Analysis of factors affecting fertility rate

출산율에 영향을 미치는 요인 분석

Data analysis project with R

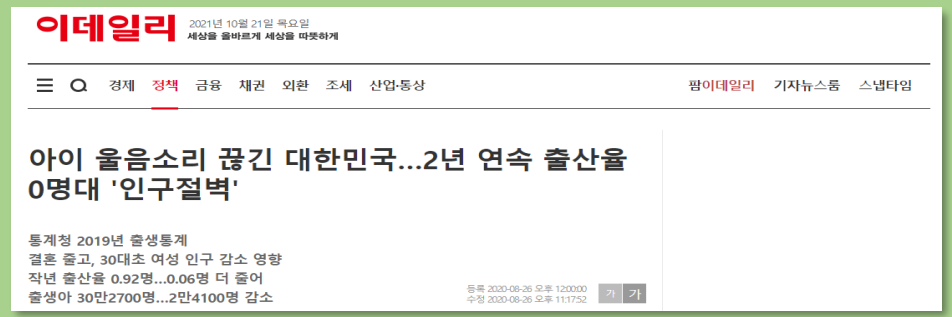
안준호

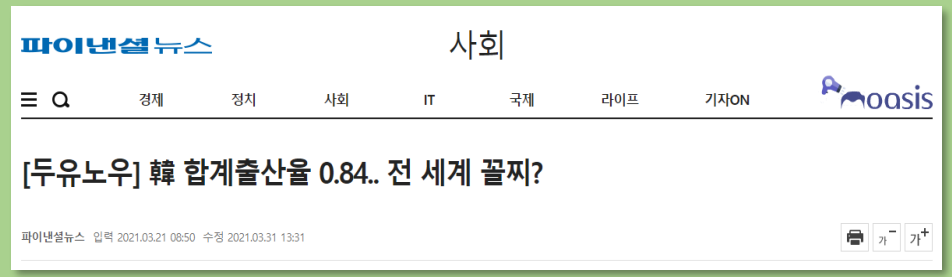
Contents

1.연구 배경


한국강사신문
KOREA LECTURER NEWS
강사뉴스 책과 사람 북세미나 인터뷰 칼럼
KBS 다큐온, 2020년 대한민국 합계출산율 0.84명 '아이, 낳을 수 있을까?'


BabyTimes
국방신문
전체 출산육아 산업 금융·재테크 생활문화 정치·사회 오피니언 포토영상 행사이벤트
[인터뷰] 대한민국 저출산 이대로 괜찮은가?


이데일리
아이 울음소리 끊긴 대한민국...2년 연속 출산율 0명대 '인구절벽'
통계청 2019년 출생통계
결혼 줄고, 30대초 여성 인구 감소 영향
작년 출산율 0.92명...0.06명 더 줄어
출생아 30만2700명...2만4100명 감소


파이낸셜뉴스
사회
경제 정치 사회 IT 국제 라이프 기자ON
[두유노우] 韓 합계출산율 0.84.. 전 세계 꼴찌?

대한민국의 출산율은 세계 최하위 수준으로 평균 1명이 안되는 유일무이한 국가이며, 지속적으로 감소하는 추세를 보이고 있음 (현재 세계 평균 출산율은 약 2.4명)

출처 : 네이버 뉴스

그렇다면, 출산율이 낮은 나라들이 가지고 있는 공통적인 특징이 있을지?

