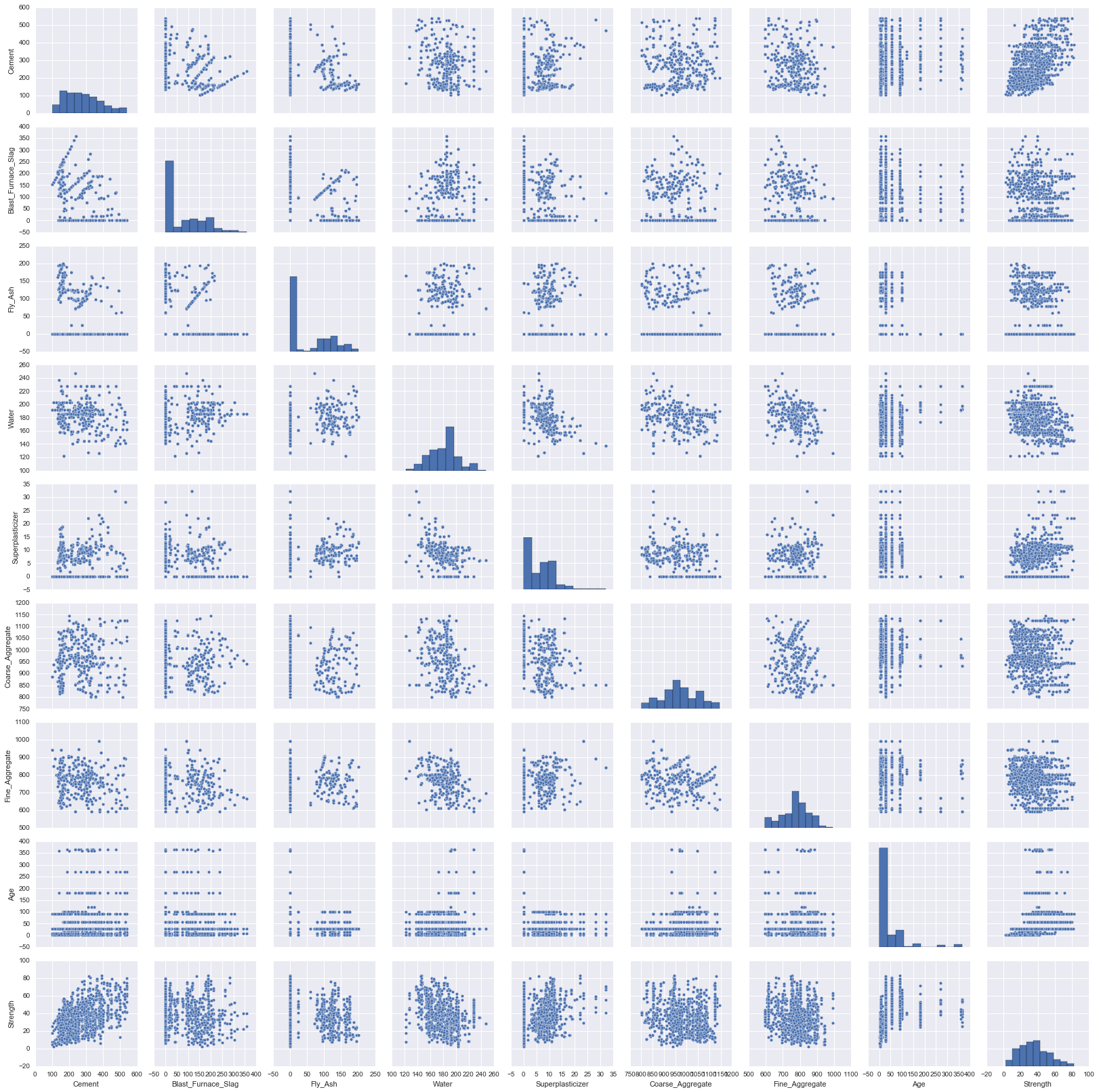
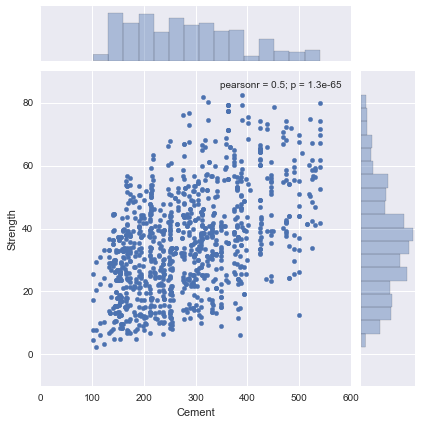
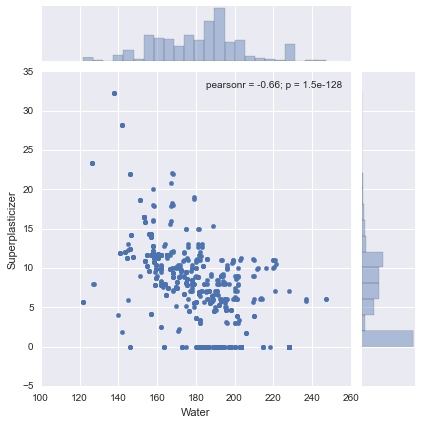
**Assignment01**

