**회귀분석을 통한 의료비 예측**



|  |  |
| --- | --- |
| **과목** | 빅데이터마이닝  :머신러닝 |
| **담당교수** | 손대순 교수님 |
| **학과,이름** | 데이터테크전공 20183226신희진 |
| **제출일** | 2020.12.13 |

**목 차**

**Ⅰ. 서론**

1. 분석 개요
2. 분석 의미

* 분석 이유

**Ⅱ. 본론**

1. 데이터셋 설명

* 변수 설명

1. 데이터 분석

-데이터 전처리

-데이터 요약(상관관계/산포도)

1. 회귀분석

**Ⅲ. 결론**

1. 데이터분석 결과 및 해석

-회귀분석

1. 새로 알게 된 사실

**Ⅳ. 출처**

**I. 서 론**

**ⅰ. 분석 개요**

미국의 환자 가상 의료비를 포함한 데이터 셋(insurance)을 가지고 분석 예측을 할 것이다. 의료비를 종속 변수로 잡고, 나머지 변수들을 독립 변수로 둔 후 회귀분석을 할 예정이다. 그것을 통해 종속 변수가 독립 변수들에 적합한지에 대해 알아볼 것이다.

**ⅱ. 분석 의미**

-분석 이유

회귀분석으로 적합성을 확인하기 위해 먼저 주제를 정해야 했다. 의료관련 데이터를 가지고 회귀분석을 하고 싶었기에 그에 맞는 데이터를 찾던 중, 미국의 가상 의료비가 포함되어있는 데이터를 발견하였다. 의료비가 성별, 연령 등에서 적합성을 띄고 있는지 확인하고 싶었다. 그래서 “insurance”의 데이터를 찾아 회귀분석으로 타당성을 판단하기 위해 분석을 하게되었다.

