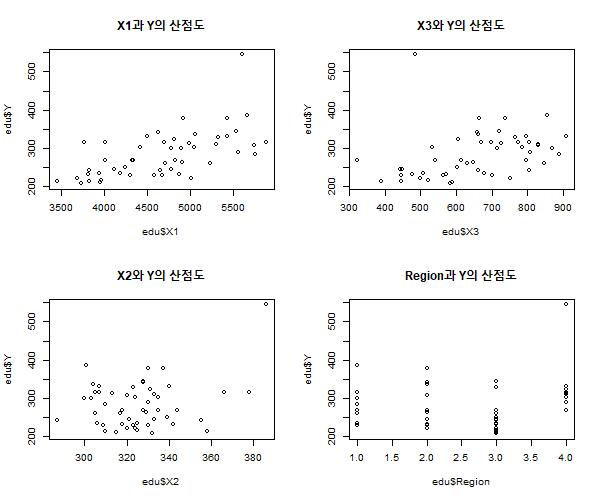
**회귀분석입문 HW3**

데이터사이언스융합전공2020312505 김민

1. **데이터 소개**

분석한 자료는 1975년의 미국 주별 교육비 지출에 대한 데이터이다. 분석을 통해 예측하고자 하는 값인 **반응변수 Y는** **인구 1인당 공립 교육 지출액**이다. 설명변수로는 인구 1인당 평균 소득 (X1), 인구 1000명당 18세 이하 인구의 수 (X2), 인구 1000명당 도시에 거주하는 인구의 수 (X3), 지역의 위치(Region)의 네 가지가 있다. Region변수는 Northeast(1), North Central(2), South(3), West(4)로 나뉘는 범주형 자료이다. Y, X1, X2, X3는 전부 수치형 자료이다. 총 관찰값은 50개이고, 결측치는 없다.



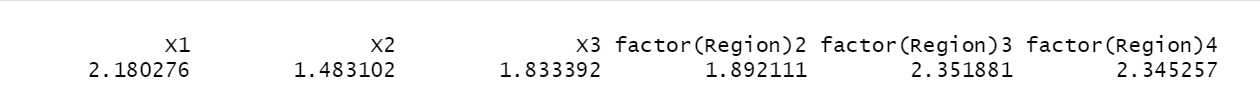
반응변수 Y와 각각의 설명변수들에 대한 산점도를 보면, X1과 X3는 Y에 대해 양의 선형관계를 갖고 있는 것처럼 보이지만, X2와 Y 사이에는 명확한 상관관계를 찾기 어렵다. 범주형 자료인 Region의 경우 1, 2, 3 지역에 비해 4지역의 특징이 두드러지는 것을 확인할 수 있다.