|  |
| --- |
| * 코드는 함께 업로드 된 Jupyter notebook에 작성 * 코드를 작성하고 화면에 print된 것을 그대로 제출 * 결과는 제시된 질문에 따라 이 워드 파일에 기입 * 최종적으로 notebook 파일과 이 워드 파일을 e-class에 같이 제출 * 제출시 파일명 뒤에 학번\_이름을 추가 (‘Assignment01\_학번\_이름’) |

1. 선형회귀(linear regression)를 이용하여 콘크리트의 성분의 따른 강도를 예측하는 모형을 개발하고자 한다. 사용 데이터는 ‘Concrete\_Data.txt’이며 타겟(target) 변수는 ‘Strength’로 마지막 열에 해당하고 나머지 변수는 설명 변수로 사용한다.

(1) 미리 데이터의 분포를 알아보기 위해서 모든 변수에 대한 pairplot을 그리시오. pairplot의 결과로부터 선형회귀를 이용해서 타겟을 예측하는데 가장 유용할 것이라고 판단되는 변수와 선형 관계 가정을 가장 위배하는 것처럼 보이는 변수를 각각 하나씩 기술하시오.



>> 타켓을 예측하는데 가장 유용할 것으로 판단되는 변수는 “Cement”. 타켓변수 ‘Strength’와 양의 상관관계를 갖는 것으로 보인다. Jointplot을 이용해 두 변수의 관계를 따로 살펴보았을 때 피어슨 상관계수가 0.5로 다소 양의 선형관계가 있다고 할 수 있다. 다른 설명변수의 피어슨 상관계수 값을 비교해도 가장 높은 것을 알 수 있다.  
한편 선형관계 가정을 가장 위배하는 것처럼 보이는 변수는 설명변수 간 가장 상관계수가 높은 변수라고 할 수 있다. 눈으로는 절대 모르겠다 싶어서, 반복문을 이용해 각각의 변수 중에서 피어슨 상관계수 절대값이 가장 큰 설명변수 쌍을 찾았다. 결과는 -0.6575 ('Water', 'Superplasticizer')