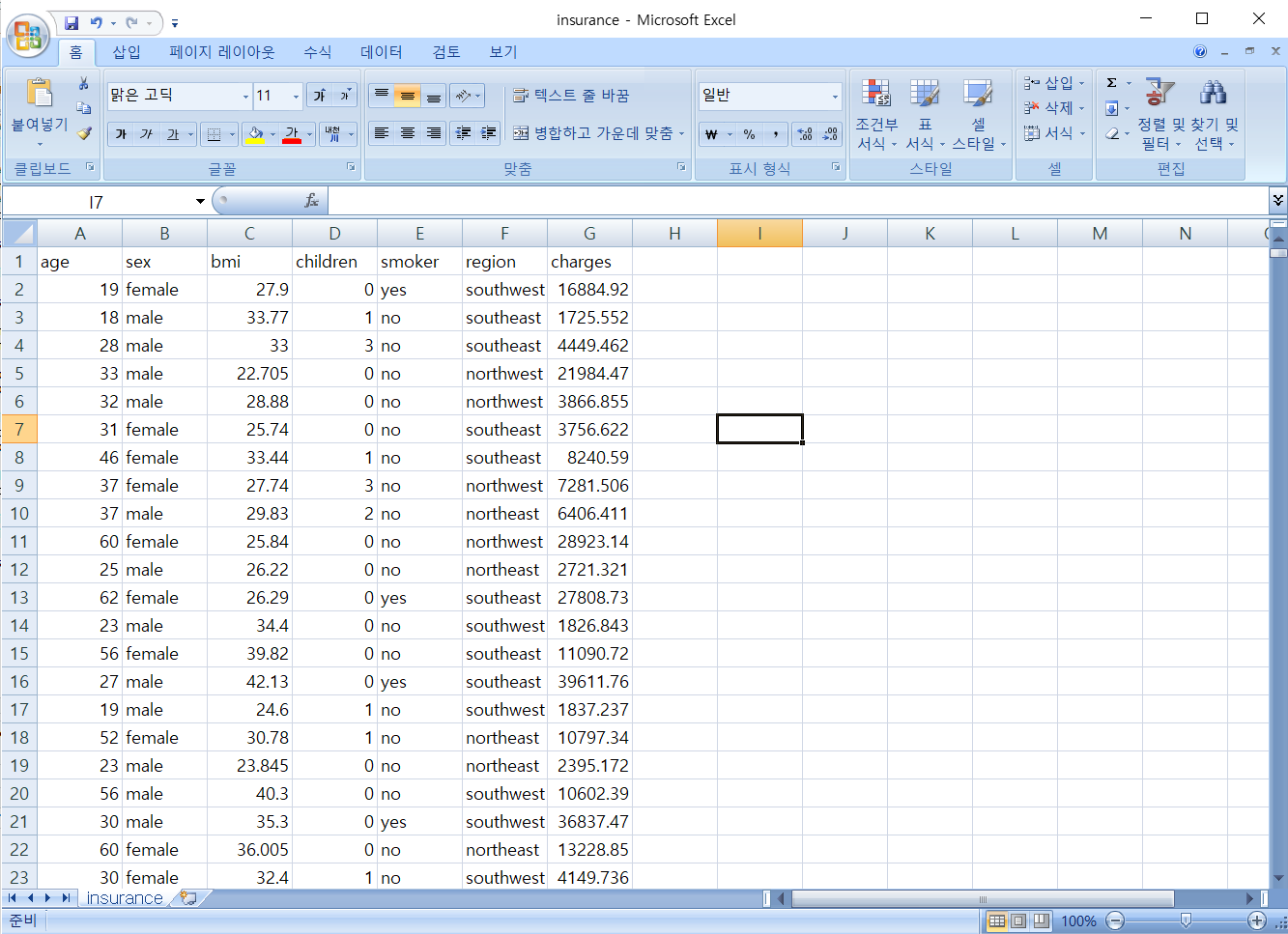
**Ⅱ. 본 론**

**ⅰ. 데이터 설명**

사용할 데이터는 미국 환자의 가상 의료비가 들어있는 데이터셋이다. 데이터셋에는 의료보험을 가입한 총 1338명의 수익자에 대한 정보이다. 총 7개의 변수를 가지고 있다. 사용할 데이터 “insurance.csv”을 구글에서 찾아 저장해준다.

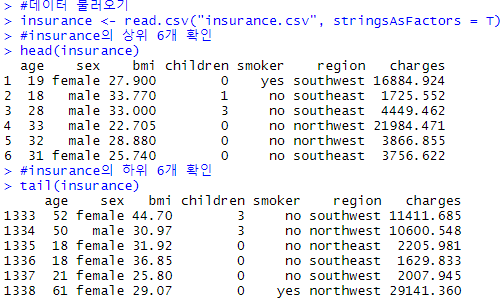
* Age : 수익자의 연령
* Sex : 계약자 성별(female/male)
* Bmi : 몸무게를 키의 제곱으로 나눈 값, 과체중 혹은 저체중인 사람의 키와 상관관계를 보여주는 신체 용적 지수
* Children : 의료보험이 적용되는 자녀 수/ 부양가족 수
* Smoker : 피보험자의 흡연여부(yes/no)
* Region : 미국 내 수익자 거주 지역
* Charges : 의료비

로 이루어져 있다. 사용할 종속 변수는 “charges”이고, 나머지 변수들(“age”, “sex”, “bmi”, “children”, “smoker”, “region”)이 독립 변수로 사용 되었다.

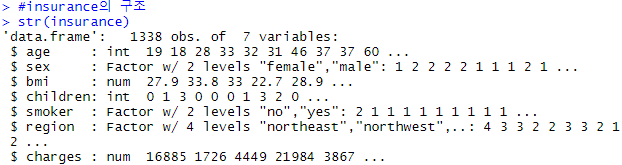


**ⅰⅰ. 데이터 분석**

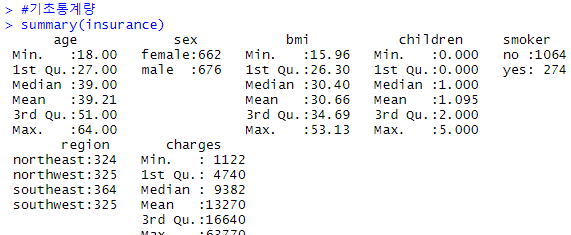
사용할 데이터 “insurance.csv”을 불러와 insurane에 저장해주고, 상위/하위 6개씩 확인한다.



Str을 사용하여 insurance의 구조를 살펴본다.