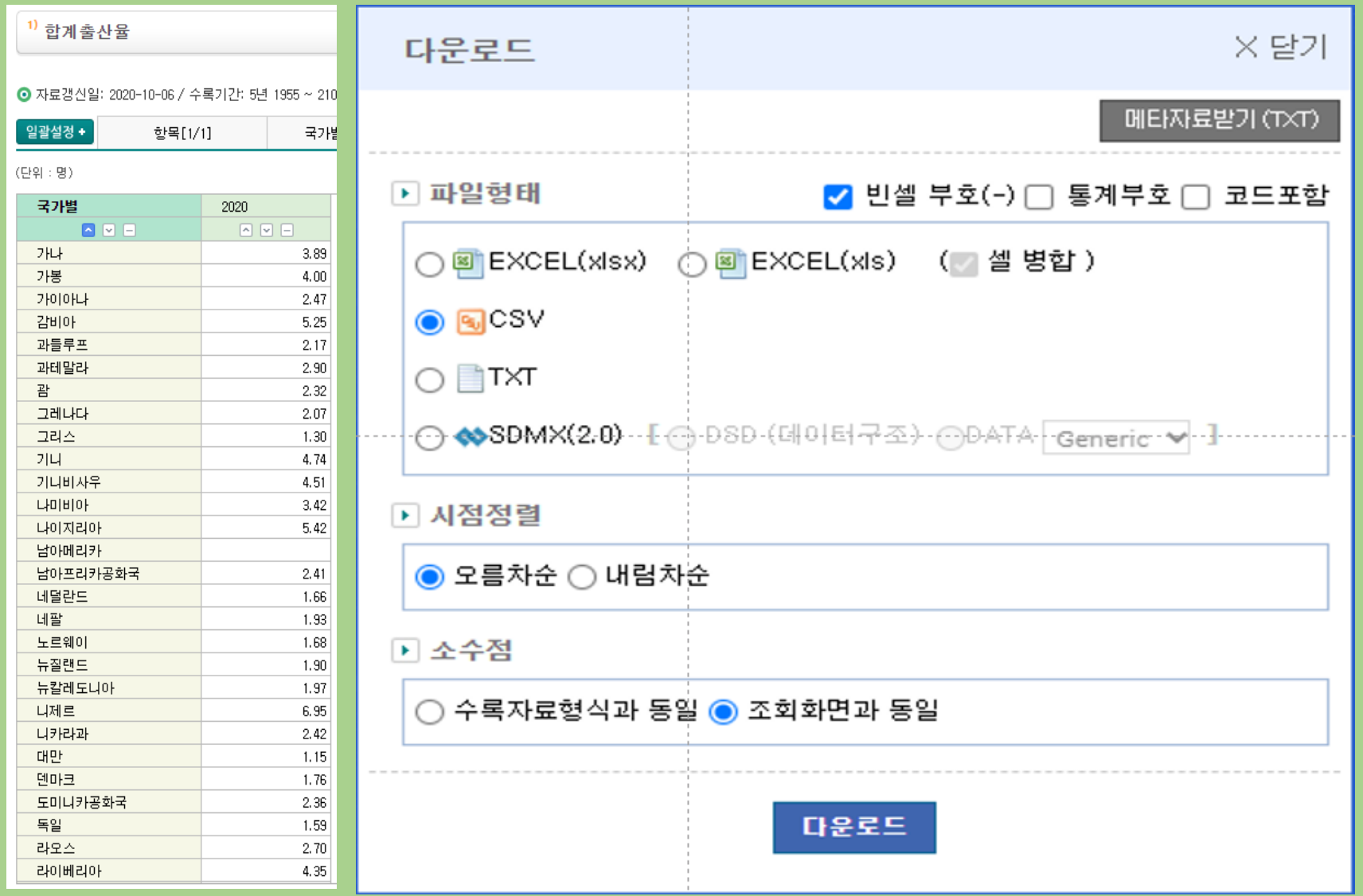
- 국민 소득 수준이 높은 국가일수록 출산율이 낮을까?
- 도시화율이 높은 국가일수록 출산율이 낮을까?
- 경제활동을 하는 여성의 비율이 높은 국가일수록 출산율이 낮을까?
- 성비가 높은 국가일수록 출산율이 낮을까?
- 지니 계수가 높은 국가일수록 출산율이 낮을까?
- 국민 소득 대비 주택 가격이 클수록 출산율이 낮을까?

