|  |
| --- |
| **빅데이터개론11주차 실습과제** |
| **실습 #2 보고서** |

|  |  |
| --- | --- |
| **이름** | 박범철 |
| **학번** | 20185241 |
| **소속**  **학과/대학** | 빅데이터전공 |

## <주의사항>

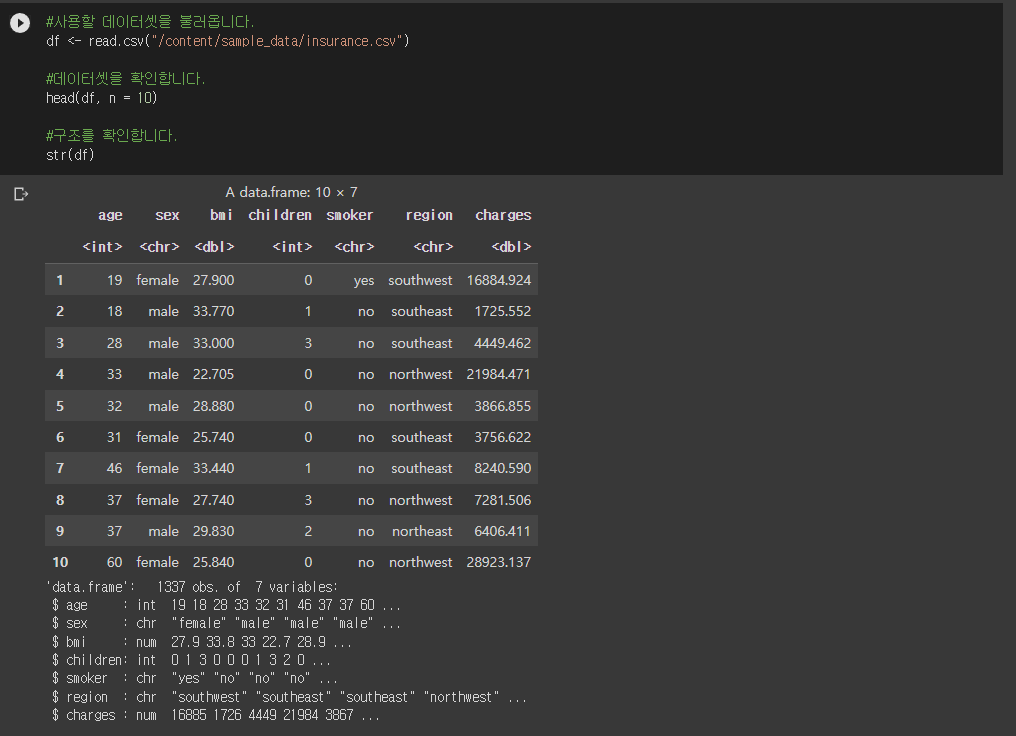
* 각각의 문제 바로 아래에 답을 작성 후, Bigdata Project.ipynb 파일을 SmartLEAD에 제출할 것
* SMARTLEAD 제출 데드라인: 2022. 11. 07(월요일) ~ 11.13(일요일) 23:59 까지
  + 부정행위 적발 시, 원본(보여준 사람)과 복사본(베낀 사람) 모두 0점 처리함
  + 예외 없음
* 해당 문제의 코드와 출력 결과를 반드시 기록할 것
  + 둘 중 하나라도 없으면 해당 문제 0점 처리

