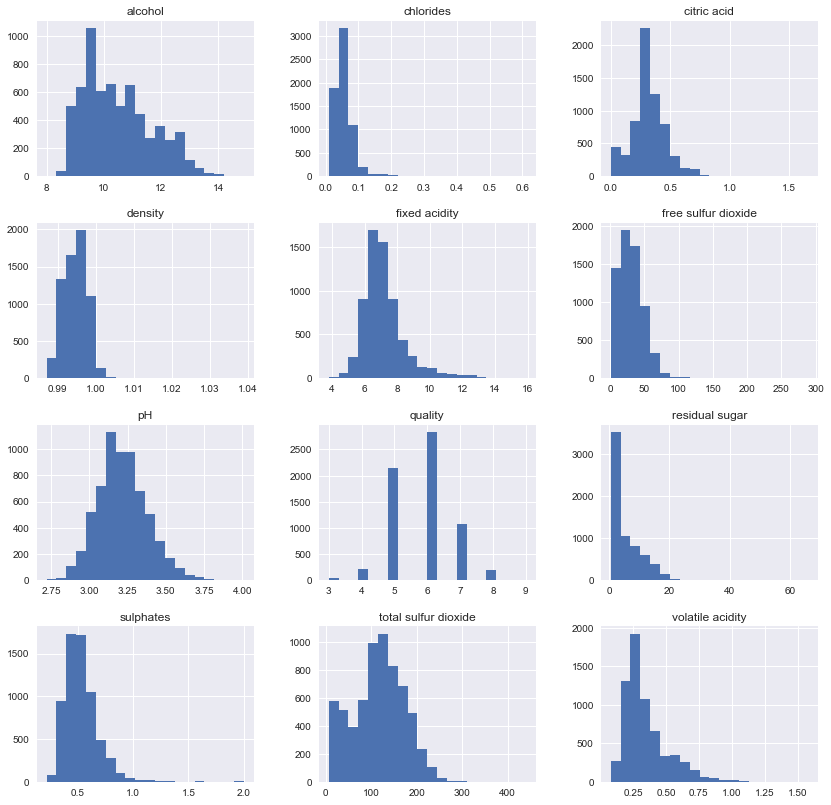
**HW – DM Process** 둘리조

<<,

**1. Data**

UCI 머신러닝 리포지토리의 와인 품질 데이터를 사용하였다. 해당 데이터는 Cortez et al. (2009)에 의하여 고안된 데이터로 포르투갈 와인 전체 생산량의 15%를 차지하고, 북부 및 북서부 미뉴 강 주변에서 주로 생산되는 ‘비뉴 베르드’ 와인 관련 데이터이다. 비뉴 베르드는 알코올 도수는 중간 정도이며 숙성되지 않아 ‘젊은 와인’이라고 불리기도 하며 여름철의 신선함으로 인해 많은 사랑을 받는다. Cortez et al. (2009)은 보다 와인 품질을 쉽게 예측하기 위하여 비뉴 베르드와 관련된 구연산 함유량, pH값 등과 같은 물리화학적 정보를 2004년 5월부터 2007년 2월까지 수집하였다. 또한, 와인의 품질은 세 명의 와인 감별사가 0(매우나쁨)부터 10(매우좋음)으로 평가한 중간 값을 사용하였다. 단, 화이트와인과 레드와인의 맛 평가 기준은 상이하기 때문에 분리하여 평가하였다.

관측치는 레드와인 1,599개, 화이트와인 4,898개로 구성되어 있고, 변수는 독립변수 11가지와 종속변수 한 가지, 총 12가지로 이루어져 있다. 독립변수는 모두 수치형 데이터로 구성되어 있으며, 변수의 종류로는 고정산(fixed acidity), 휘발산(volatile acidity), 구연산(citric acid), 잔당(residual sugar), 염화물(chlorides), FSO2(free sulfur dioxide), 총SO2(total sulfur dioxide), 밀도(density), pH, 황산염(sulphates), 알코올 도수(alcohol), 품질(quality)이 있다. 총SO2를 제외하고는 모두 right-skewed된 형태이고 특히 염화물의 왜도가 가장 컸다. 구체적인 측정 단위 및 변수의 통계량은 아래와 같다.



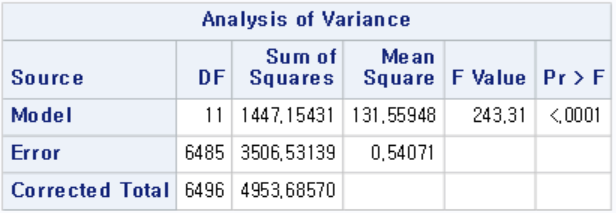
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **변수 (단위)** | **min** | **mean** | **max** | **sd** | **skewness** |
| 고정산 (g(주석산)/dm3) | 3.8 | 7.2153 | 15.9 | 1.2964 | 1.7233 |
| 휘발산 (g(초산)/dm3) | 0.08 | 0.3397 | 1.58 | 0.1646 | 1.4951 |
| 구연산 (g/dm3) | 0 | 0.3186 | 1.66 | 0.1453 | 0.4717 |
| 잔당 (g/dm3) | 0.6 | 5.4432 | 65.8 | 4.7578 | 1.4354 |
| 염화물 (g(염화나트륨)/dm3) | 0.009 | 0.0560 | 0.611 | 0.0350 | 5.3998 |
| FSO2 (g/dm3) | 1 | 30.525 | 289 | 17.749 | 1.2201 |
| 총SO2 (g/dm3) | 6 | 115.74 | 440 | 56.522 | -0.0012 |
| 밀도 (g/dm3) | 0.9871 | 0.9947 | 1.039 | 0.0003 | 0.5036 |
| pH | 2.72 | 3.2185 | 4.01 | 0.1608 | 0.3868 |
| 황산염 (g(황산칼륨)/dm3) | 0.22 | 0.5313 | 2 | 0.1488 | 1.7973 |
| 알코올 (%) | 8 | 10.492 | 14.9 | 1.1927 | 0.5657 |

와인은 일반적으로 적절한 신맛과 단맛의 조화가 가장 중요한 품질 요소 중 하나라고 알려져 있다. 고정산은 가열해도 증발되지 않는 산이고, 휘발산은 가열하면 증발하는 산인데 함유량이 낮아야 좋은 와인이라고 알려져 있다. 구연산은 상쾌한 신맛이 나는 산이다. 잔당은 자연적 발효에 의하여 생성된 당이다. 염화물은 와인의 짠맛을 결정하고 이산화황은 와인의 일관성과 품질을 위해 필수적이라는 의견은 있지만 유해한 성분이자 와인의 섬세한 풍미를 죽일 수도 있다는 의견이 있는 물질이다. pH는 산도와는 약간 차이가 나는 개념으로 용액에서 해리된 수소이온의 양만 측정한 값이다. 황산염은 박테리아를 예방하기 위하여 필요한 물질이며 마지막으로 알코올은 와인의 강도를 결정하는 물질이고 강한 단맛을 가지고 있고 복합적인 맛 형성 및 향의 운반체 역할을 하기도 한다.