→ 국민 소득 수준, 도시화율, 여성 경제활동 인구 비율, 성비, 지니 계수, 국민 소득 대비 주택 가격 등 총 6개를 독립변수로 고려

→ 6개의 후보 독립 변수 중 국민 소득 수준, 도시화율, 여성 경제활동 인구 비율 변수에 대해서 dataset 수집이 가능하여 총 3개의 독립변수로 분석 진행

*도시화율: 전체 인구 중에서 도시에 거주하는 인구가 차지하는 비율
*지니계수: 빈부격차와 소득불균형의 척도, 0에 가까울수록 평등하고 1에 가까울수록 불평등함을 의미

2. 데이터 수집



1) 합계출산율

자료갱신일: 2020-10-06 / 수록기간: 5년 1955 ~ 210

일괄설정 + | 항목[1/1] | 국가별

(단위 : 명)

국가별	2020
가나	3.89
가봉	4.00
가이아나	2.47
감비아	5.25
과들루프	2.17
과테말라	2.90
괌	2.32
그레나다	2.07
그리스	1.30
기니	4.74
기니비사우	4.51
나미비아	3.42
나이지리아	5.42
남아메리카	
남아프리카공화국	2.41
네덜란드	1.66
네팔	1.93
노르웨이	1.68
뉴질랜드	1.90
뉴칼레도니아	1.97
니제르	6.95
니카라과	2.42
대만	1.15
덴마크	1.76
도미니카공화국	2.36
독일	1.59
라오스	2.70
라이베리아	4.35

각 변수에 대한 dataset(2015년, 2020년 자료)을 각각 kosis(국가통계포털)에서 csv 형태로 다운로드

	A	B
1	국가별	출산율
2	가나	3.89
3	가봉	4
4	가이아나	2.47
5	감비아	5.25
6	과들루프	2.17
7	과테말라	2.9
8	괌	2.32
9	그레나다	2.07
10	그리스	1.3
11	기니	4.74
12	기니비사우	4.51
13	나미비아	3.42
14	나이지리아	5.42
15	남아메리카	
16	남아프리카	2.41
17	네덜란드	1.66
18	네팔	1.93
19	노르웨이	1.68
20	뉴질랜드	1.9
21	뉴칼레도니	1.97
22	니제르	6.95
23	니카라과	2.42
24	대만	1.15
25	덴마크	1.76
26	도미니카공	2.36
27	독일	1.59
28	라오스	2.7
29	라이베리아	4.35
30	라트비아	1.72
31	러시아	1.82
32	레바논	2.09
33	레소토	3.16
34	레위니옹	2.27

	A	B
1	국가별	도시화율
2	가나	57.3
3	가봉	90.1
4	가이아나	26.8
5	감비아	62.6
6	과들루프	98.5
7	과테말라	51.8
8	괌	94.9
9	교황청	100
10	그레나다	36.5
11	그리스	79.7
12	기니	36.9
13	기니비사우	44.2
14	나미비아	52
15	나이지리아	52
16	남아메리카	
17	남아프리카	67.4
18	네덜란드	92.2
19	네팔	20.6
20	노르웨이	83
21	뉴질랜드	86.7
22	니제르	16.6
23	니카라과	59
24	덴마크	88.1
25	도미니카공	82.5
26	도미니카연	71.1
27	독일	77.5
28	동티모르	31.3
29	라오스	36.3
30	라이베리아	52.1
31	라트비아	68.3
32	러시아	74.8
33	레바논	88.9
34	레소토	29

	A	B
1	국가	1인당 소득
2	가나	2230
3	가봉	6970
4	가이아나	6600
5	감비아	750
6	과테말라	4490
7	그레나다	8740
8	기니	1020
9	기니비사우	760
10	나미비아	4520
11	나이지리아	2000
12	남아메리카	
13	남아프리카	5410
14	네팔	1190
15	노르웨이	78250
16	니제르	540
17	니카라과	1850
18	덴마크	62720
19	도미니카공	7260
20	도미니카연	6870
21	독일	46980
22	동티모르	1830
23	라오스	2480
24	라이베리아	530
25	러시아	10690
26	레바논	5510
27	레소토	1100
28	루마니아	12570
29	르완다	780
30	리비아	4850
31	마다가스카	480
32	말라위	580
33	말레이시아	10580