## <실습 과제>

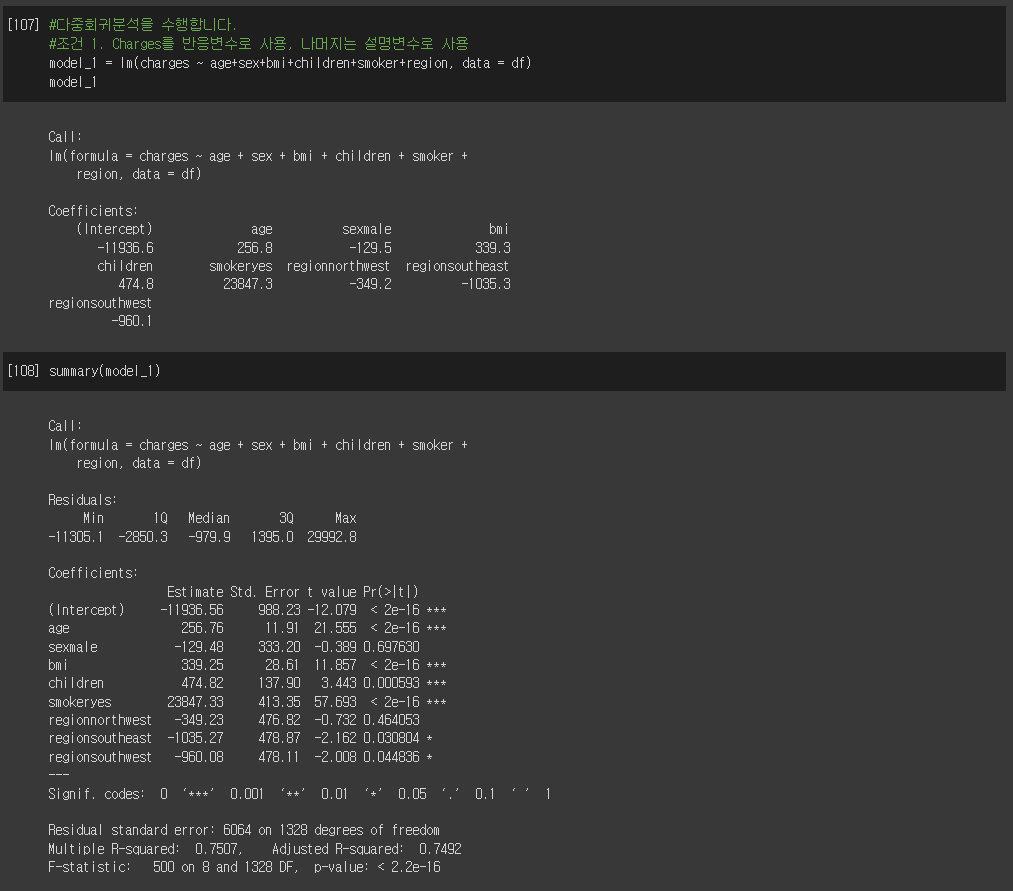
|  |
| --- |
| [Q 1] **[50점] – 다중회귀분석(Multiple Linear Regression)**  **<사용되는 데이터 – insurance.csv >**  **Kaggle에서 제공하는 의료비 예측 데이터셋을 이용하여 다중 회귀분석을 해보자!**  <https://www.kaggle.com/datasets/mirichoi0218/insurance>  **<조 건>**   1. **Charges를 반응변수로 사용하고 그 외 나머지를 설명변수로 사용.** 2. **createDataPartition을 이용하여 Train/Test set 분리 후 회귀모델 생성.** 3. **Summary함수를 사용하여 회귀분석 결과 해석을 작성.** 4. **분석 결과에 기반하여Feature Selection(변수 선택) 및 새로운 회귀모델 생성.** 5. **Test set을 이용한 Predict(예측)결과 비교.** |

답변:

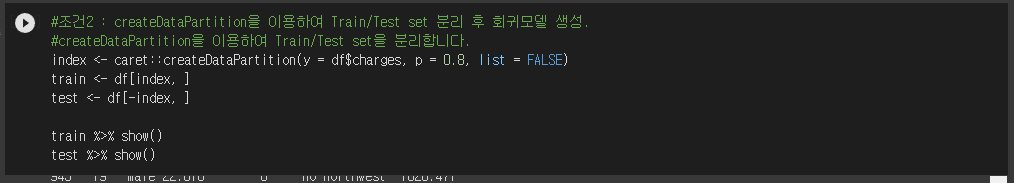
1. **소스 코드 및 결과 (30 점)**

****

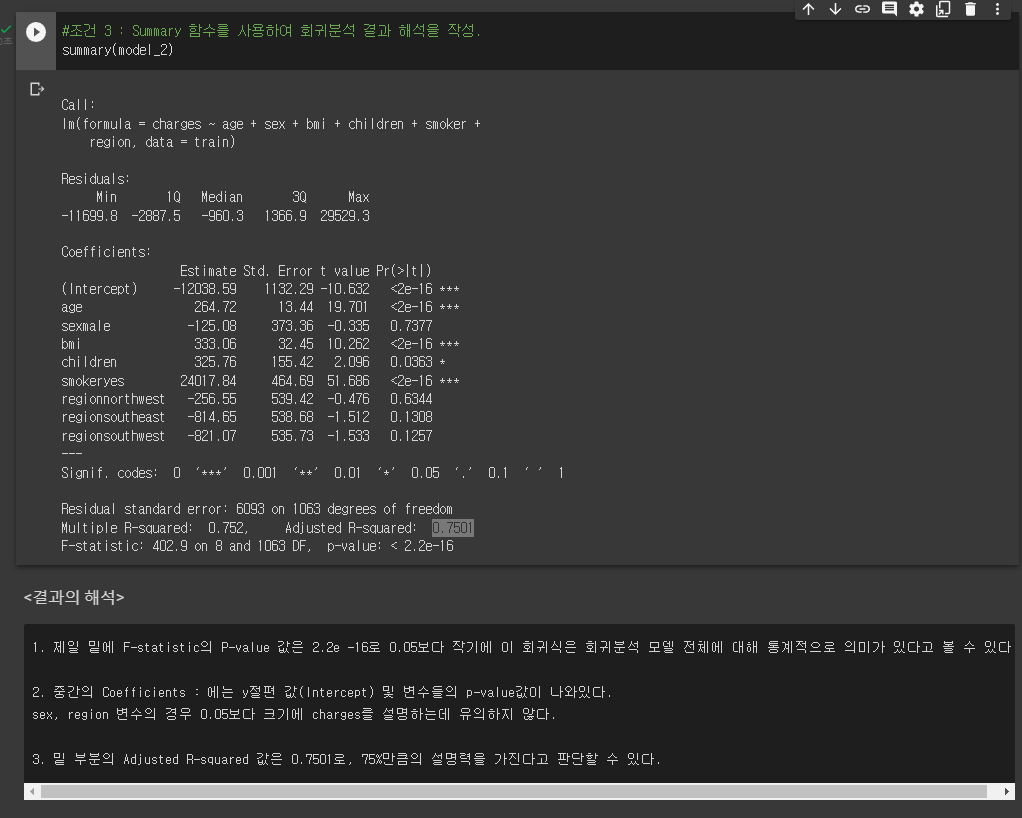
* + - 1. **데이터 셋을 불러오고 데이터셋이 제대로 동작하는지 확인합니다.**

****

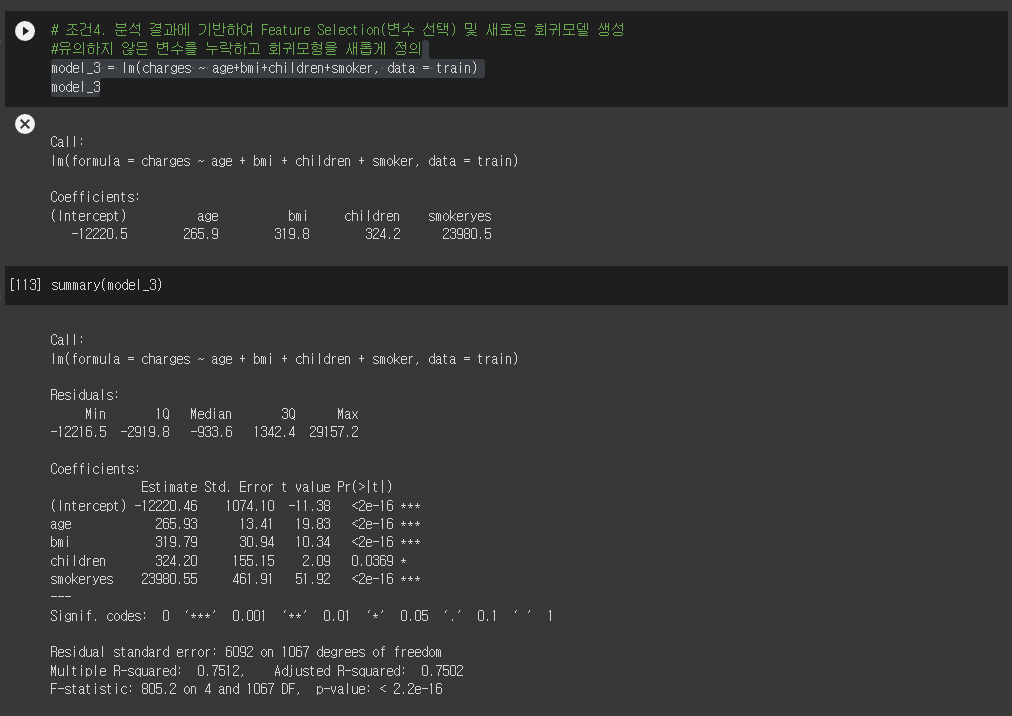
* + - 1. **첫번째 조건인 Charges를 반응변수 나머지를 설명변수로 사용합니다.  
         lm()함수를 사용하여 lm(반응변수 ~ 설명변수)형식으로 구성합니다.  
         summary()함수를 사용하여 제대로 회귀분석이 됐는지 확인합니다.**