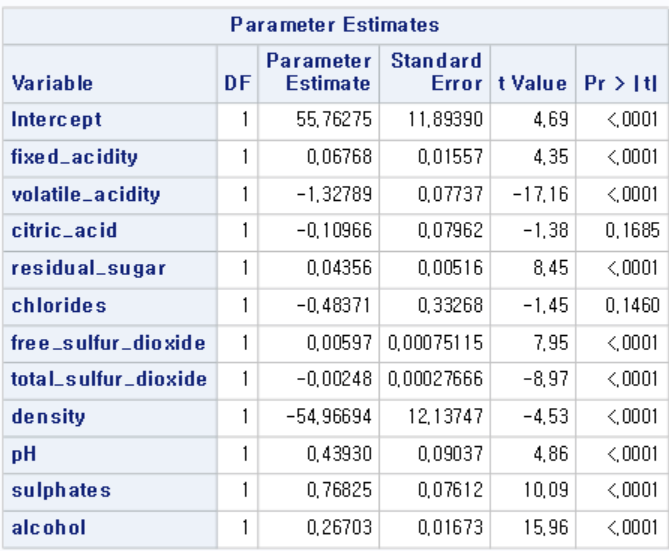
**2. Variable Selection**

11개의 설명변수로 wine의 quality를 설명하는 모델을 적합해 보았다. 총 6가지의 변수선택 기법인 forward selection, backward elimination, stepwise selection, ridge regression, lasso regression, elastic net regression을 이용해 분석해 보았다.

먼저, 가장 기본적인 OLS를 이용한 회귀 분석 결과 값이다.







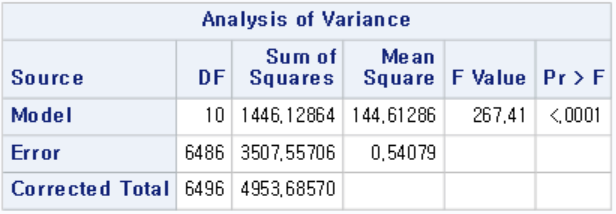
추정된 회귀식은 다음과 같다

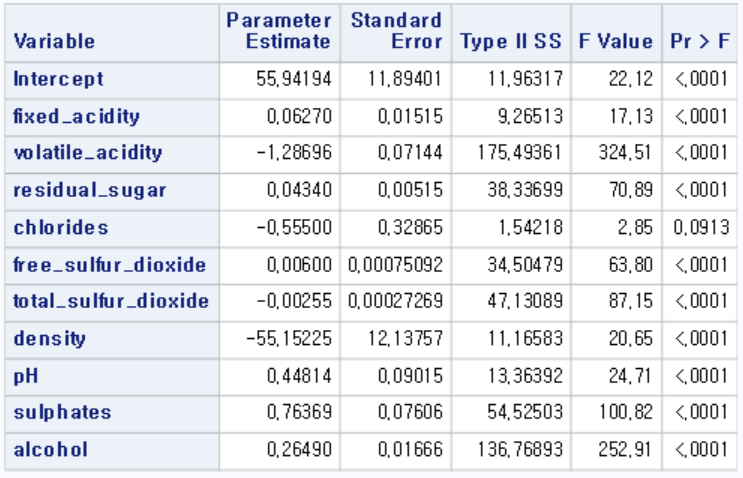
이 모형의 결정계수는 0.2921이며, 분산분석의 F통계량의 값은 243.31이었고 p값은 0.0001보다 작은 것으로 보아 유의한 모델임을 알 수 있었다. 개별 변수 별로 보면 citric\_acid와 chlorides의 p값이 각각 0.1685,0.1460으로 유의수준 0.1하에서 유의하지 않은 변수라는 것을 알 수 있었다. Fixed\_acidity, residual\_sugar, free\_sulfur\_dioxide,pH,sulphates,alcohol은 wine의 quality에 좋은 영향을 주는 변수이고 volatile\_acidity, citric\_acid, chlorides, total\_sulfur\_dioxide, density는 quality에 좋지 않은 영향을 주는 변수임을 알 수 있었다.

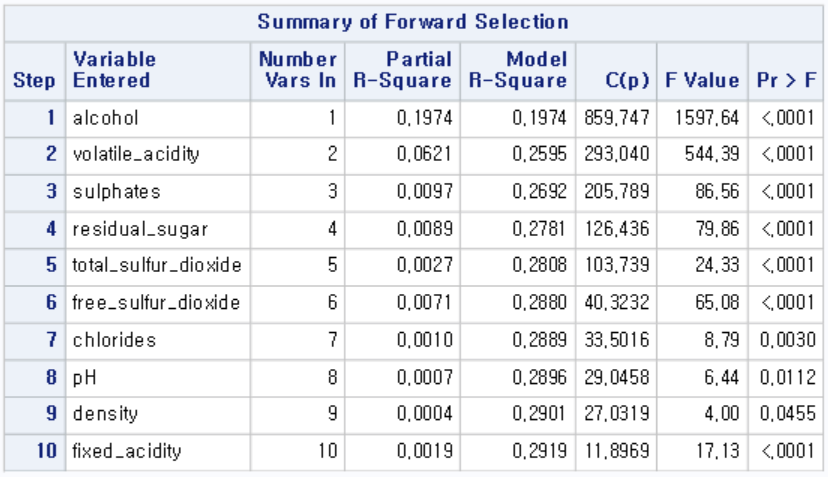
**1. forward selection**

유의수준을 0.1로 설정한 후, forward selection을 진행하였다. 분석 결과, 첫 번째 스텝에서 ‘alcohol’이 선택되었고 R Square값은 0.1974로 나타나 와인의 품질을 결정하는데 알코올이 가장 중요한 요소라는 사실을 추정할 수 있었다. 그 이후로 유의수준 하에서 아홉 개의 변수들이 추가될 때 증가하는 R Square값이 유의하여 추가되었고 ‘citric acid’이 포함되었을 때 증가하는 R Square값은 너무 작아 유의한 값이 아니기 때문에 변수 선택에서 제외되었다. 최종적인 R square의 값은 0.2919로 나타났고 추정된 회귀식은 다음과 같이 나타났다.