	A	B
1	국가	여성 경제활동 비율
2	가나	46.44
3	가봉	40.319
4	과테말라	33.046
5	그리스	43.824
6	나이지리아	44.822
7	남아프리카	45.471
8	네덜란드	46.269
9	노르웨이	46.995
10	뉴질랜드	47.608
11	니카라과	38.68
12	덴마크	47.374
13	독일	46.324
14	라이베리아	47.414
15	라트비아	49.939
16	러시아	48.57
17	레바논	24.451
18	루마니아	42.907
19	룩셈부르크	45.999
20	리비아	33.995
21	리투아니아	50.082
22	마다가스카	48.894
23	말레이시아	38.45
24	멕시코	37.827
25	모로코	24.113
26	몬테네그로	43.586
27	몰도바	49.446
28	몰타	40.864
29	미국	46.008
30	바레인	19.775
31	바베이도스	49.48
32	바하마	47.403
33	베냉	49.223

	A	B	C	D	E	F
1	Year	국가	출산율	도시화율	여성 경제활동 비율	1인당 소득
2	2020	가나	3.89	57.3	46.44	2230
3	2020	가봉	4	90.1	40.319	6970
4	2020	가이아나	2.47	26.8		6600
5	2020	감비아	5.25	62.6		750
6	2020	과들루프	2.17	98.5		
7	2020	과테말라	2.9	51.8	33.046	4490
8	2020	괌	2.32	94.9		
9	2020	그레나다	2.07	36.5		8740
10	2020	그리스	1.3	79.7	43.824	
11	2020	기니	4.74	36.9		1020
12	2020	기니비사우	4.51	44.2		760
13	2020	나미비아	3.42	52		4520
14	2020	나이지리아	5.42	52	44.822	2000
15	2020	남아프리카공화국	2.41	67.4	45.471	5410
16	2020	네덜란드	1.66	92.2	46.269	
17	2020	네팔	1.93	20.6		1190
18	2020	노르웨이	1.68	83	46.995	78250
19	2020	뉴질랜드	1.9	86.7	47.608	
20	2020	뉴칼레도니아	1.97			
21	2020	니제르	6.95	16.6		540
22	2020	니카라과	2.42	59	38.68	1850
23	2020	대만	1.15			
24	2020	덴마크	1.76	88.1	47.374	62720
25	2020	도미니카공화국	2.36	82.5		7260
26	2020	독일	1.59	77.5	46.324	46980
27	2020	라오스	2.7	36.3		2480
28	2020	라이베리아	4.35	52.1	47.414	530
29	2020	라트비아	1.72	68.3	49.939	
30	2020	러시아	1.82	74.8	48.57	10690
31	2020	레바논	2.09	88.9	24.451	5510
32	2020	레소토	3.16	29		1100
33	2020	레위니옹	2.27	99.7		