****

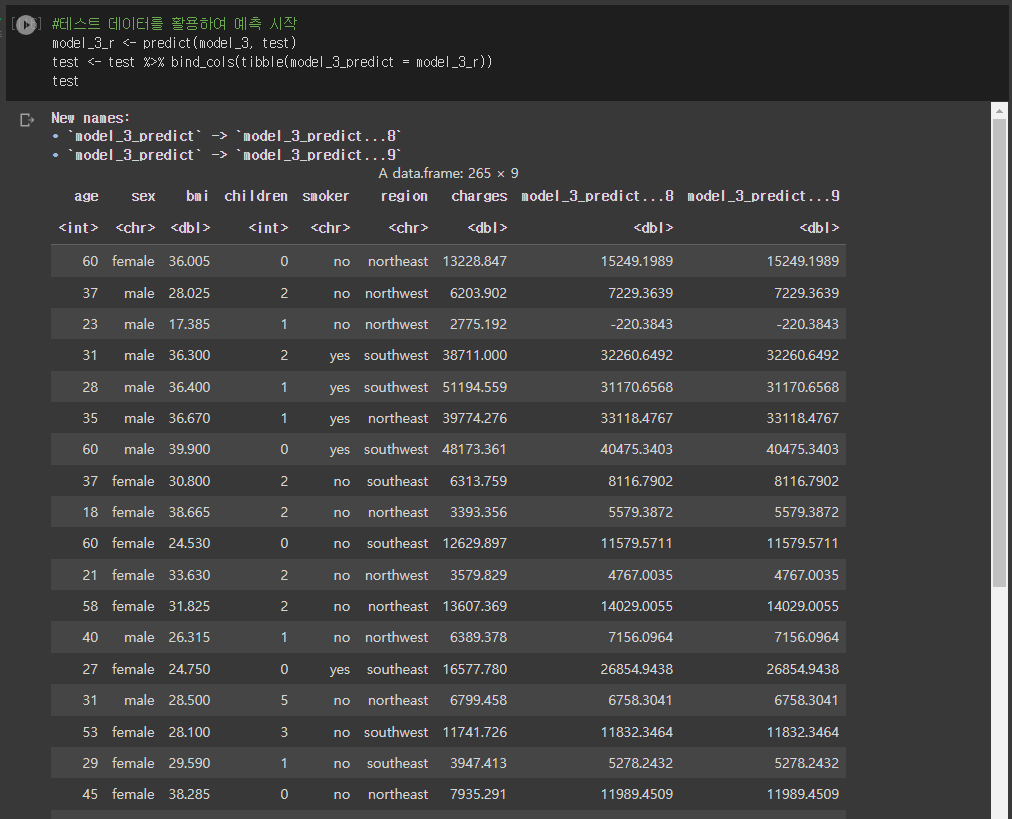
* + - 1. **두번째 조건인 caret라이브러리의 createDataPartition을 사용하여 Train set과 Test set을 분리하여 회귀모델을 생성합니다.  
         p = 0.8로 설정하여 8:2 비율로 분류한다.**

****

* + - 1. **Summary 함수를 사용하여 회귀분석 결과 해석을 작성하시오.  
         P-value 값은 0.05보다 크면 통계적 의미가 없다고 본다.  
         그러므로 0.05보다 큰 sex, region값은 charges를 설명하기 유의하지않다.**
      2. **R-squared 값은 정확율을 나타낸다. 0.7501이므로 75%만큼의 설명력을 가진다.**

****

* + - 1. **유의하지 않는 변수는 누락한다 (sex, region)  
         그 후 새로운 회귀모형을 새롭게 정의한다.**

****

* + - 1. **조건 5번 테스트 데이터를 활용한다.**

**predict함수를 사용하여 예측을 진행했다. Model3의 값과 charges의 값을 비교해보면 비슷하게 예측하는 것을 볼수 있다.**

1. **결과 설명 (20 점)**

추정치 부분에서는 각 변수에 따른 회귀계수이다. 단순회귀 분석과 달리 다중회귀분석은 여러 개의 독립변수가 존재한다. 그것의 유의수준에 따라서 결과값이 달라질 수 있다. 그렇기에 우리가 실제로 필요한, 중요도에 따라 다르게 데이터를 구성하여 가져온다. Model\_3같은경우 sex, region을 제외하고 회귀분석을 실시했다. P값pr(>|t|)은 작으면 작을수록 회귀계수가 유의하다는 것을 나타낸다. 상관관계를 알고싶으면 확인하면 된다. 다중회귀분석 모델인 model\_3를 사용하여 charges와 비교해본다. 모델의 성능은 75%을 보여준다. 직접 비교해도 비슷한 값을 유추해내고 있으며 25%정도는 근소한 차이, 큰 차이를 보여주고 있다. Adjusted R-squared : 0.7502 🡪 설명변수 개수에 대한 패널티를 적용한 결정계수를 설명하고 있다. F