이 모형의 결정계수는 0.2919였고, 분산분석의 F통계량의 값은 267.41이고 p값은 0.0001보다 작은 것으로 보아 유의한 모델임을 알 수 있었다. 앞서 말했듯, 일반적으로 와인은 신맛과 단맛의 조화가 가장 중요한 품질 요소의 하나라고 평가 하는데 분석하는 변수 중 단맛의 요인은 alcohol, residual\_sugar로 신맛의 요인은 citric\_acid, volatile\_acidity, fixed\_acidity이다. 실제 결과도 partial R square 값을 고려 할 때 alcohol과 volatile\_acidity이 각각 0.1974, 0.0621으로 상대적으로 큰 값을 나타내어 가장 먼저 선택되었음을 알 수 있다. 2개의 변수만으로 0.2595의 R Square 값을 가짐을 알 수 있고, 이는 최종적인 R Square값이 0.2919임을 고려 할 때 89%를 차지할 정도로 큰 설명력을 가지고 있음을 나타낸다. 따라서 이 두 변수가 와인의 품질에 큰 영향을 끼친다는 것을 뒷받침 할 수 있다. citric\_acid은 신맛에 영향을 주는 요소라고 알려져 있기는 하지만 와인의 품질 결정에 큰 영향을 주는 요소가 아닌 것으로 보인다. sulphates는 0.76의 양의 추정 계수를 가지는데 이는 sulphates가 건강에 유해하다고 판단되어 법적으로 양에 제한이 있기는 하지만 많이 함량 될 수록 와인의 품질에도 좋은 영향을 끼치는 물질임을 알 수 있었다. Chlorides, volatile\_acidity가 음수를 나타내는데 실제로 volatile\_acidity의 경우 역한 신맛을 유발하여 작을수록 와인의 품질이 좋다고 알려져 있다. chlorides는 와인의 일관성과 품질을 위해 필수적이라는 의견과 유해한 성분이자 와인의 섬세한 풍미를 죽일 수도 있다는 의견이 둘 다 있는데 -0.555의 음의 값을 가지는 것으로 보아 와인의 품질을 결정하는데 적을수록 좋다는 것을 알 수 있었다. free\_sulfur\_dioxide와 total\_sulfur\_dioxide의 경우 각각의 추정 계수 값과(0.006, -0.00255) 데이터의 평균(30.525,115.74) 보면 이산화황이 와인의 일관성과 품질을 위해 필수적이지만, 와인의 품질을 떨어뜨릴 수 도 있음을 알 수 있었다. 마지막으로 상태수가 448.16이 나온 것으로 보아 다중공선성 문제가 어느 정도 있는 것으로 보인다. 분석 결과 표는 아래와 같다.





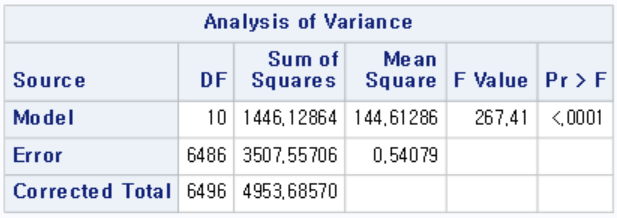


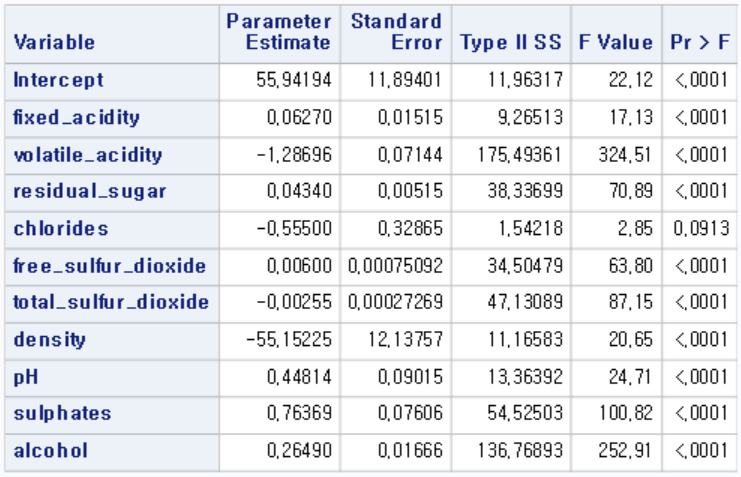
**2. backward elimination**

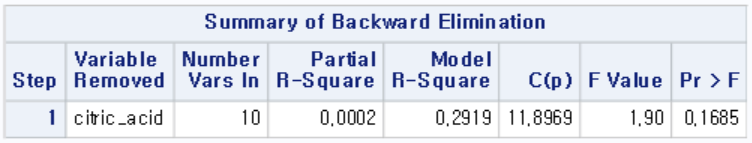
유의수준을 0.1로 설정한 후, backward elimination을 진행하였다. 분석 결과, 첫 번째 스텝에서 ‘citric\_acid’가 제거되었다. Citric\_acid가 제거 되었을 때, R Square값이 0.0002 줄어들었지만 너무 작은 값으로 유의하지 않아 포함되지 않았다. 첫 번째 스텝에서 citric\_acid를 고려한 이유는, 다른 모든 변수들을 이용해서 모델을 적합한 다음 citric\_acid를 추가했을 때 증가하는 SSR값이 다른 변수들과 비교해서 가장 작았기 때문이다. 그 이후로 유의수준 하에서 제거되는 변수는 없었다. 최종적인 R square의 값은 0.2919로 나타났고 추정된 회귀식은 다음과 같이 나타났다.

