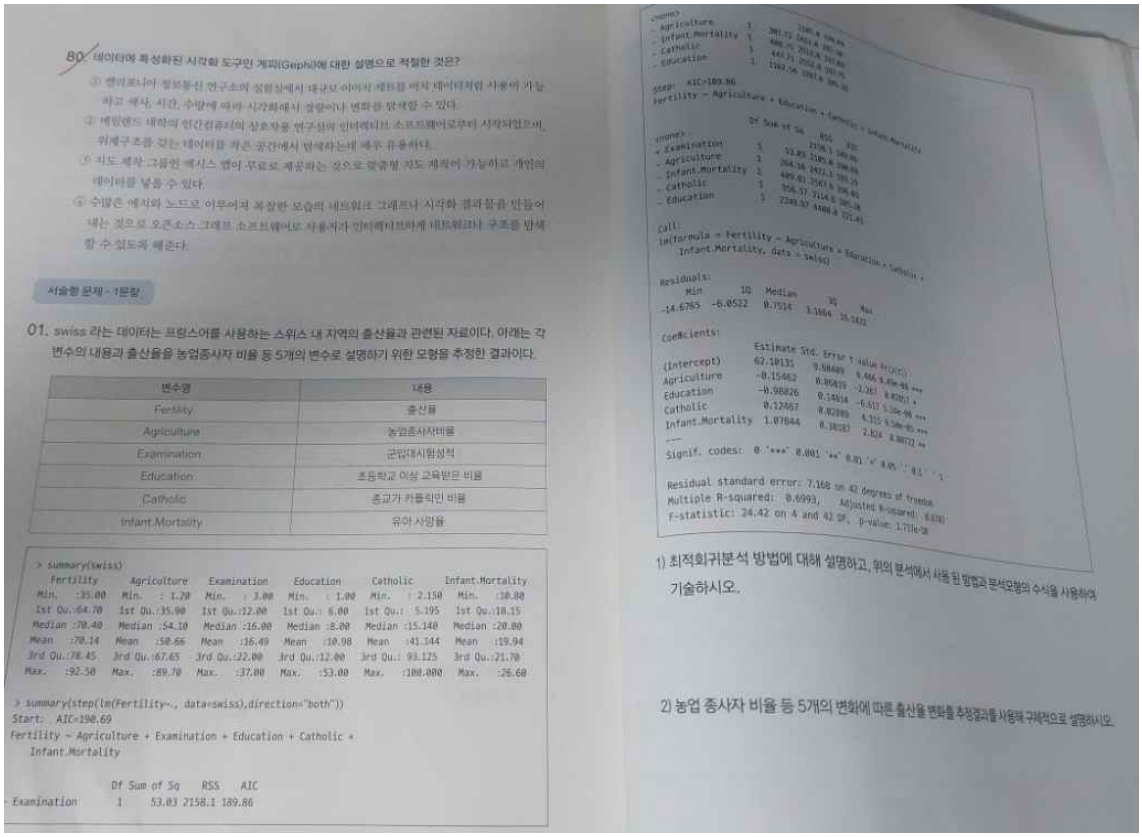

80 데이터에 특성화된 시각화 도구인 게피(Gephi)에 대한 설명으로 적절한 것은?

① 캘리포니아 정보통신 연구소의 실험실에서 대규모 이미지 세트를 비롯 데이터처럼 사용이 가능하고 색상, 시간, 수량에 따라 시각화해서 설정이나 변화를 탐색할 수 있다.

② 메릴랜드 대학의 인간컴퓨터의 상호작용 연구실의 인터랙티브 소프트웨어로부터 시작되었으며, 위계구조를 갖는 데이터를 자른 공간에서 탐색하는데 매우 유용하다.

③ 지도 제작 그룹인 맥시스 랩이 무료로 제공하는 강으로 맞춤형 지도 제작이 가능하고 개인의 데이터를 넣을 수 있다.

④ 수많은 에지와 노드로 이루어져 복잡한 모습의 네트워크 그래프나 시각화 결과물을 만들어 내는 것으로 오픈소스 그래프 소프트웨어로 사용자가 인터랙티브하게 네트워크나 구조를 탐색할 수 있도록 해준다.

서술형 문제 - 1문항

01. swiss 라는 데이터는 프랑스어를 사용하는 스위스 내 지역의 출산율과 관련된 자료이다. 아래는 각 변수의 내용과 출산율을 농업종사자 비율 등 5개의 변수로 설명하기 위한 모형을 추정한 결과이다.

변수명	내용
Fertility	출산율
Agriculture	농업종사자비율
Examination	군입대시험성적
Education	초등학교 이상 교육받은 비율
Catholic	종교가 카톨릭인 비율
Infant.Mortality	유아 사망율

```
> summary(swiss)
   Fertility      Agriculture     Examination      Education        Catholic      Infant.Mortality
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00   Min.   :  2.150   Min.   :10.80
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00   1st Qu.:  5.195   1st Qu.:18.15
 Median :70.40   Median :54.10   Median :16.00   Median :8.00    Median :15.140    Median :20.00
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98   Mean   :41.144    Mean   :19.94
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00   3rd Qu.: 93.125   3rd Qu.:21.70
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00   Max.   :100.000   Max.   :26.60

> summary(step(lm(Fertility~., data=swiss),direction="both"))
Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

              Df Sum of Sq    RSS    AIC
- Examination  1     53.03 2158.1 189.86
```

1) 최적회귀분석 방법에 대해 설명하고, 위의 분석에서 사용된 방법과 분석모형의 수식을 사용하여 기술하시오.