1. **변수 선택 과정**

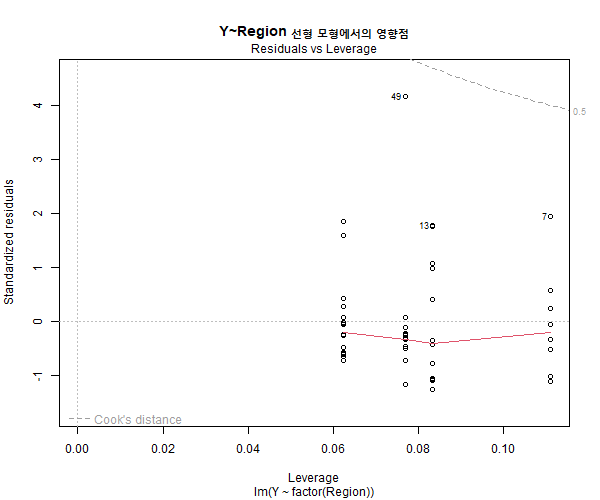
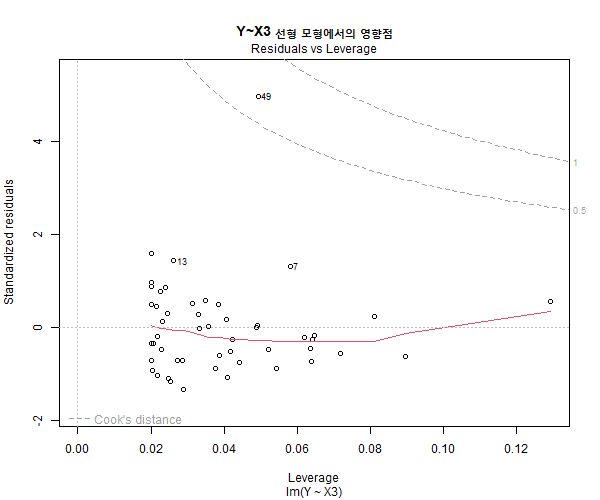
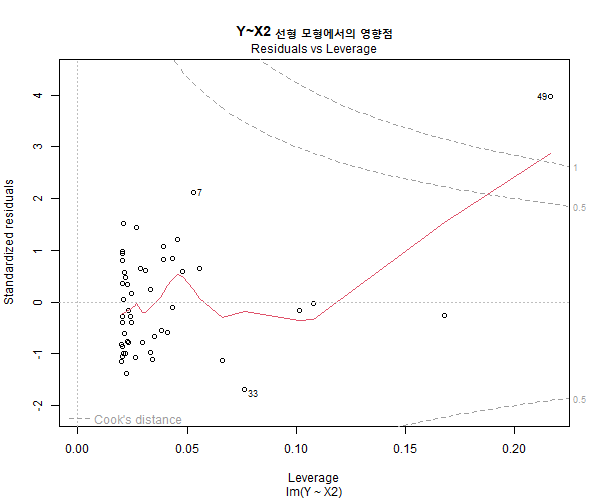
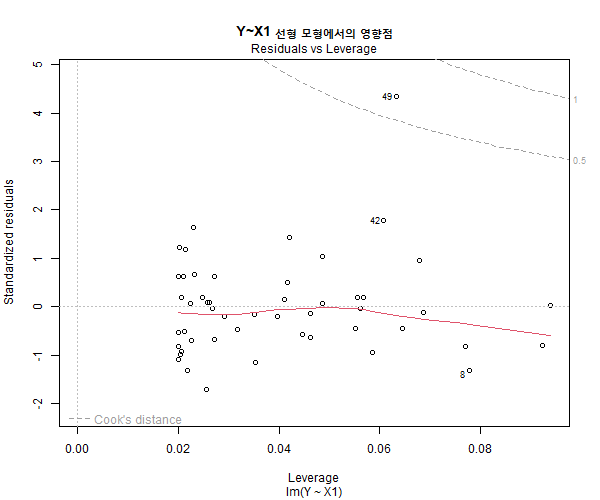
텍스트이(가) 표시된 사진

자동 생성된 설명

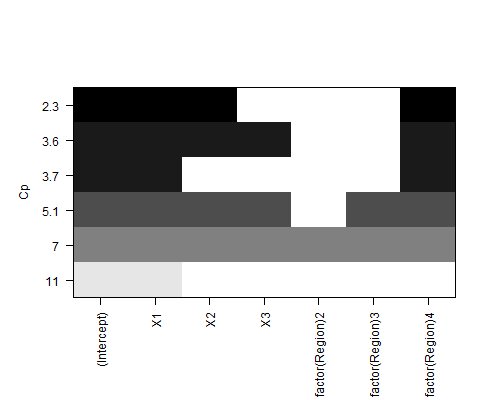
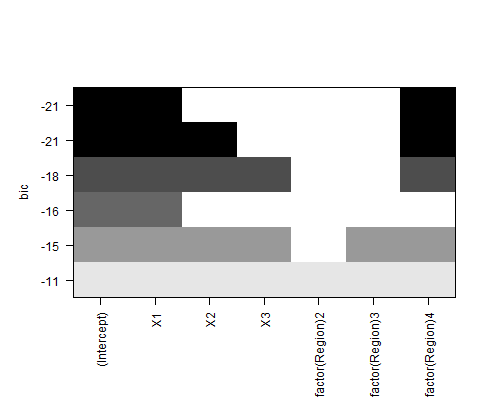
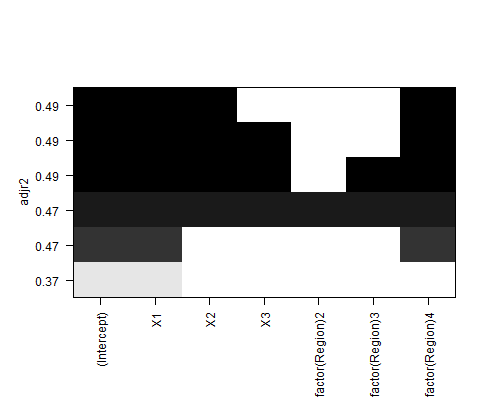
위 결과는 전체 변수를 전부 포함한 선형회귀모형에 대한 요약 보고서이다. 회귀모형에 대한 F통계량은 유의하지만, 개별 변수에 대한 t통계량은 절편, X1, X2를 제외하고는 유의하지 않다. 전체 변수를 포함했음에도 설명력이 0.63정도로 높지 않고, 유의하지 않은 변수들의 p값이 크기 때문에 변수 선택을 진행하여 회귀 모형을 적합하는 것이 좋다고 판단했다.