=55.94194+0.0670fixed\_acidity-1.28696volatile\_acidity+0.0434residual\_sugar-0.555chlorides+0.006free\_sulfur\_dioxide-0.00255total\_sulfur\_dioxide-55.15225density+0.44814pH+0.76369sulphates+0.2649alcohol

이 모형의 결정계수는 0.2919였고, 분산분석의 F통계량의 값은 267.41이고 p값은 0.0001보다 작은 것으로 보아 유의한 모델임을 알 수 있었다. 회귀식을 보면 위의 forward selection기법과 일치함을 알 수 있다. 따라서 변수 해석 결과가 위의 forward selection기법에서의 해석과 동일하고 결과 표는 아래와 같다.



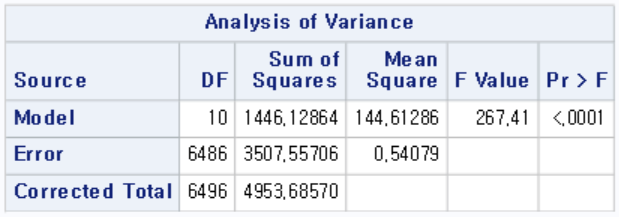


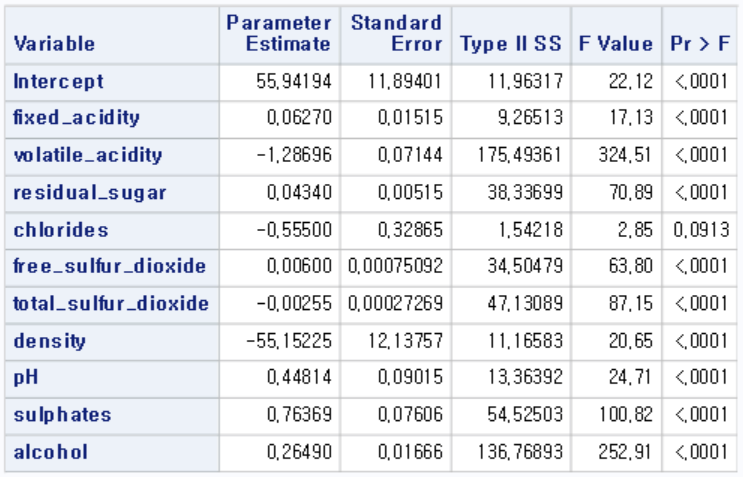


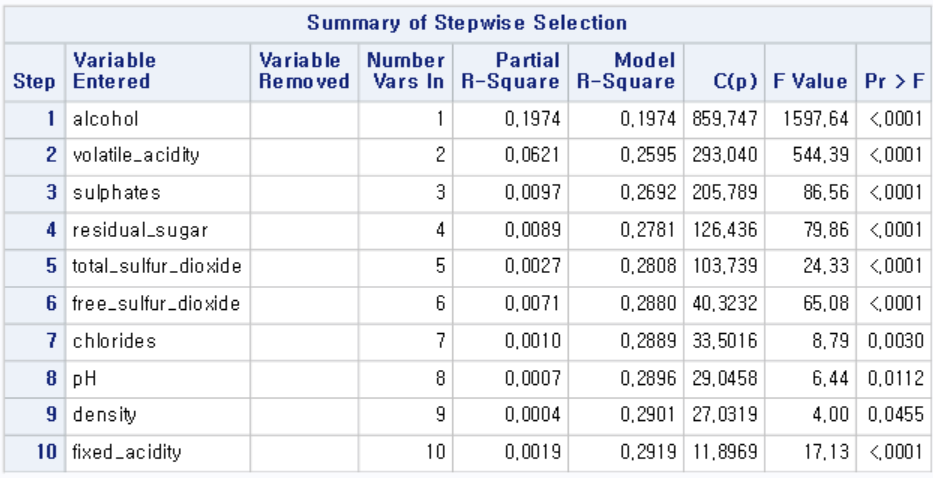
**3. stepwise selection**

유의수준을 0.1로 설정하여 stepwise selection을 이용해 분석한 결과 첫 번째 스텝에서 ‘alcohol’이 선택되었고 R Square값은 0.1974으로 나타났다. 그 이후로 유의수준 하에서 아홉 개의 변수들이 추가될 때 증가하는 R Square값이 유의하여 추가되었고 ‘citric acid’이 포함되었을 때 증가하는 R Square값은 너무 작아 유의한 값이 아니기 때문에 변수 선택에서 제외되었다. 또한 변수 제거 과정에서 제거되는 변수는 없었다. 모형에 투입되는 변수들의 순서는 alcohol, volatile\_acidity, sulphates, residual\_sugar, total\_sulfur\_dioxide, free\_sulfur\_dioxide, chlorides, pH, density, fixed\_acidity 순이였다. 최종적으로는 R Square값이 0.2919가 나왔고 추정된 회귀식은 다음과 같다.

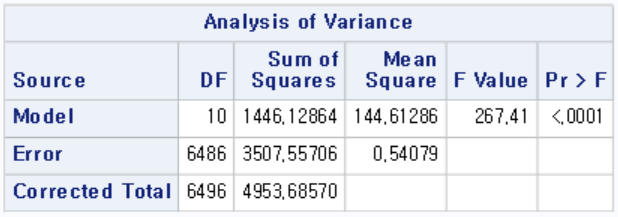
이 모형의 결정계수는 0.2919였고, 분산분석의 F통계량의 값은 267.41이고 p값은 0.0001보다 작은 것으로 보아 유의한 모델임을 알 수 있었다. 회귀식을 보면 위의 forward selection, backward elinimation기법과 일치함을 알 수 있다. 따라서 변수 해석 결과가 위의 forward selection, backward elimination 기법에서의 해석과 동일하고 결과 표는 아래와 같다.







\*\*\*\*Adjusted R Square 값을 비교하기 위해 위의 세 방법에서 제거된 변수 citric\_acid를 제외하고 OLS회귀 분석을 해보았다.





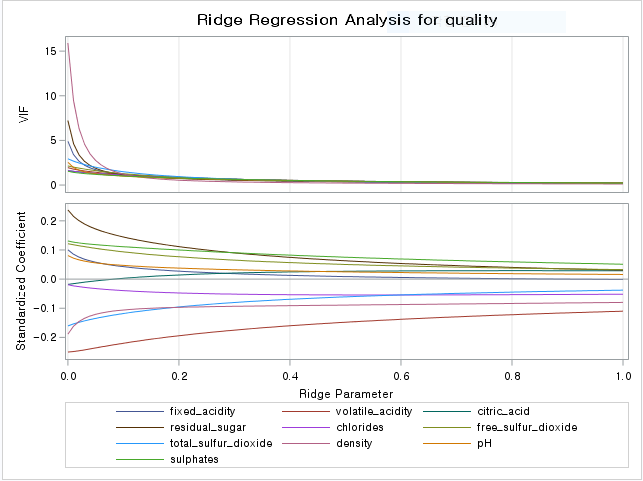
Adjusted R Square 값이 0.2908로 모든 변수를 이용해서 구한 값인 0.2909보다 낮은 값을 보였다. R Square, Adjusted R Square를 지표로 이용한다면 주어진 데이터 셋에서 유의수준 0.1의 forward selection, backward elimination, stepwise selection은 좋은 방법이 아니라고 판단된다.