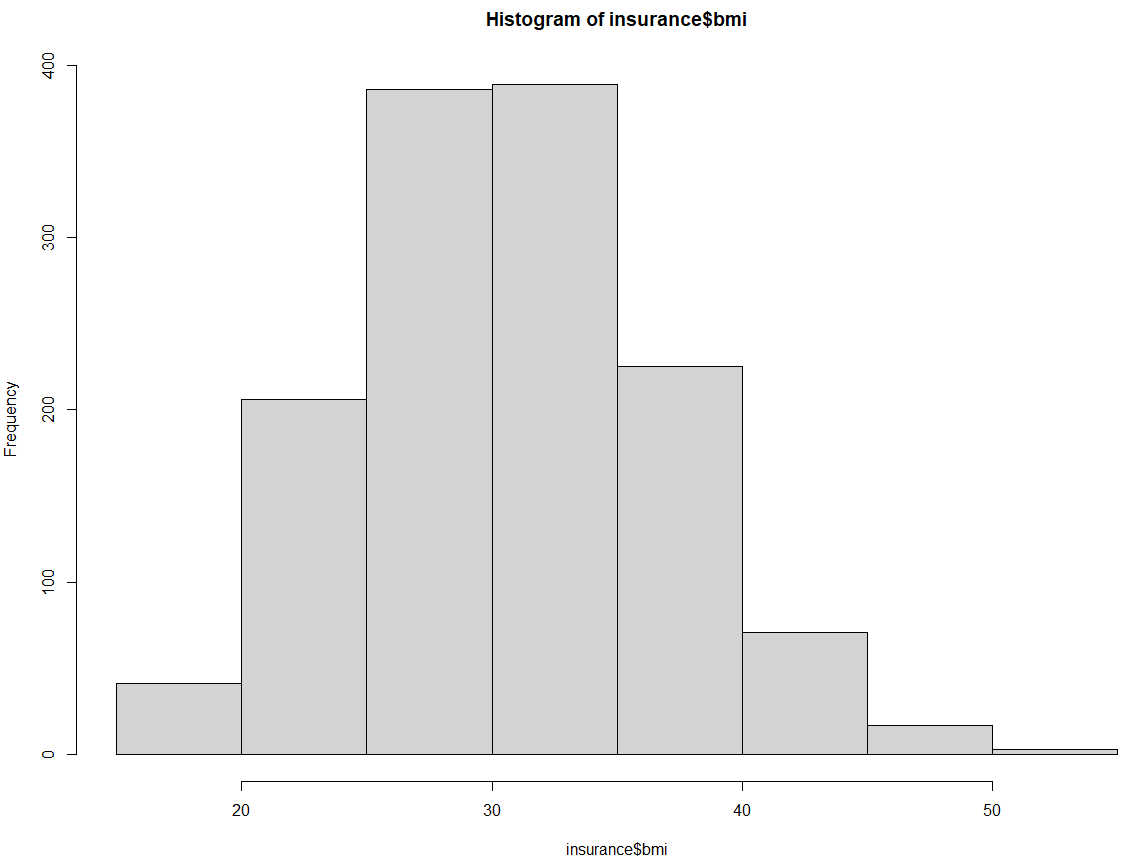


* 구조를 확인해본 결과, “age”, “bmi”, “children”, “charges”는 수치형 데이터로, “sex”, “smoker”, “region”은 범주형 데이터로 구성되어있는 것을 알 수 있다.