## <실습 과제>

|  |
| --- |
| [Q 2] **[50점] – 로지스틱 회귀분석(Logistic Regression)**  **<사용되는 데이터 - Stroke Prediction Dataset.csv >**  **Kaggle에서 제공하는 뇌졸증 예측 데이터셋을 이용하여 로지스틱 회귀분석을 해보자!**  <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>  **<참고사항>**   1. **Stroke Prediction 데이터를 로드 후 다음과 같은 전처리 과정을 거친다.**  * **결측값 제거(bmi 컬럼값이 N/A로 작성되어 있는 값을 결측값으로 지정함)** * **결측값 제거 후 bmi 데이터타입 변경** * **Stroke 값이 0인 데이터를 랜덤으로 209개 추출** * **추출된 데이터와 stroke 값이 1인 데이터 병합**  1. **병합한 Stroke Prediction 데이터를 caret 패키지의 createDataPartition을 이용하여 train / test 으로 분리할 것** 2. **Train dataset을 사용하여glm 함수를 사용하여 모델을 생성하고 반응변수는 stroke로 설정하고, 설명변수는 stroke를 제외한 모든 것을 사용한다.**  * **step, update을 생성한 모델을 새롭게 정의할 것.** * **생성한 모델과 새롭게 정의한 모델의 summary결과를 확인한다.**  1. **Test set을 사용해서 모델을 예측한다. 예측 결과와 원본 데이터와 비교하여 모델의 적합성에 대하여 기술한다.** |

답변:

1. **소스 코드 및 결과 (30 점)**

**조건 1-1에 현재 결측값이 N/A로 되어있는 결측값인 NA로 바꿔준다. 현재 N/A라고 되어있는 함수를 전체함수를 바꿔주는 []를 사용하여 없애준다. 결측값을 없애기 위해 na.omit 함수를 사용하여 없애고, colSums로 결측값이 제거 됐는지 확인합니다. 그 후 조건 1-2의 bmi의 데이터 타입을 변경하기 위해 as.numeric 함수를 사용합니다. 조건1-3를 하기위해 stroke 값이 0인 것을 새로 저장 후 랜덤으로 209개를 추출하기위해 sample함수를 사용합니다. 조건 1-4를 하기위해 stroke 값이 1인 데이터를 따로 뽑아 낸후 미리 뽑아낸 0값을 가지고있는 변수 와 병합합니다.**

**조건 2인 createDataPartition함수를 사용하기위해 caret패키지를 설치 후 train데이터와 test데이터를 분류합니다. 그 후 조건 3을 하기위해 glm 함수를 사용합니다. 여기서 stroke를 반응변수, 나머지 전체를 설명변수로 설정합니다. 코드가 길어지는 것을 방지하기 위해 .을 사용하여 모든 설명변수를 나타냅니다. Family = “binomal”을 사용하여 종속 변수가 어떤 분포를 따르고 있는지 옵션을 주는 것을 확인합니다. 조건 3-1 을하기위해 m2, m3를 step,update함수를 사용하여 새로운 모델을 정의합니다. 새로운 모델을 predict함수를 사용해 모델을 예측합니다.**

1. **결과 설명 (20 점)**

**Test set 을 활용하여 결과를 확인했을 때 사건을 stroke 1 100%로 나타냈을 때 0.5 이상 “사건이 일어날 확률” 을 예측이 잘 됐음을 확인 할 수 있습니다. Stroke 0 0%로 나타냈을 때 0.5 이하 “사건이 일어나지 않을 확률”을 예측이 잘 됐음을 확인 할 수 있습니다. 하지만 중간 중간에 예측을 벗어난, 0인데 0.5를 넘거나 1인데 0.5이하의 예측을 보일 때가 있습니다. 예측된 데이터를 기반으로 하기 때문에 오차가 생길 수 있습니다.**

텍스트, 전자기기, 컴퓨터, 디스플레이이(가) 표시된 사진

자동 생성된 설명텍스트, 전자기기이(가) 표시된 사진

자동 생성된 설명