2) 농업 종사자 비율 등 5개의 변화에 따른 출산율 변화를 추정결과를 사용해 구체적으로 설명하시오.

1) 최적회귀분석 방법에 대해 설명하고 위의 분석에서 사용된 방법과 분석모형의 수식을 사용하여 기술하시오.

최적회귀분석 방법은 분석 데이터에 가장 잘 맞는 모형을 찾아내는 방법으로서 R 프로그램에서는 step 함수를 통해 종속변수에 대해 설명변수가 없을 경우부터 모든 설명 변수가 포함될때의 회귀모형을 비교해 최적의 회귀방정식을 도출할 수 있다. 또, R 프로그램에서 step 함수 안의 direction에서 'both'는 단계적 선택법(모든 가능한 독립변수들의 조합에 대한 회귀모형을 생성한 뒤 가장 적합한 회귀모형을 선택하는 방법), 'forwad'는 전진선택법. 'backward'는 후진선택법을 의미한다. 위의 분석에서 direction이 both로 입력되어 단계적 선택법을 사용했다. 위의 분석결과를 아래와 같은 순서로 단계를 나누어 결과를 해석할 수 있다.

@1단계 : 변수선택법을 결정하고 초기 모형을 설정한다.

위의 분석 결과에서 dtrection이 both로 설정되어 변수선택법을 단계적 선택법으로 선정했음을 확인할 수 있다. 또, 초기 모형은 Fertility ~ 으로 설명변수가 모두 포함된 상태에서부터 시작함을 의미한다.

@2단계 : 선택된 최적 모형의 AIC를 계산한다.

분석 결과에서 시작 모형은 Fertility~ 이 최적 모형으로 설정되어 있으며 start에서 AIC값이 190.69로 계산되어 있다.

@3단계 : 선택된 모형에서 변수를 추가/삭제 할 경우의 각 모형의 AIC를 계산한다.

Fertility~ 모형에 대하 설명변수 5개에 대한 각각의 AIC값을 계산하여 자유도 등과 함께 나타낸다. Examination의 AIC 값이 189.86, Agriculture의 AIC 값은 195.10 등으로 나타난다. 그리고 모형은 Fertility~ Agriculture + Examination + Education + Catholic +

Infant.Mortality 으로 나타나 있다.
@4단계 : 각 모형에서 최소의 AIC 모형을 선택하여 최적 모형을 선정한다.
계산된 AIC값을 비교하여 190.69보다 작은 설명변수인 Exmination을 제거하여 최적 모형으로 선정한다.
@5단계 : 2~4단계를 반복하여 AIC가 더 이상 줄어들지 않을 때 최종모형을 최적의 모형으로 선정한다.
위의 과정을 반복하여 최적의 모형을 선정하고 마지막 Steop에서 AIC가 189.86으로 계산되고 이 값보다 작은 값이 없어 변수를 모형에 추가, 삭제하지 않고 최적의 모형을 Fertility~ Agriculture + Examination + Education + Catholic + Infant.Mortality 로 선정했다.
@6단계 : 다변량회귀분석에서 종속변수인 출산률에 대한 설명변수들간의 모형에 대한 통계적 타당성을 가설검정한다.
귀무가설 : 설명변수는 모두 0이다.
대립가설 : 적어도 하나의 설명변수는 0이 아니다.
F-통계량은 24.42이며 p-value 값이 1.717e-10으로 귀무가설의 기각역인 0.05보다 작게 나타남에 따라 유의수준 5%하에서 대립가설을 채택하게 된다. 그러므로 추정된 회귀모형은 통계적으로 매우 유의함을 알 수 있다.
@7단계 : 다변량 회귀분석에 활용된 각 설명변수들의 계수들에 대한 통계적 타당성을 가설검정한다.
- 첫 번째 설명변수인 Agriculture에 대한 통계적 가설검정을 실시한다.
t-통계량은 -2.267이며 p-value 값이 0.02857아므로 귀무가설의 기각역인 0.05보다 작게 나타남에 따라 유의수준 5%하에서 대립가설을 채택하게 된다. 그러므로 추정된 회귀모형의 첫 번째 설명변수인 Agriculture는 통계적으로 유의함을 알 수 있다.
두 번째 설명변수인 Education의 경우 t-통계량은 -6.617이며 p-value값이 5.14e-08이므로 귀무가설의 기각역인 0.05보다 작게 나타남에 따라 유의수준 5%하에서 대립가설을 채택하게 된다. 그러므로 추정된 회귀모형의 모든 설명변수는 통계적으로 유의함을 알 수 있다.
@8단계 : 통계적으로 유의성을 확인한 다변량 회귀모형이 전체 데이터를 얼마나 잘 설명하는지 확인하기 위해 결정계수()를 확인한다.
- 결정계수를 확인하기 위해 Multiple R-squared와 Adjusted R-squared를 확인한결과, 0.6993과 0.6707로 나타났으며, 이는 전체 데이터를 설계된 다변량 회귀모형이 69.33%, 67.07%를 설명하고 있다고 해석할 수 있다.

2) 농업종사자 비율 등 5개의 변화에 따른 출산율 변화를 추정한 결과를 사용해 구체적으로 설명하시오.
최종적으로 다변량 회귀분석 결과를 종합해보면 다변량 회귀식이 추정된다. 회귀식을 통해 Education, Agriculture가 증가할수록 출산율(Fertility)는 감소하고 Catholic 등은 증가할수록 출산율이 증가하는 것을 확인할 수 있다. 그리고 출산율에 가장 영향을 많이 미치는 변수는 Infant이기 때문에 다른 변수들에 비해 많은 신경을 써야한다.