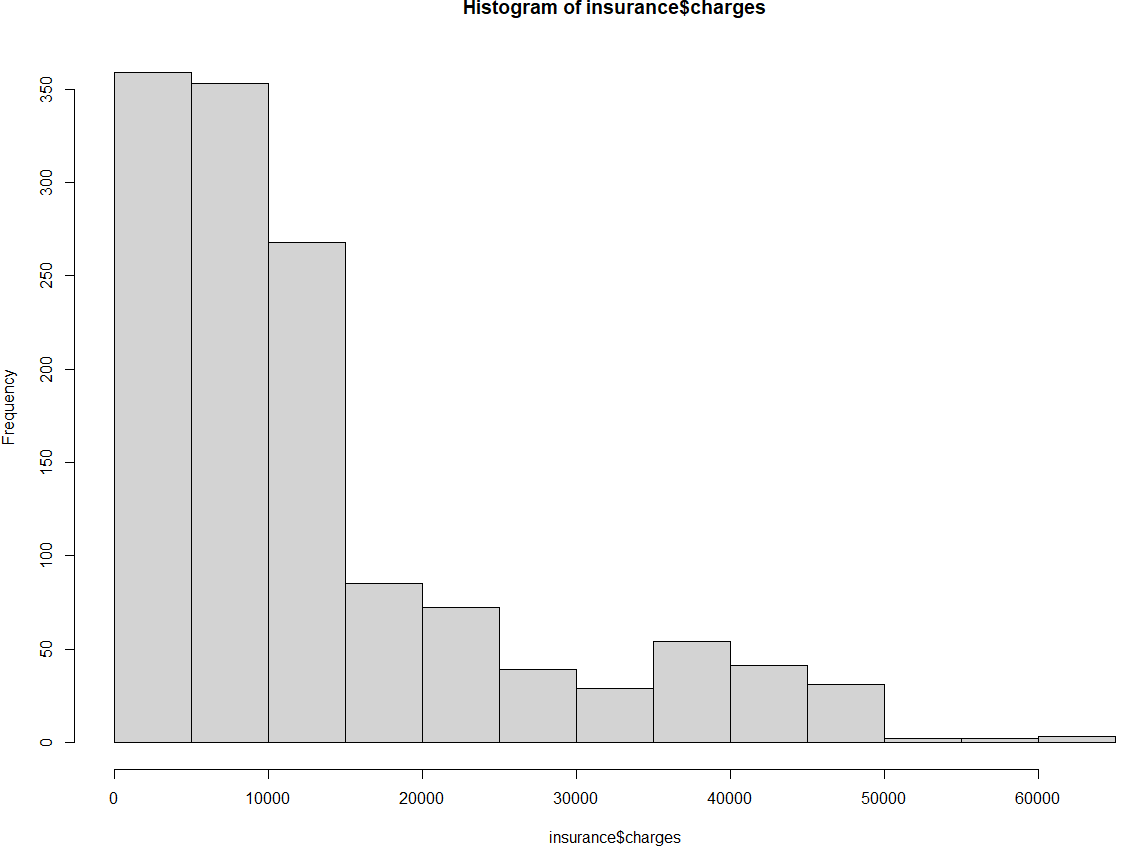
* 각각 7개의 변수의 최솟값, 제 1사분위수, 중앙값, 평균, 제 3사분위수, 최댓값이 포함되어 나와있다. “sex”와 “smoker” 그리고 “region”은 범주형 데이터로 성별의 수, 흡연여부, 지역사람의 수 등이 표기되어있는 것을 볼 수 있고, 나머지는 수치로 나타나있다.

데이터를 가지고 히스토그램과 상자그림을 그려서 시각화 해보았다. 시각화에 적합한 데이터는 “bmi”와 “charges”이므로 두 개의 변수의 그래프만 나타내보았다. 



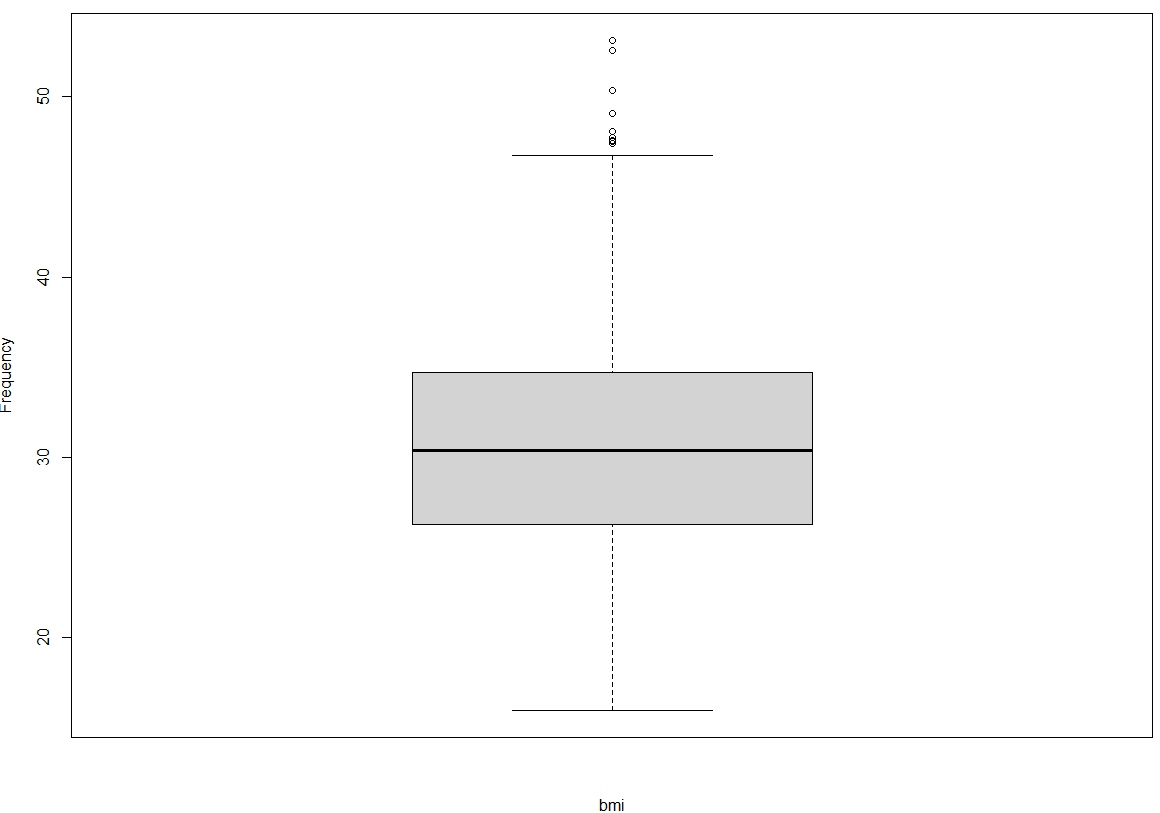
* Bmi의 정상 범위는 18.5 ~ 24.9 이므로 그래프를 확인해보면 과체중과 비만의 비율이 대체적으로 높다는 사실을 알 수 있다.