(2) 모든 설명 변수를 이용하여 강도(Strength)를 예측하는 모형을 학습하고 아래 표를 채우시오. 일 때 0이라는 가설을 기각할 수 없는 변수는 무엇인가?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 변수명 | 추정 계수() |  |  |  |
| Intercept | -23.3312 | 2.6585e+01 | -0.8775 | 0.1901 |
| Cement | 0.1198 | 8.4890e-03 | 14.1128 | 0.0 |
| Blast\_Furance\_Slag | 0.1038 | 1.0135e-02 | 10.2474 | 0.0 |
| Fly\_Ash | 0.0879 | 1.2583e-02 | 6.9881 | 2.5098e-12 |
| Water | -0.1499 | 4.0177e-02 | -3.7314 | 0.0001 |
| Superplasticizer | 0.2922 | 9.3424e-02 | 3.1279 | 0.0009 |
| Coarse\_Aggregate | 0.0181 | 9.3922e-03 | 1.9256 | 0.0272 |
| Fine\_Aggregate | 0.0201 | 1.0701e-02 | 1.8866 | 0.0297 |
| Age | 0.1142 | 5.4271e-03 | 21.0464 | 0.0 |

>> “Intercept”, “Coarse\_Aggregate”, “Fine\_Aggregate” 변수의 계수는 일 때 이라는 가설을 기각할 수 없다고 볼 수 있다. 그 외의 모든 변수의 계수는 일 때 0이라는 가설을 기각할 수 있다고 볼 수 있다. 각각의 t-통계량을 통해 p-value는 거의 0에 수렴하기 때문이다.

(3) 추정된 모형에 대한 ANOVA 표를 완성하시오. 전체적으로 이 모형을 강도를 예측하는데 쓸모가 있다고 할 수 있는가?

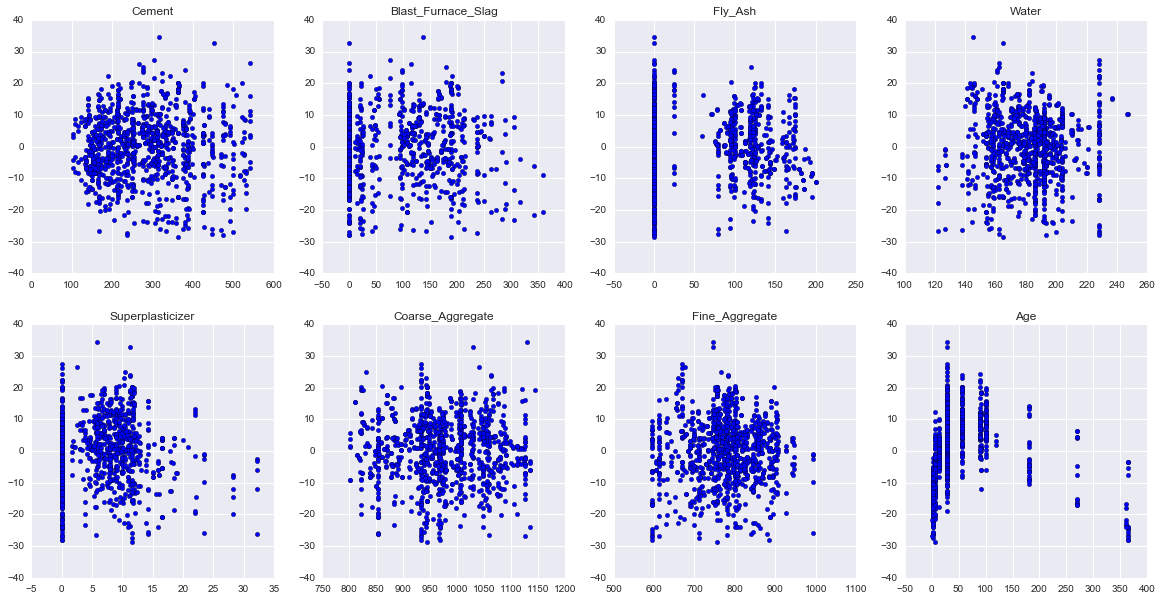
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SS | Degree of freedom | MS | F |  |
| Model | 176762.0339 | 8 | 22095.2542 | 204.3167 | 1.1102e-16 |
| Residual | 110413.1531 | 1021 | 108.1421 |  |  |
| Total | 287175.1871 | 1029 |  |  |  |

>> 그렇다. 이 모형은 강도를 예측하는데 쓸모가 있다고 할 수 있다. p-value가 일반적인 유의수준 0.05보다 훨씬 작은 것을 통해 이 모형의 모든 계수가 0이 아니라고 할 수 있기 때문이다.