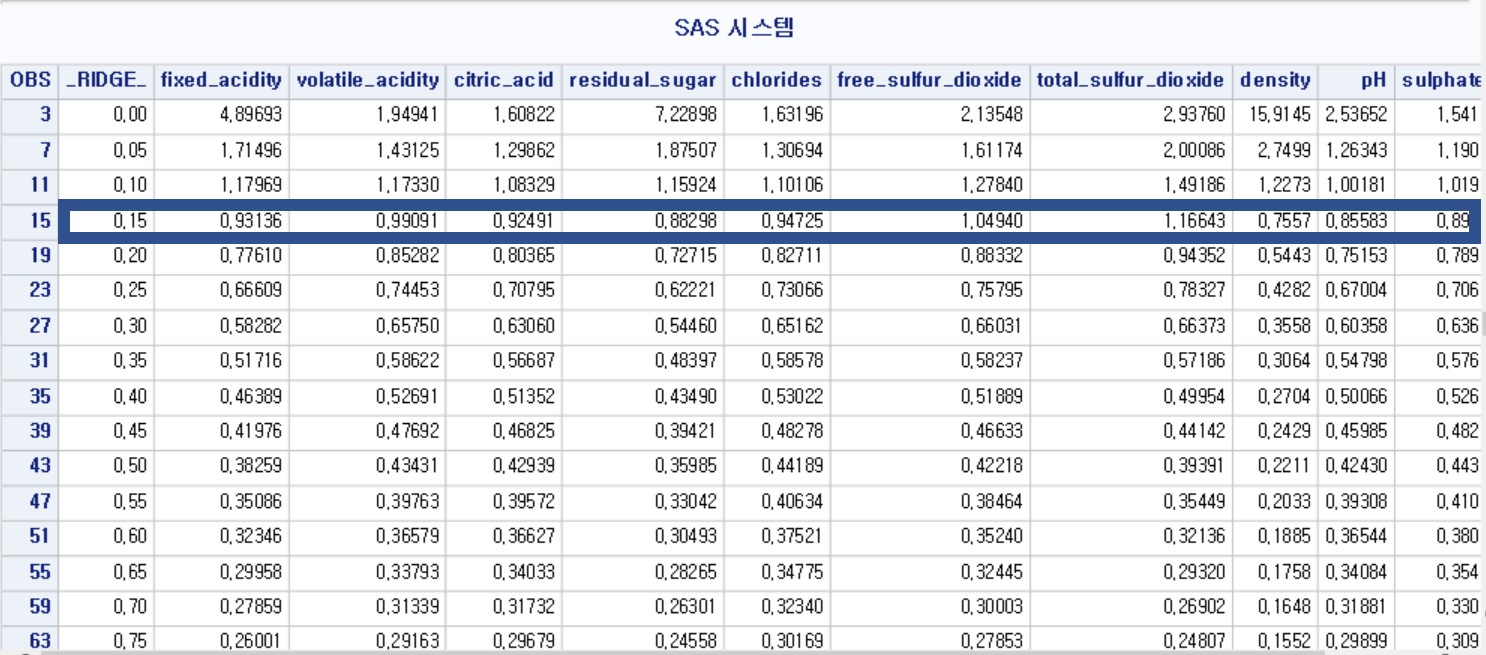
**4. ridge regression**

Ridge regression은 계수의 제곱에 대한 패널티가 존재하며, ridge regression을 통해 다중선형회귀 모형의 선형 계수 값의 크기를 감소시켜 다중공선성에 의한 예측 오차를 줄일 수 있다. Ridge 식은 다음과 같다.

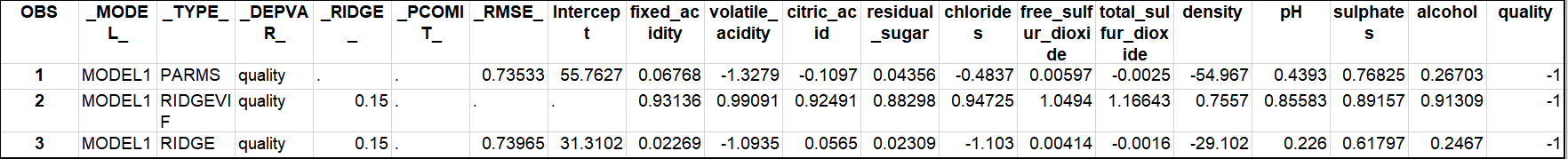


변수들의 다중공선성을 줄이기 위해서 ridge parameter값을 변경해 보면서(0에서 10까지 0.05단위로 변경해 보았다) VIF값이 1근처가 되는 값을 찾아 보았다.(VIF 값이 1 근처일때 변수들의 다중공선성이 가장 줄어드는 것으로 알려져 있다.)

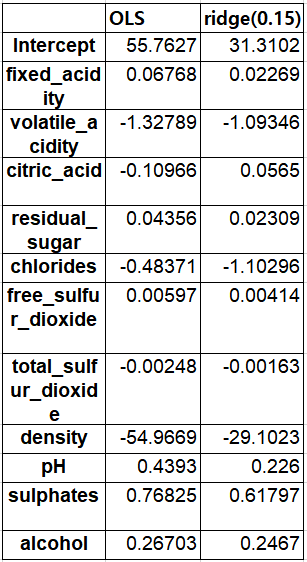




위의 표와 그래프에서 알 수 있듯이 ridge parameter가 0.15일 때 VIF값들이 1에 가장 가까웠다. 다음은 ridge parameter=0.15로 회귀 분석했을 때의 변수의 계수 값들이다.



추정된 회귀식은 다음과 같다.