다운로드한 각각의 변수에 대한 dataset을 Excel의 VLOOKUP 함수를 이용하여 하나의 dataset으로 통합

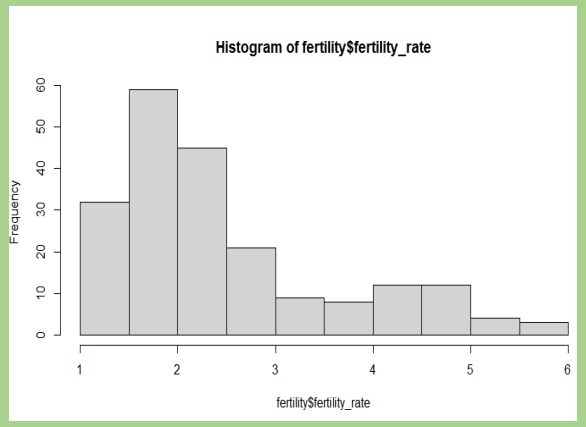
3. 데이터 분석

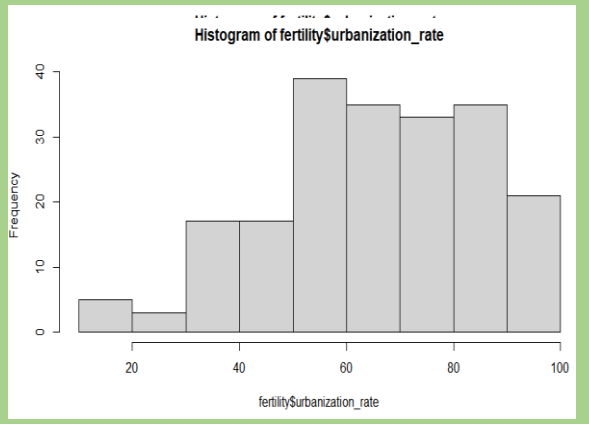
```r
# 데이터셋 불러오기
fertility <- read.csv('fertility_rate.csv')
head(fertility,10)

# 종속변수/독립변수 추출 및 결측치 제거
fertility <- fertility[c(3:6)] # 종속변수: 출산율 / 독립변수: 도시화율, 여성 경제활동인구 비율, 1인당 국민소득 열만 추출
names(fertility) <- c('fertility_rate', 'urbanization_rate', 'female_worker_ratio', 'national_income') # 열 이름 영문으로 수정
fertility <- na.omit(fertility) # 결측치가 존재하는 행 제거
str(fertility) # 'data.frame':  205 obs. of  4 variables:
```

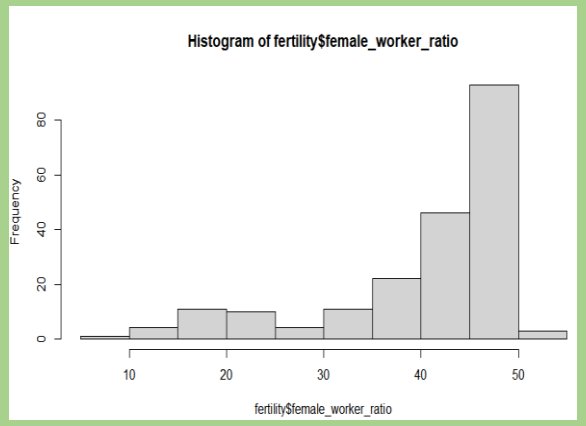
➢ 데이터셋 로드 및 결측치가 존재하는 행 제거 -> 제거 후 총 205개의 행

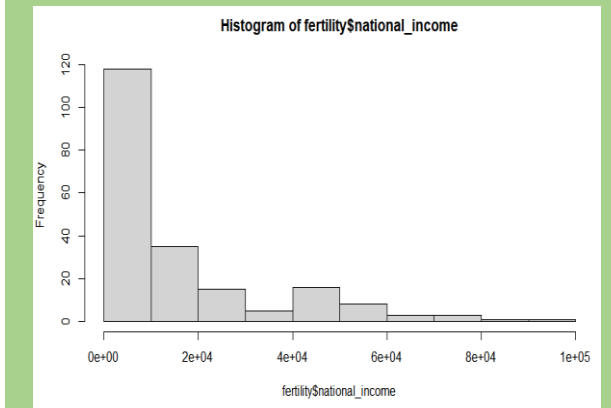
```r
# 각 변수별 분포, 통계량 확인
library(moments)

# 1) fertility_rate(출산율)
summary(fertility$fertility_rate)
#  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 1.110   1.680   2.090   2.491   2.900   6.000
hist(fertility$fertility_rate) # 왼쪽으로 치우친 형태(주로 1.5~2.5)
skewness(fertility$fertility_rate) # 왜도 = 1.183909 > 0
kurtosis(fertility$fertility_rate) # 첨도 = 3.428276 -> 정규분포(3)보다 약간 뾰족한 형태

# 2) urbanization_rate(도시화율)
summary(fertility$urbanization_rate)
#  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 13.00   53.80   67.20   65.51   81.30  100.00
hist(fertility$urbanization_rate) # 오른쪽으로 약간 치우친 형태
skewness(fertility$urbanization_rate) # 왜도 = -0.3967547 < 0
kurtosis(fertility$urbanization_rate) # 첨도 = 2.572506 -> 정규분포(3)보다 뭉툭한 형태

# 3) female_worker_ratio(여성 경제활동인구 비율)
summary(fertility$female_worker_ratio)
#  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
# 8.111  38.536  44.119  40.545  47.125  50.465
hist(fertility$female_worker_ratio) # 오른쪽으로 많이 치우친 형태
skewness(fertility$female_worker_ratio) # 왜도 = -1.497028 < 0
kurtosis(fertility$female_worker_ratio) # 첨도 = 4.284562 -> 정규분포(3)보다 뾰족한 형태

# 4) national_income(1인당 국민소득)
summary(fertility$national_income)
# Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#  480    3090    6970   15752   20140   92910
hist(fertility$national_income) # 왼쪽으로 많이 치우친 형태
skewness(fertility$national_income) # 왜도 = 1.723397 > 0
kurtosis(fertility$national_income) # 첨도 = 5.413992 -> 정규분포(3)보다 매우 뾰족한 형태
```