(4) 와 값을 구하시오.

>> = 0.6155 = 0.6125

(5) 모든 설명 변수에 대해 개별 설명 변수를 x축으로 하고 y축을 잔차(residual)로 하는 산포도(scatter plot)을 그리시오. 산포도를 보고 알 수 있는 사항에 대해서 기술하시오.



>> 잔차의 평균은 0이다. 선형회귀식의 기본가정 중 오차와 관련된 가정으로 등분산성(Homoscedasticity)과 정규성(Normality)이 있다. 선형회귀식에서의 오차는 정규분포를 따른다는 특징이 정규성이고, 그 오차는 x값과 상관없이 동일한 분산을 갖는다는 특징이 등분산성이다. 잔차에 대한 개별 변수에 대한 산포도를 그려서 확인할 수 있는 부분은 등분산성이다.

결과를 확인해 보면, 평균 0을 기준으로 대체적으로 비슷하게 잔차가 퍼져 있는 것을 확인 할 수 있다. 하지만 “Age” 변수의 잔차 산포도를 보면, 선형이 아닌 2차 곡선으로 bias되었다는 것을 알 수 있다. 이를 해결하기 위해서 새로운 변수를 생성해서 다시 모델링을 하는 것이 좋다고 한다. 다른 변수에 대해서는 눈으로 확인하는게 한계가 있기 때문에 오차에 대한 분산이 설명 변수에 의해 영향을 받나, 안받나 모델을 만들어서 검정을 해야한다.

(6) 잔차에 대해서 Jarque-Bera test를 수행하시오. 잔차는 정규분포를 따른다고 볼 수 있는지 기술하시오.

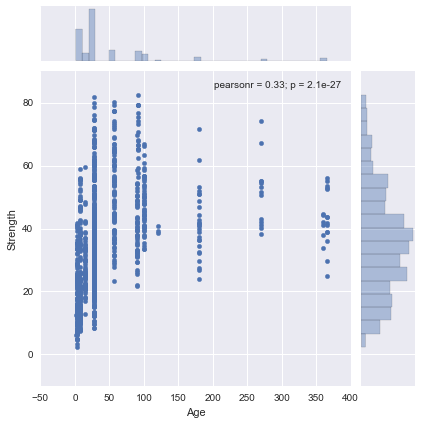
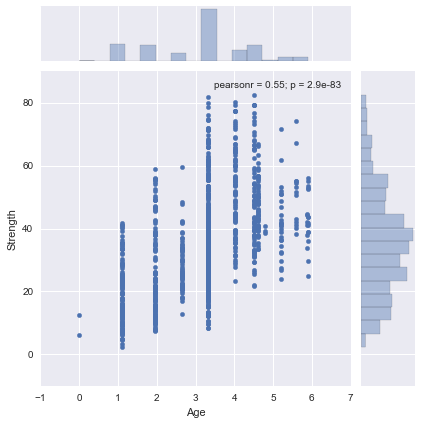
>> 그렇다. 잔차는 정규분포를 따른다고 할 수 있다. p-value가 0.07로 크다고 생각할 수 있지만, Jarque-Bera 검정은 귀무가설이 이고, 귀무가설의 조건이 정규분포의 조건이기 때문에 귀무가설을 채택할 때 잔차가 정규분포를 따른다고 할 수 있다.

(7) 잔차에 대해서 Breusch-Pagan test를 수행하시오. 설명 변수에 무관하게 등분산성을 보인다고 볼 수 있는지 기술하시오.