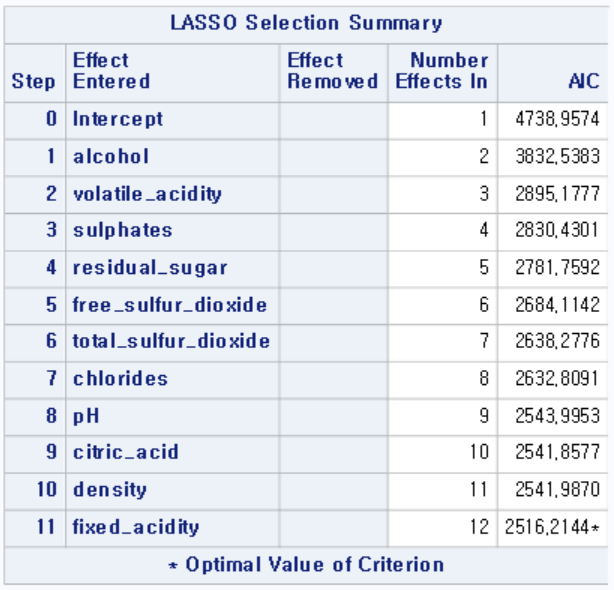
Ridge를 이용하여 구한 계수 들은 OLS를 이용하여 구한 계수들보다 더 작은 값은 가진 다는 것을 알 수 있었다. 하지만, ridge 회귀의 경우 선형 계수가 0이 될 수 없으므로 계수의 절댓값이 더 작아지기는 했어도 변수 제거 효과가 나타나지는 않았다. 설명변수들의 다중공선성이 높아 회귀계수의 과잉추정이 우려 될 때 ridge regression을 통해 계수의 크기를 축소하여 문제를 해결 할 수 있다.

**5. lasso model**

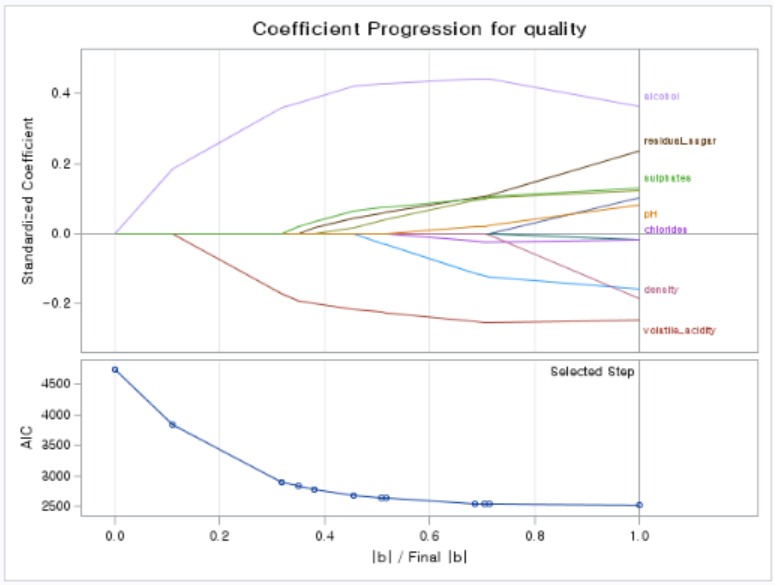
Lasso regression은 계수의 절댓값에 대한 패널티가 존재하며, ridge regression의 선형 계수가 0이 될 수 없다는 단점을 보완하여 변수선택이 가능하도록 제안된 축소추정 방법이다. Lasso 회귀는 설명변수를 하나씩 투입해 보면서 AIC가 최저인 지점을 찾고, 그 때의 설명변수들로만 모형을 적합한다. 변수들을 추가 할수록 설명력은 높아지지만, AIC식에는 변수 수에 대한 penalty항도 존재한다. Lasso식은 다음과 같다.



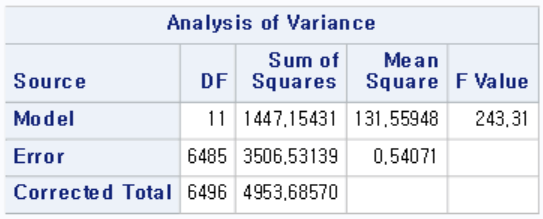
다음은 lasso회귀를 한 결과이다.

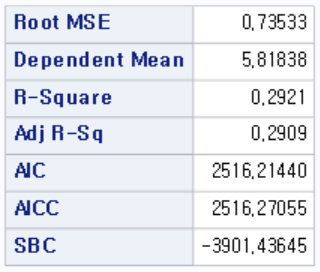


11개의 설명변수가 모두 이용되었을 때 AIC가 최저값을 가진다는 것을 알 수 있었고, 모든 설명변수가 모델 적합에 이용된다는 걸 알 수 있었다.



x axis = normalized tuning point(= the normalized lambda, 변수를 추가해 나갈수록 1.0에 가까워진다)







추정된 회귀식은 다음과 같다

=55.94194+0.0670fixed\_acidity-1.28696volatile\_acidity+0.0434residual\_sugar-0.555chlorides+0.006free\_sulfur\_dioxide-0.00255total\_sulfur\_dioxide-55.15225density+0.44814pH+0.76369sulphates+0.2649alcohol

OLS로 회귀를 한 것과 같은 결과가 나왔다. 모든 설명변수가 이용되었을 때 AIC값이 가장 작았기 때문에 변수선택의 효과가 없었고, 결국 OLS와 같은 결과가 나온 것이다. 이를 통해서 모든 설명 변수에 대하여 AIC값을 측정함에 있어서 변수 수에 의한 penalty효과 보다 설명력 증가의 효과가 더 컸다는 것을 알 수 있었다.

