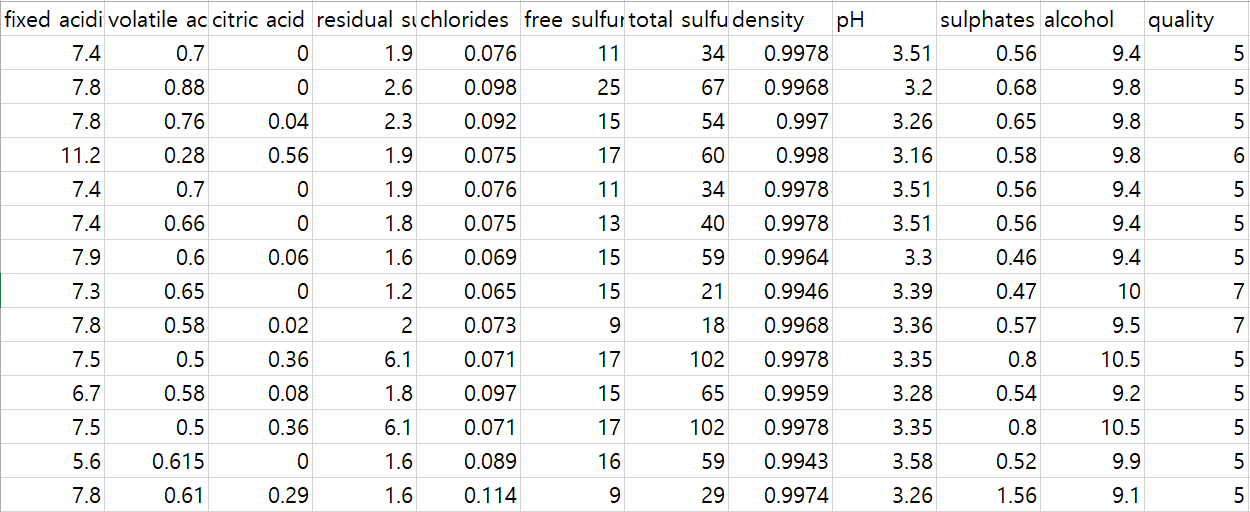
와인 데이터

1. 데이터 소개



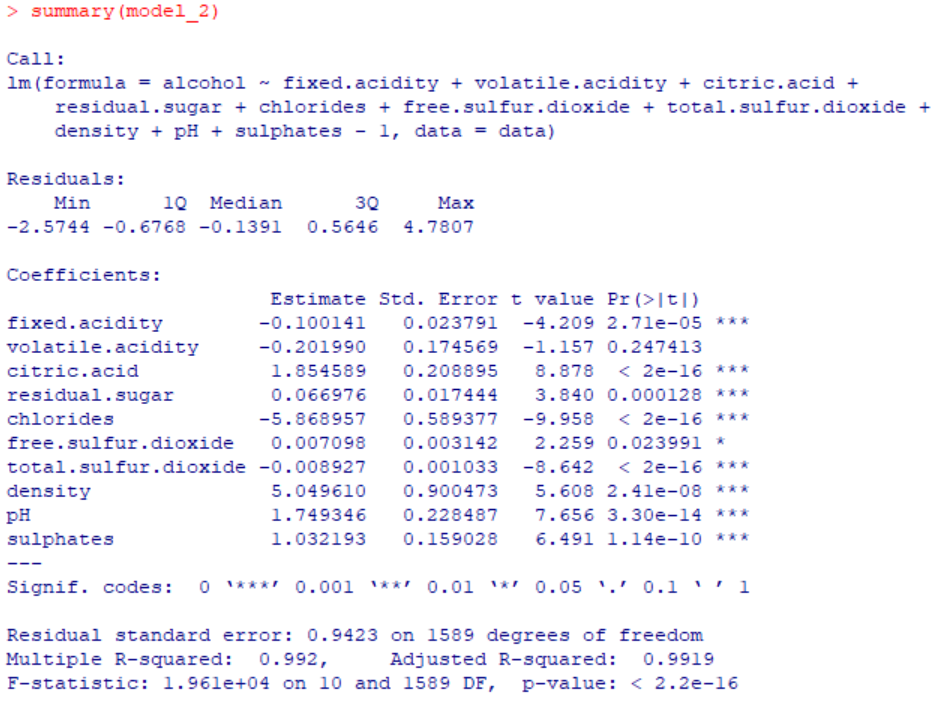
화이트 와인의 품질 분류에 관한 데이터로 다른 변수들이 품질에 대해 어떠한 영향을 미치는 지에 대한 데이터