>> 그렇다. 잔차는 설명 변수에 무관하게 등분산성을 보인다고 볼 수 있다. 잔차 제곱에 대한 회귀식의 계수 가 모두 0일 때 기존 회귀식에서 오차는 설명변수에 영향을 받지 않는다고 해설 할 수 있다. Breusch-Pagan 검정은 위 잔차 제곱에 대한 회귀식에서의 F-검정을 통해 확인 할 수 있다. 한편, 위 과정은 을 통해 근사화해 카이제곱 검정을 통해 확인 할 수도 있다. 이때 귀무가설 을 채택해야 등분산성을 만족할 수 있다고 할 수 있다.

2. ‘Age’ 대신에 ‘Age’에 log를 취한 값을 변수로 사용하여 새로운 모형을 학습하려고 한다.

(1) ‘Age’와 ‘Strength’ 사이의 joint plot과 log(‘Age’)와 ‘Strength’ 사이의 joint plot을 서로 비교하고 log를 취했을 때 달라지는 점이 무엇인지 기술하시오.

>> ‘Age’ 변수의 histogram 분포를 보았을 때 대부분의 값이 0과 100안팎에 몰려 있고, 간격이 크게 벌어져 있는 것을 확인 할 수 있다. log(‘Age’) 변수는 그에 비해 전체 분포구간에서 중앙에 최대 likelihood가 존재하는 변화가 생긴 것을 알 수 있다. 그에 따라 대부분의 plot 좌표가 우상향으로 이동한 점이 달라졌다고 할 수 있다. 피어슨 상관계수 값도 0.55로 양의 상관관계가 있다고 할 수 있는 수치로 변화하였다.

(2) 기존 변수 대신에 바뀐 변수를 포함하여 선형 회귀 모형을 학습하고 아래 표를 완성하시오. 를 기준으로 계수가 0이라는 가설을 기각할 수 없는 변수는 무엇인가?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 변수명 | 추정 계수() |  |  |  |
| Intercept | -77.5333 | 1.8330e+01 | -4.2296 | 1.2751e-05 |
| Cement | 0.1348 | 5.8493e-03 | 23.0551 | 0.0 |
| Blast\_Furance\_Slag | 0.1159 | 6.9649e-03 | 16.6469 | 0.0 |
| Fly\_Ash | 0.0943 | 8.6247e-03 | 10.9442 | 0.0 |
| Water | -0.1288 | 2.7564e-02 | -4.6733 | 1.6799e-06 |
| Superplasticizer | 0.1188 | 6.4366e-02 | 1.8460 | 0.0325 |
| Coarse\_Aggregate | 0.0296 | 6.4620e-03 | 4.5930 | 2.4564e-06 |
| Fine\_Aggregate | 0.0359 | 7.3658e-03 | 4.8858 | 5.9747e-07 |
| log(Age) | 8.7465 | 1.9176e-01 | 45.6103 | 0.0 |

>> “Superplasticizer” 변수를 제외하고, 모든 변수의 계수는 일 때 0이라는 가설을 기각할 수 있다고 볼 수 있다. 각각의 t-통계량을 통해 p-value는 거의 0에 수렴하기 때문이다.