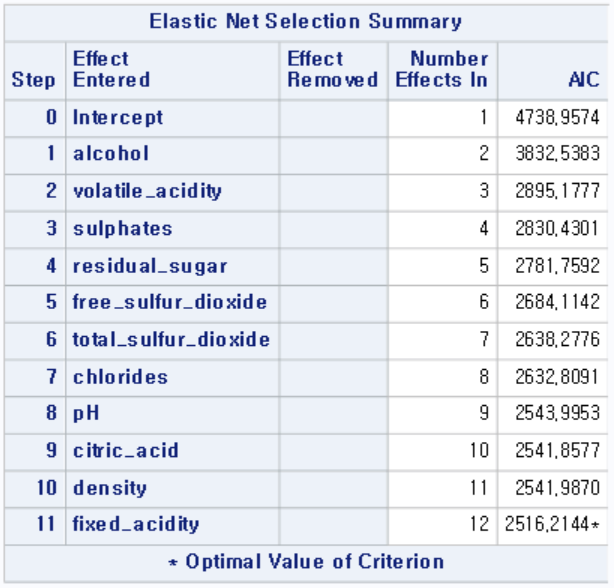
**6. elastic net model**

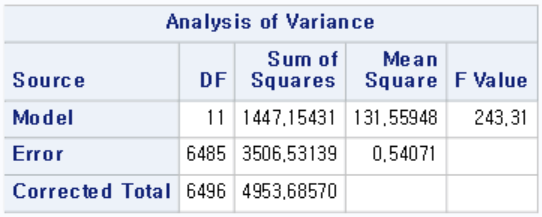
elastic net은 lasso와 ridge의 절충안으로 나온 모델로 상관관계가 있는 변수들 중에서 하나의 변수만을 흔히 선택하는 lasso의 단점을 보완하는 형태이다. Sas에서는 이를 L1과 L2를 이용하여 조정하는데 l1은 lasso penalty를 l2는 ridge penalty를 의미한다. 따라서 이를 조정하여 elastic net식 만으로 ridge, lasso 모델 적용이 모두 가능하다. elastic net의 식은 다음과 같다.

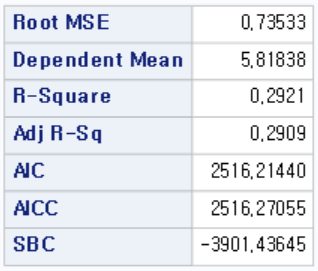


위 식에서 λ2 = l2의 경우 l2low와 l2high를 이용하여 0~1사이의 값이 선택되도록 하였다. 이렇게 하면 sas에서 GLMSELECT를 이용하여 자동으로 l2의 값을 조절하며 기준으로 선택한 AIC에 가장 적합한 l1 값을 선택해 준다. (λ1,λ2는 각각 lasso, ridge관련 패널티 계수이다)GLMSELECT에 대한 자세한 정보는 따로 첨부 하였다.

분석 결과는 아래와 같다.

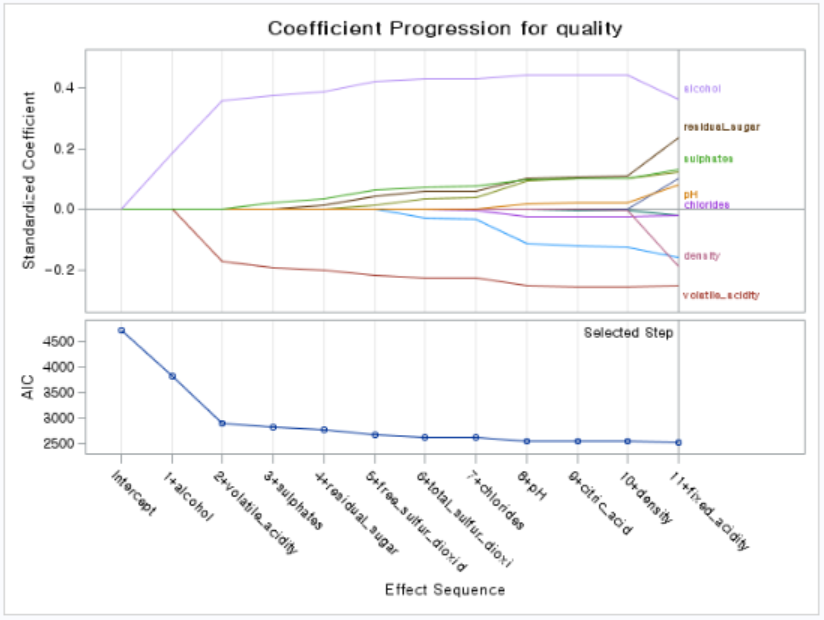








추정된 회귀식은 다음과 같다.

=55.94194+0.0670fixed\_acidity-1.28696volatile\_acidity+0.0434residual\_sugar-0.555chlorides+0.006free\_sulfur\_dioxide-0.00255total\_sulfur\_dioxide-55.15225density+0.44814pH+0.76369sulphates+0.2649alcohol

그래프를 해석해 보면 AIC가 가장 낮은 점인 ‘fixed\_acidity’가 최저점으로 선택되었고 결과적으로 모든 변수가 선택되었다. 따라서 모든 변수가 의미 있는 변수이고 가장 설명력이 있는 변수로는 ‘alcohol’이 가장 낮은 변수로는 ‘fixed\_acidity’로 나타났다. 결과적으로 어떤 변수도 제거되지 않았기 때문에 위의 lasso regression과 같은 결과가 나왔다. Ridge 관련 패널티 효과도 없었고 lasso 관련 패널티 효과도 없었기 때문에 위의 lasso, OLS와 같은 결과가 나왔다. Elastic net은 lasso를 보완하여 긴밀한 설명변수의 그룹을 유지하기 위하여 고안되었는데, 애초에 lasso로 인해 제거되는 설명변수가 없었기 때문에, lasso와 같은 결과를 보인다.

**비교**

Forward selection, backward elimination, stepwise selection을 했을 때 모두 변수 citric\_acid가 제거 되었고 같은 회귀 결과가 나왔다. Ridge를 이용하여 회귀 분석을 한 결과 제거되는 변수들은 없었지만, 계수들의 절댓값의 크기가 줄어들었다. Lasso, elastic net을 사용해서 회귀 분석을 한 결과 제거되는 변수가 없었고 OLS와 같은 결과를 보였다. 변수 제거를 통해서 다중공선성을 줄이는 효과보다 있는 설명 변수를 모두 이용하여 설명력을 높이는 효과가 크기 때문인 것 같다. 변수수가 더 많고 변수 간의 다중공선성이 더 컸더라면 제거되는 변수가 생겼을 것이고 ridge, lasso, elastic net을 통해서 더 좋은 설명력을 가진 모델을 적합할 수 있었을 것이라고 생각한다.’

