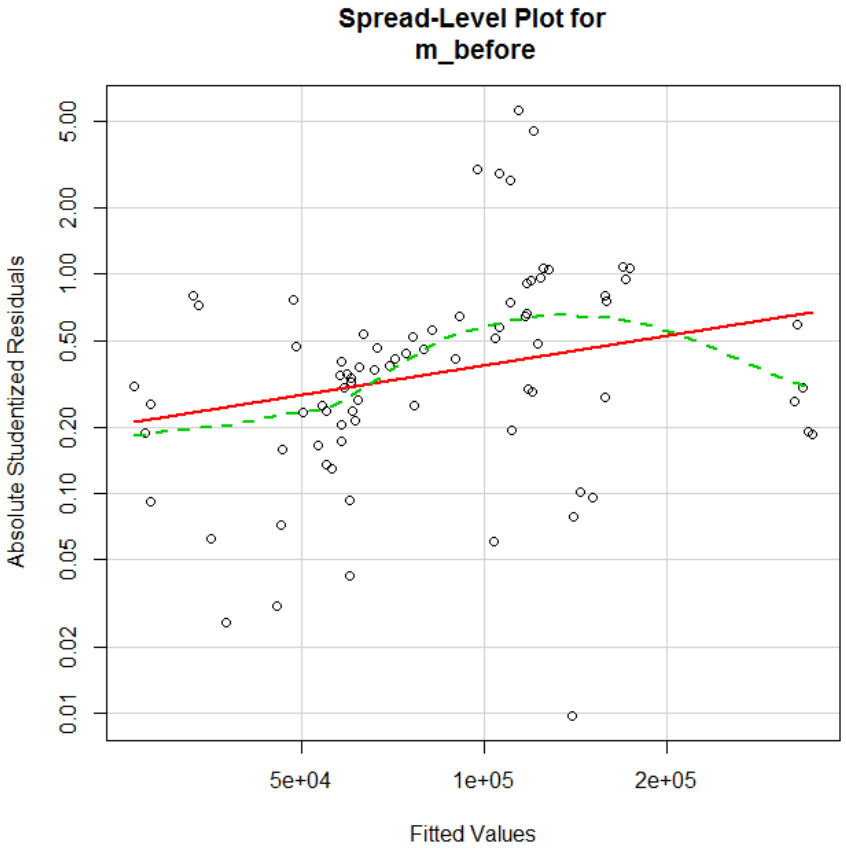
회귀진단 – 이상치

```
> outlierTest(m_before)
   rstudent unadjusted p-value Bonferonni p
72 5.597542         3.4939e-07   2.7951e-05
56 4.478348         2.6837e-05   2.1470e-03
```

✓ 72번과 56번 관측치가 이상치 임을 알 수 있다

회귀모형 교정

```
> ncvTest(m_before)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.084231    Df = 1     p = 0.2977529
> spreadLevelPlot(m_before)

Suggested power transformation:  0.5527039
```



✓ 등분산성 개선
✓ 결과가 유의하지 않으므로 등분산성 가정을 만족했다고 볼 수 있음

분석 결과

변수들의 중요성

```
> data2 = as.data.frame(data[,c("crime", "price", "unemployment",
+         "population", "stress")])
> zdata = as.data.frame(scale(data2))
> zfit = lm(crime~price+unemployment+population+stress, data=zdata)
> coef(zfit)
  (Intercept)         price  unemployment    population        stress
 2.490798e-16 -4.184284e-02  1.889271e-01  5.089953e-01  1.184380e-01
```

- ✓ 변수들의 표준화된 회귀계수 비교
- ✓ 평균이 1, 표준편차가 1로 표준화한 후 회귀분석
- ✓ 인구가 표준편차만큼 증가하면 물가지수, 실직률, 스트레스 인지율이 일정할 때 범죄 발생수가 표준편차의 0.5배 증가
- ✓ 인구가 가장 중요한 변수

추정과 예측

- ✓ 물가지수 96, 실직률 3, 인구 5000, 스트레스 인지율 27일 때와
  물가지수 98, 실직률 5, 인구 18000, 스트레스 인지율 30일 때
  범죄 발생 건수 예측
- ✓ 범죄 발생의 95% 신뢰 구간

```
> d = data.frame("price"=c(96,98), "unemployment"=c(3,5),
+                "population"=c(5000,18000), "stress"=c(27,30))
> predict(m_before, d, interval="confidence")
       fit       lwr      upr
1 142004.2  96112.76 187895.6
2 373639.4 301457.61 445821.3
```

Reference

- ✓ 국가 통계 포털 KOSIS


THANK
YOU