➢ 각 변수에 대한 통계량 및 대략적인 분포 확인

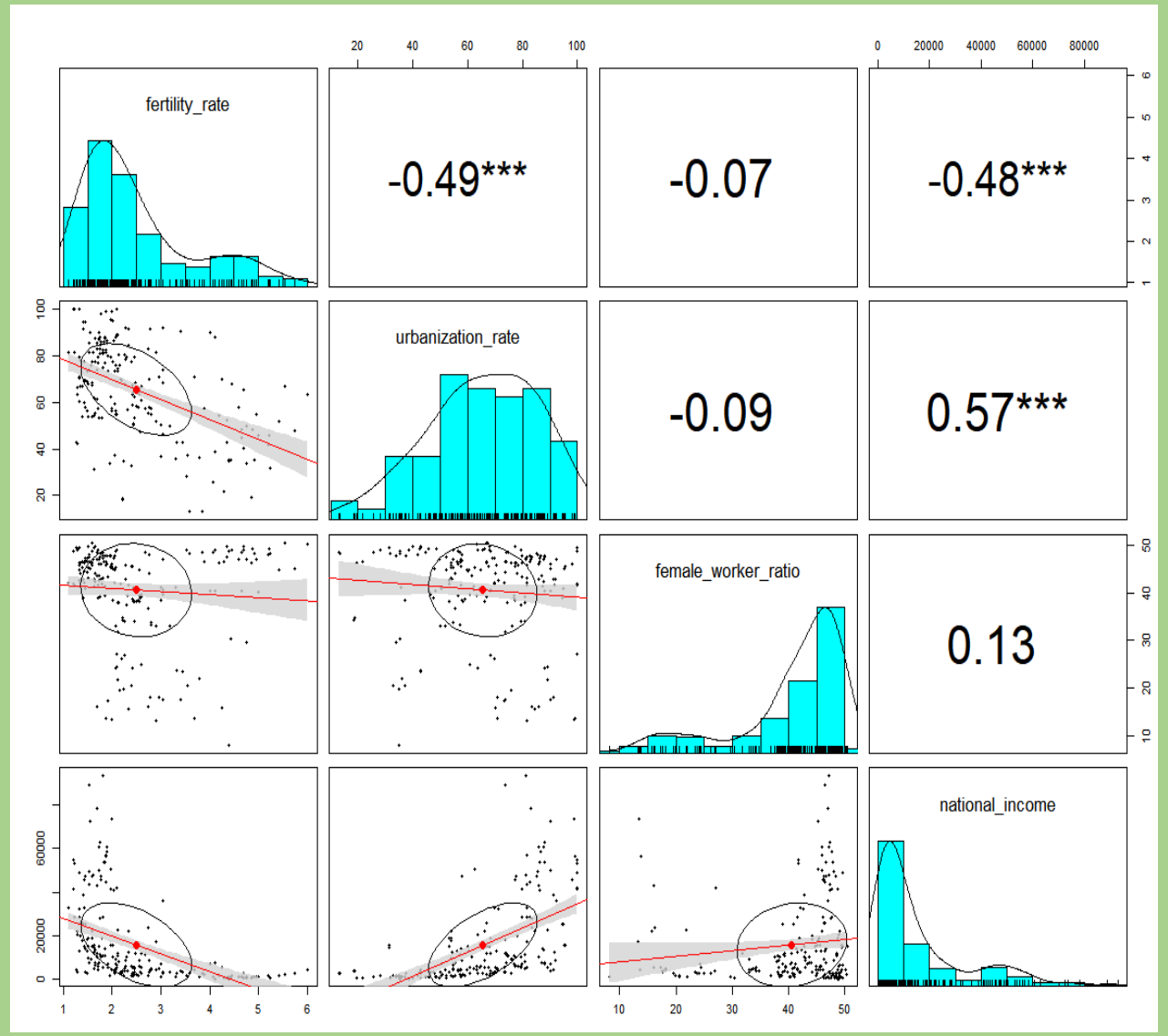








```r
# 상관계수 확인
COR <- cor(fertility)
COR["fertility_rate",]

# fertility_rate   urbanization_rate female_worker_ratio     national_income
#     1.00000000          -0.49020925          -0.07164783          -0.47659585
# -> 여성 경제활동 인구 비율은 상관관계가 거의 없는 것으로 보임


# 상관계수 및 회귀선 시각화
library(psych)
pairs.panels(fertility, stars = TRUE, lm = TRUE, ci = TRUE)
```