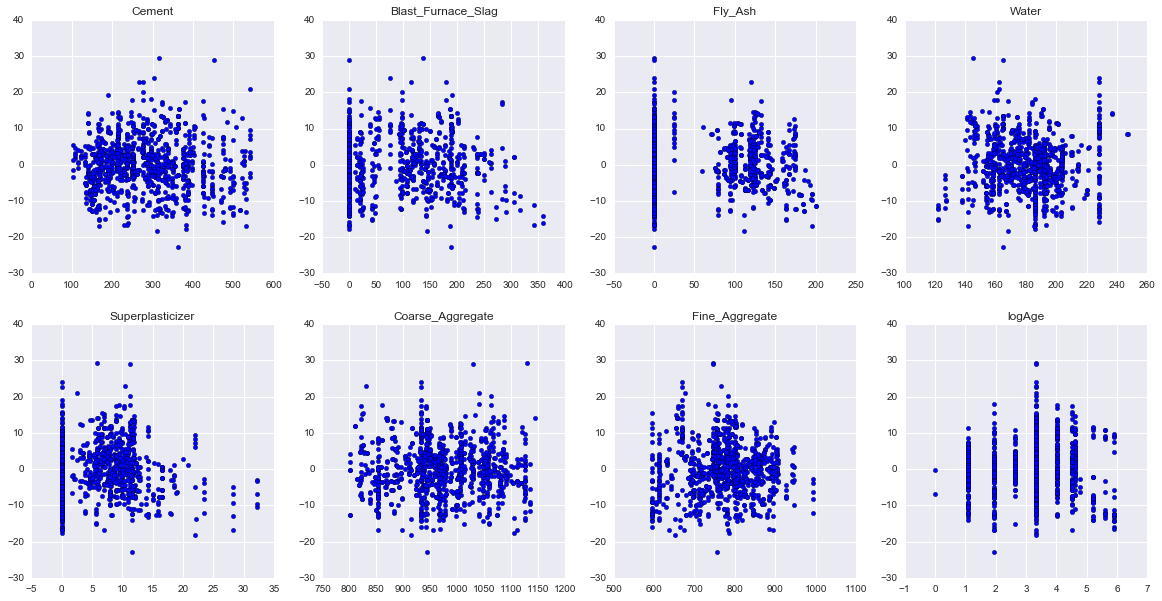
(3) 추정된 모형에 대한 ANOVA 표를 완성하시오. 전체적으로 이 모형을 강도를 예측하는데 쓸모가 있다고 할 수 있는가?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | SS | Degree of freedom | MS | F |  |
| Model | 235055.3619 | 8 | 29381.9202 | 575.5763 | 1.1102e-16 |
| Residual | 52119.8251 | 1021 | 51.0478 |  |  |
| Total | 287175.1871 | 1029 |  |  |  |

(4) 와 값을 구하시오. 기존 것과 비교했을 때 어느 쪽이 다 낫다고 볼 수 있는가?

>> 새 모델 = 0.8185 = 0.8170   
한편 기존 모델 = 0.6155 = 0.6125 새 모델의 설명력이 더 낫다고 볼 수 있다.

(5) 모든 설명 변수에 대해 개별 설명 변수를 x축으로 하고 y축을 잔차(residual)로 하는 산포도(scatter plot)을 그리시오. 산포도를 보고 변화한 사항에 대해 기술하시오.



>> 결과를 비교해 보면 “Age” 변수의 잔차 산포도의 변화가 있었다. 2차 곡선 모양 산포도가 0을 기준으로 직선으로 변화하는 것처럼 보인다. 하지만 등분산성인지에 대해서는 잘 모르겠다.

(6) 잔차에 대해서 Jarque-Bera test를 수행하시오. 잔차는 정규분포를 따른다고 볼 수 있는지 기술하시오.

>> 그렇지 않다. 잔차는 정규분포를 따른다고 할 수 없다. Jarque-Bera 검정은 귀무가설이 이고, 귀무가설의 조건이 정규분포의 조건이기 때문에 귀무가설을 채택할 때 잔차가 정규분포를 따른다고 할 수 있다. 그렇기 때문에 대립가설을 채택하기 때문에 잔차는 정규분포를 따르지 않는다고 볼 수 있다.

(7) 잔차에 대해서 Breusch-Pagan test를 수행하시오. 설명 변수에 무관하게 등분산성을 보인다고 볼 수 있는지 기술하시오.

>> 그렇지 않다. 을 통해 근사화해 카이제곱 검정을 통해 확인 할 수도 있다. 이때 귀무가설 을 채택해야 등분산성을 만족할 수 있다고 할 수 있는데, 대립가설을 채택하기 떄문에 등분산성을 보인다고 할 수 없다.