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **OLS** | **Forward** | **Backward** | **Stepwise** | **Lasso** | **Elastic** |
| **R Square** | 0.2921 | 0.2919 | 0.2919 | 0.2919 | 0.2921 | 0.2921 |
| **F value** | 243.31 | 267.41 | 247.41 | 267.41 | 243.31 | 243.31 |
| **DF** | 11 | 10 | 10 | 10 | 11 | 11 |
| **Mean Square** | 131.56 | 144.61 | 144.61 | 144.61 | 131.56 | 131.56 |
| **AIC** |  |  |  |  | 2516.21 | 2516.21 |

**3. Chi-square Test**

다음으로는, 여러 변수간의 관계를 검정하기 위하여 카이스퀘어 독립성 검정을 진행하였다. 변수 종류에 따라 각각 다른 방식으로 이산화 과정을 거쳤다. 우선, 독립성 검정을 할 변수로 짠 맛을 결정하는 염화물, 신 맛을 어느정도 대표하는 pH, 발효를 위해 필요한 황산염, 단 맛과 향에 영향을 주는 알코올, 와인 색을 골랐고, 그 변수들의 몇 가지 관계를 관찰하고자 하였다. 색이나 품질을 제외한 변수들이 모두 연속형 수치형 데이터이기 때문에, 다양한 독립성 검정을 수행하기 위하여 변수들의 이산화 과정을 거쳤다.

염화물의 경우에는 중앙값이 0.047이어서 비슷한 비율로 나누어지게끔 0.05를 기준으로 두 범주로 나누었다. pH의 경우에는 도메인 지식을 활용하여 이산화를 진행하였다. 3 이하의 와인은 품질이 좋지 않다고 알려져 있어 한 범주로 묶어주었고 평균적으로 와인의 pH는 3.3이라고 알려져 있어 0.3 단위로 묶어주었다. 마지막으로, 3.6 이상인 데이터 수가 적어 3.6 이상을 한 범주로 묶었다. 황산염의 경우에는 염화물과 마찬가지로 중앙값이 0.51이어서 비슷한 비율로 나누어지게끔 0.5를 기준으로 두 범주로 나누었고 도수는 1% 단위로 범주를 나누었다. 따라서, 독립성 검정을 진행하기 위하여 변형을 완료한 상태는 아래와 같다.

|  |  |
| --- | --- |
| **변수** | **값** |
| 염화물 | ~0.05 / 0.05~ |
| pH | ~3.0 / 3.0~3.3 / 3.3~3.6 / 3.6 |
| 황산염 | ~0.5 / 0.5~ |
| 알코올 | 8 / 9 / 10 / 11 / 12 / 13 / 14 |
| 색 | red / white |

위 다섯 가지 변수의 모든 경우에 대한 카이스퀘어 독립검정 테이블 통계량은 아래와 같다.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **염화물 X pH** | | | **염화물 X 황산염** | | |
| **통계량** | **자유도** | **값** | **prob** | **자유도** | **값** | **prob** |
| 카이제곱 | 3 | 141.3970 | <.0001 | 1 | 430.3187 | <.0001 |
| 우도비 카이제곱 | 3 | 141.1185 | <.0001 | 1 | 437.3384 | <.0001 |
| Mantel-Haenszel 카이제곱 | 1 | 8.6040 | 0.0034 | 1 | 430.2524 | <.0001 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **염화물 X 알코올** | | | **염화물 X 색** | | |
| **통계량** | **자유도** | **값** | **prob** | **자유도** | **값** | **prob** |
| 카이제곱 | 6 | 684.9069 | <.0001 | 1 | 2369.8346 | <.0001 |
| 우도비 카이제곱 | 6 | 721.7389 | <.0001 | 1 | 2744.4937 | <.0001 |
| Mantel-Haenszel 카이제곱 | 1 | 588.1262 | <.0001 |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **pH X 황산염** | | | **pH X 알코올** | | |
| **통계량** | **자유도** | **값** | **prob** | **자유도** | **값** | **prob** |
| 카이제곱 | 3 | 280.1860 | <.0001 | 1 | 283.9448 | <.0001 |
| 우도비 카이제곱 | 3 | 287.6125 | <.0001 | 1 | 212.9670 | <.0001 |
| Mantel-Haenszel 카이제곱 | 1 | 0.1784 | 0.6728 | 1 | 1.5663 | 0.2107 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **pH X 색** | | | **황산염 X 알코올** | | |
| **통계량** | **자유도** | **값** | **prob** | **자유도** | **값** | **prob** |
| 카이제곱 | 3 | 619.7982 | <.0001 | 6 | 158.1465 | <.0001 |
| 우도비 카이제곱 | 3 | 605.2027 | <.0001 | 6 | 160.8037 | <.0001 |
| Mantel-Haenszel 카이제곱 |  |  |  | 1 | 0.5180 | 0.4717 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **황산염 X 색** | | | **알코올 X 색** | | |
| **통계량** | **자유도** | **값** | **prob** | **자유도** | **값** | **prob** |
| 카이제곱 | 1 | 619.7982 | <.0001 | 6 | 166.2632 | <.0001 |
| 우도비 카이제곱 | 1 | 605.2027 | <.0001 | 6 | 211.1244 | <.0001 |
| Mantel-Haenszel 카이제곱 |  |  |  |  |  |  |

모든 경우에 대해 두 범주형 변수간의 관계가 없다, 즉 두 변수는 서로 독립이다는 귀무가설 하에서 계산된 검정통계량 값의 p-value가 0.0001 이하이다. 따라서 유의수준 0.05하에서 귀무가설을 모두 기각할 수 있고 두 범주형 변수는 종속이라는 대립가설을 채택할 수 있다. 다시 말하면, 염화물에 따라 pH, 황산염, 알코올, 색이 차이 날 수 있고 색과 알코올, 황산염과 pH 등 다른 변수들간의 관계도 종속이라고 할 수 있다.

특히, pH와 황산염, pH와 알코올, 황산염과 알코올을 제외한 경우에는 Mantel-Haenszel 카이제곱 통계량 값의 p-value가 유의수준보다 낮다는 것을 알 수 있다. 색을 제외한 변수들이 현재 모두 순서형이므로, 유의수준 0.05하에서 두 변수간 어떠한 연관도 없다는 귀무가설을 기각하고, pH와 황산염, pH와 알코올, 황산염과 알코올을 제외한 모든 변수간에는 선형관계가 존재한다는 대립가설을 채택할 수 있다. 즉, pH와 황산염, pH와 알코올, 황산염과 알코올은 서로 선형 관계에 있지는 않지만 종속 관계에 있다.

**REFERENCES**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009.

SAS elastic net

<http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_glmselect_details12.htm>

GLMSELECT

http://www2.sas.com/proceedings/sugi31/207-31.pdf