위 결과는 전체 변수를 전부 포함한 선형회귀모형에 대해 VIF(분산팽창요인)을 구한 것이다. VIF가 10이 넘는 변수는 없기 때문에 다중공선성 문제는 없는 것으로 보인다.



각각의 설명변수를 하나씩만 사용하여 Y를 예측하는 단순선형회귀모형을 만들었을 때, 4개의 변수 모두에서 **49번 데이터가 이상치이자 영향점으로 작용했다**. 영향점은 모형의 예측 결과에 부정적으로 작용할 수 있기 때문에 해당 데이터를 제거했다. 다시 동일한 조건에서 진단한 결과, 등분산성과 정규성이 영향점 제거 전보다 개선되어 보다 정확한 예측을 위한 회귀모형의 설계가 가능해졌다. 영향점을 제거한 뒤, 최적의 변수 조합을 찾기 위해 best subset selection을 진행하였다. 이는 전체 변수 조합 중 가장 Y에 대한 예측력이 좋은 변수 조합을 찾는 기법이다.



위 그래프는 각 평가 지표에 따른 변수 조합을 나타낸 것으로, y축 기준 위쪽으로 갈수록 예측력이 좋다. (수정결정계수(adjr2)는 클수록, BIC와 Cp 통계량은 작을수록 좋다.) 흰색은 x축에 위치한 해당 변수가 빠진 것이고 검은색은 해당 변수가 포함된 모형이라는 의미이다.

위 통계량에 따라 **절편, X1, X2, Region=4만을 최종 선형회귀모형의 변수 조합으로 선택하였다**. **이 모형의 수정결정계수는 0.4937으로,** **전체 데이터에 대해 50% 정도의 설명력을 갖는다**. 전체 변수를 포함한 선형회귀모형의 수정결정계수가 0.5774였던 것과 비교하면, 변수를 절반으로 줄였음에도 설명력을 비슷한 수준으로 유지했다는 점에서 의미가 있다. 이 때 각 변수들의 계수는 다음 표와 같다.

|  |  |
| --- | --- |
| **변수** | **계수** |
| Intercept (절편) | -191.24601386 |
| X1 | 0.05433759 |
| X2 | 0.64988002 |
| X3 | 0 |
| Region (= 4) | 27.04761021 |

1. **모형 해석**

앞에서 설계한 선형 회귀 모형을 식으로 표현하면 다음과 같다.

(▲ Region이 1, 2, 3인 경우)

(▲ Region이 4인 경우)

위 선형회귀모형은 다음과 같이 해석할 수 있다. 다른 조건이 다 동일한 경우, 인구 1인당 평균 소득(X1)이 1 증가하면, 인구 1인당 공립교육 지출액(Y)이 약 0.0543만큼 높아진다. 다른 조건이 다 동일한 경우, 인구 1000명당 18세 이하 인구 수(X2)가 1 증가하면, 인구 1인당 공립교육 지출액(Y)이 약 0.6499만큼 높아진다. 동일한 조건일 때, region이West(4)라면, 인구 1인당 공립교육 지출액(Y)이 약 27.0476만큼 높다.

결론적으로, **평균 소득이 높을수록, 18세 이하 인구 수가 많을수록 공교육에 대한 지출액이 높아지는 경향**이 있으며, **서부지역은 다른 지역보다 공교육 지출액이 높은 경향**이 있다. 1000명당 도시거주 인구 수는 공교육 지출액에 대한 유의미한 선형 상관관계가 없는 것으로 볼 수 있다.

1. **요약 및 결론**

본 보고서는 1975년의 미국 주별 교육비 지출에 대한 데이터를 활용하여 인구 1인당 공립교육 지출액을 예측하기 위한 선형 회귀 모형을 적합하고 해석했다. 영향점인 관찰값 1개를 삭제한 뒤, best subset selection을 진행하여 X1, X2, Region=4 세 개의 변수를 선택하였다. 이 회귀모형은 약 50%의 설명력을 갖는다. 이 회귀모형에 따르면 평균 소득이 높을수록, 18세 이하 인구 수가 많을수록 공교육에 대한 지출액이 높아지는 경향이 있으며, 서부지역은 다른 지역보다 공교육 지출액이 높은 경향이 있다.