- 종속변수와 독립변수 간의 상관관계 분석 결과 여성 경제활동 인구 비율은 약 -0.07로 종속변수와 거의 상관관계가 없는 것처럼 보임

```r
# 독립변수간 다중공선성 확인
library(car)
vif(f_lm) > 10 # 분산팽창계수가 10을 초과하는지 확인

# urbanization_rate female_worker_ratio     national_income
#             FALSE                FALSE               FALSE
# -> 독립변수간 다중공선성 문제 없음
```

- 분산팽창계수가 10을 초과하는 독립변수가 없으므로 다중공선성 문제도 없는 것으로 확인됨

```
# 회귀모델 생성 및 확인
f_lm = lm(formula = fertility_rate ~., data = fertility)
summary(f_lm)
```

```
Call:
lm(formula = fertility_rate ~ ., data = fertility)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7944 -0.6873 -0.1400  0.4215  3.3606

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         4.349e+00  4.100e-01  10.608  < 2e-16 ***
urbanization_rate  -1.954e-02  4.195e-03  -4.659 5.78e-06 ***
female_worker_ratio -7.941e-03  7.069e-03  -1.123 0.262604
national_income    -1.623e-05  4.346e-06  -3.734 0.000246 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9551 on 201 degrees of freedom
Multiple R-squared:  0.3026,    Adjusted R-squared:  0.2922
F-statistic: 29.07 on 3 and 201 DF,  p-value: 1.182e-15
```

- 회귀모델 생성 후 확인 결과 여성 경제활동 인구 비율은 유의하지 않은 변수로 나타남
- 도시화율, 국민 소득 변수는 매우 유의한 변수로 나타나며 모두 종속변수에 부(-)의 영향을 끼침
- p-value: 1.182e-15 < 0.05이므로 통계적으로 유의함
- Adjusted R-squared = 0.2922로 모델의 설명력은 다소 낮은 편에 속함
- RMSE = 0.9551
- MSE = RMSE^2 = 0.9551^2 = 0.912216

4. 결론

Conclusion

1. 여성 경제활동 인구 비율은 출산율에 큰 영향을 미치지 않음
2. 도시화율이 높은 국가일수록 출산율이 낮음
3. 국민의 소득 수준이 높은 국가일수록 출산율이 낮음

5. 한계점 및 느낀 점

<한계점>

- 결측치가 존재하는 행을 모두 제거하였으나 실제로는 데이터손실율을 최소화하기 위해 평균대체법, 최빈수대체법, 다중대체법 등의 기법을 고려해야함

<느낀 점>

- 본래 총 6개의 독립변수를 고려하였으나 데이터 수집에 어려움이 있어 3개의 독립변수밖에 분석하지 못하여 다양한 변수에 대한 검증을 하지 못한 아쉬움이 있음
- 유의할 것이라고 생각한 변수가 실제로 유의하지 않게 나왔으며, 사람들의 생각과 의문점을 직접 분석하여 증명하는 것이 데이터 분석가의 역할임을 깨달을 수 있었음

감사합니다 ☺