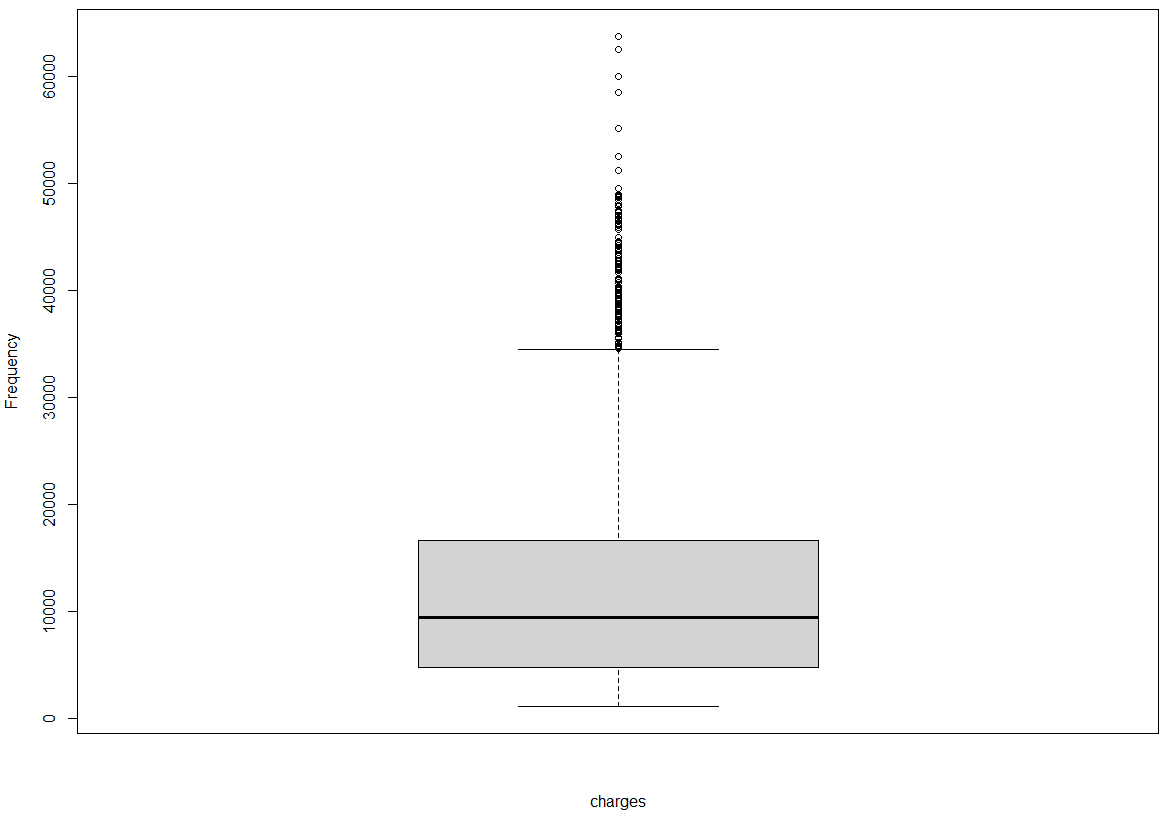
* 의료비의 그래프를 확인결과 주로 0에서 15000사이에 위치되어있다는 사실을 알 수 있다.





* 기초통계량에서 확인한 데이터로 bmi의 상자그림을 확인해 볼 수 있다.

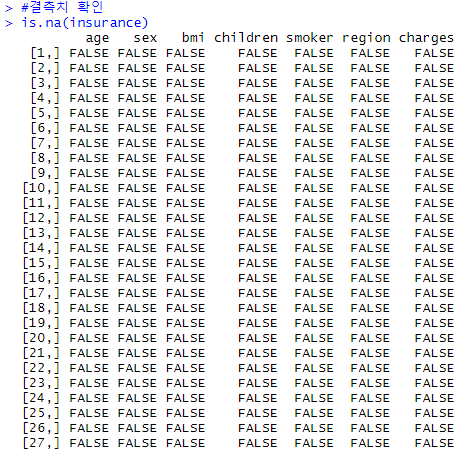




* 기초통계량에서 확인한 데이터로 charges의 상자그림을 그린 것이다. 위에 보이는 굵은 선은 데이터가 많이 몰려있어 진하게 보이는 모습이다.

**-데이터 전처리**

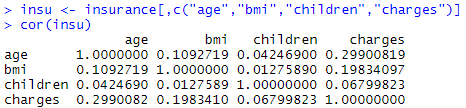
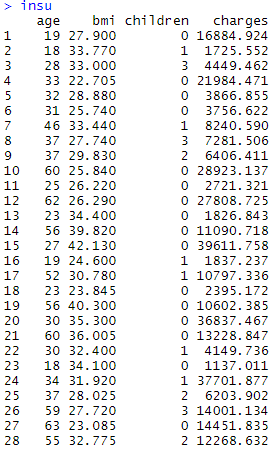
데이터에 결측치가 있으면 회귀분석이나 상관분석이 제대로 나오지 않을 수 있으므로 결측치가 존재하는지 확인해본다.



* 확인결과 모두 FALSE로 결측치는 따로 발견되지 않았다. 그러므로 전처리는 수치형 데이터만 담은 insu에 저장하여 사용하게되었다.

**-데이터 요약**

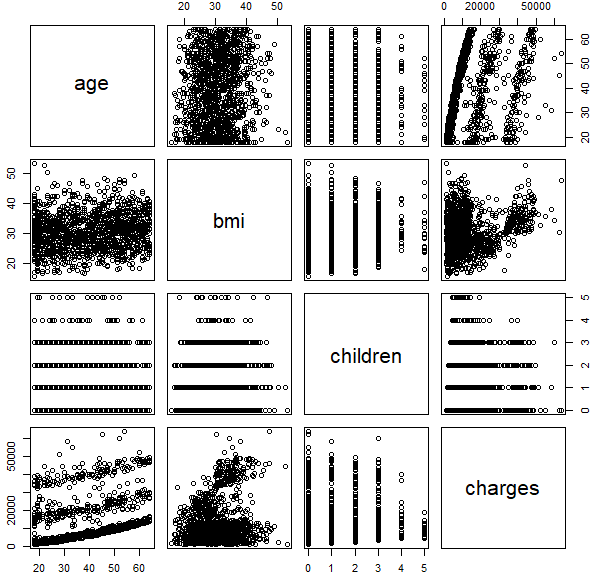
범주형 데이터를 제외한 수치형 데이터만 가지고 상관관계를 분석해보기 위해 insu에 수치형 데이터만 요약시켜 저장시켜준다.

* 결과를 보면 “age”와 “charges”가 0.299로 가장 높은 상관관계를 나타내고 있다.

뽑아온 데이터 insu를 가지고 산포도 행렬을 그려보았다.

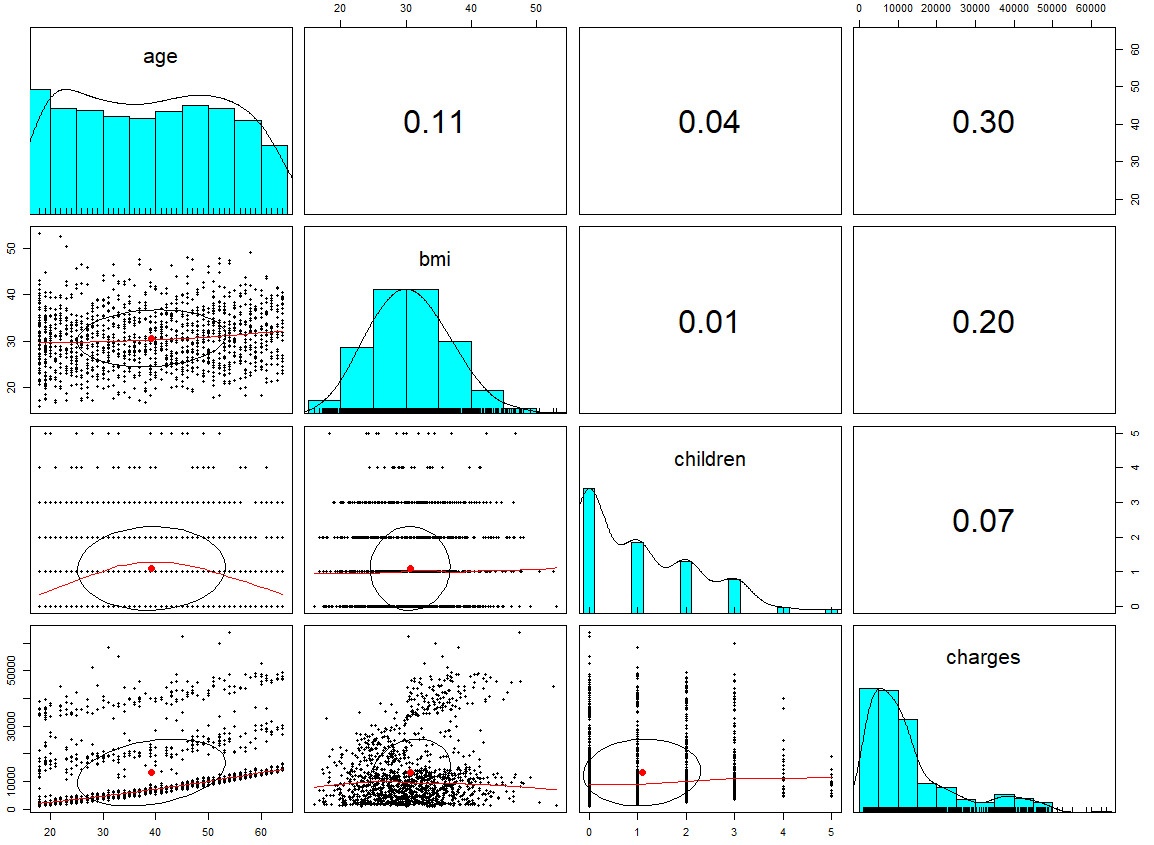




* 뚜렷하게 어떤 결과를 나타내는지 확인하기에는 문제가 있다. 하지만 “age”와 “charges”간의 관계는 일부 상대적인 직선을 나타내고 있다. 또한 “bmi”와 “charges”간에는 두 개의 다른 그룹을 갖고 있다는 사실을 알 수 있다.

위의 그래프는 자세히 보이지 않아 좀 더 유용한 시각화로 만들기 위해 다시 한번 그려보았다.

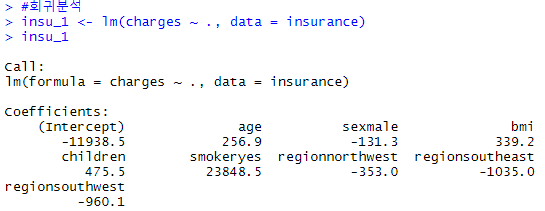




* “age”와 “charges”는 다른 변수들 간의 관계에 비해 완벽한 타원에 가까운 것을 볼 수 있다. 또한 “bmi”와 “charges”의 관계는 거의 원형에 가깝기 때문에 약한 상관관계를 갖는다고 볼 수 있다.

1. **회귀분석**

회귀분석을 하기 위해 lm(종속 변수 ~ 독립 변수, 데이터)를 이용하여 분석하였다. 종속 변수인 의료비와 6개의 독립 변수를 이용해 선형 회귀 모델을 적합화한다.

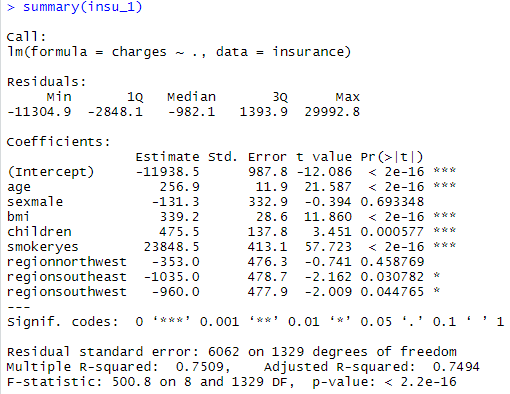


* 6개의 속성을 명시했지만 8개의 계수를 갖는다. 그 이유는 lm()함수가 모델에 포함한 각 백터 타입에 대해 자동적으로 더미 코딩이 사용되었기 때문이다. “smokeryes”, “sexmale”, “regionnortheast”, “regionnothwest”, “regionsouthwest”의 변수는 음의 상관관계를 갖는 것을 볼 수 있다. 결론적으로는 보험료는 나이가 많을수록, 비만, 흡연과 같이 부가적인 건강문제와 연관되는 경향을 보인다고 해석이 되었다.

**Ⅲ. 결론**

**ⅰ. 데이터 분석 결과 및 해석**

**-회귀분석**



* 위의 결과에서 오차 범위가 최대 30000까지 차이가 날 수 있다는 것을 의미한다. 하지만 50%의 오차가 1Q와 2Q 사이에 존재하기 때문에 예측의 대다수는 실제 값이 2580달러이상이고 1400달러 이하로 볼 수 있다. 위의 모델의 경우 R-squared값이 거의 75%에 근사 하는 값을 갖고 있으므로 나쁘지 않다고 보여진다. 또한 p-value값은 5%보다 작은 값을 가지므로 유의미한 모델이라고 해석이 된다.

**ⅰⅰ. 새로 알게 된 사실**

데이터 셋을 사용한 의료비 데이터를 사용하여 회귀분석을 해 보았다. 회귀분석을 배우면서 데이터를 모델로 정해 회귀분석을 하는 과정에서 어려움이 정말 많았다. 해석할 때 헷갈리는 점도 많고, 어떻게 해석해야 될지 막막했기 때문이다. 하지만 이번에 기말과제를 회귀분석으로 정한 이유는 그 전의 어려움을 극복하고자 하는 마음에서 회귀분석을 해보자 하는 마음도 있었다. 비록 부족한 해석을 하였지만 조금씩 코드를 써나가고 해석하는 과정에서 변수들의 상관관계와 연관성, 유의성들을 더 깊이 새롭게 알아가게 되었다. 또한 독립 변수가 어떻게 종속 변수와 관련이 되는지도 알았다. 데이터들의 변수 간의 좀 더 자세한 정보를 확인 할 수 있다는 점을 알아가게 되면서 유익한 시간 이였던 것 같다.

**-손대순 교수님에게 하고 싶은 말**

안녕하세요 교수님.

이번 학기 빅데이터 마이닝 수업을 들으면서 정말 유익한 시간 이였다고 말씀 드리고 싶습니다. 비록 매주 2개의 pdf을 통해 강의를 들었지만 이해가 쉽게 설명해주시고, 중간중간 예시를 들면서 이야기를 해주셔서 r프로그래밍으로 따라 하면서 수업을 들으니 더 쉽게 배웠습니다. 분석수업을 혼자 들으면서 따라 하면 혼자 못 따라가고 버벅거리지 않을까 고민도 되었지만 pdf로 세세하게 설명이 되어있어서 더 빨리 배울 수 있게 되었던 것 같습니다. 이번 학기 수업을 들으며 알아가는게 더 많은 수업 이였던 것 같습니다. 감사합니다

**Ⅳ. 출처**

<https://m.blog.naver.com/PostView.nhn?blogId=bestinall&logNo=221580078436&proxyReferer=https:%2F%2Fwww.google.com%2F>

<https://terms.naver.com/entry.nhn?docId=5141772&cid=60266&categoryId=60266>