2. 변수 소개

|  |  |
| --- | --- |
| **변수명** | **변수 설명** |
| fixed acidity | 비휘발성 산 |
| volatile acidity | 휘발성 산 |
| citric acid | 구연산(시트르산) |
| residual sugar | 잔당 |
| chlorides | 염화물 |
| free sulfur dioxide | 유리 이산화황 |
| total sulfur dioxide | 총 이산화황 |
| density | 밀도 |
| Ph | 산도(수소 이온 농도) |
| Sulphates | 황산염 |
| Alcohol | 알코올 |
| Quality | 와인 품질(나쁨 : 0 ~ 좋음 : 10) |

이 중 반응변수를 품질이 아닌 alcohol로 설정하여 다른 변수들이 알코올 농도에 미치는 영향을 알아보기 위해 회귀분석 실행

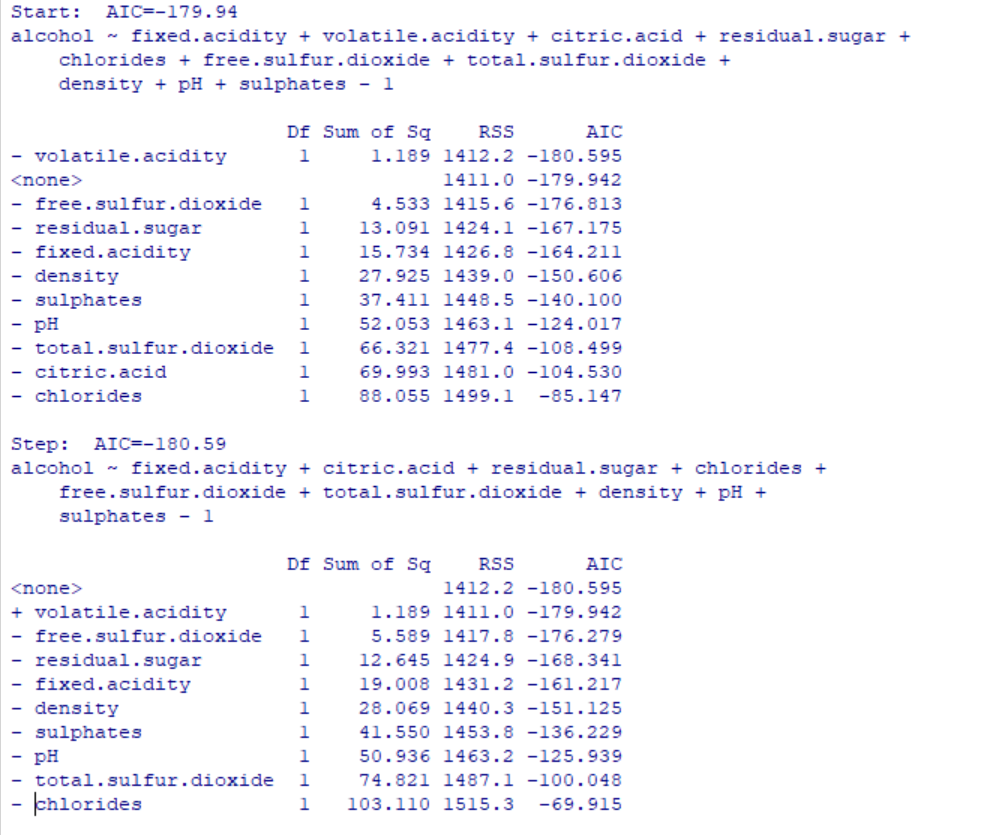
■ 다중회귀분석

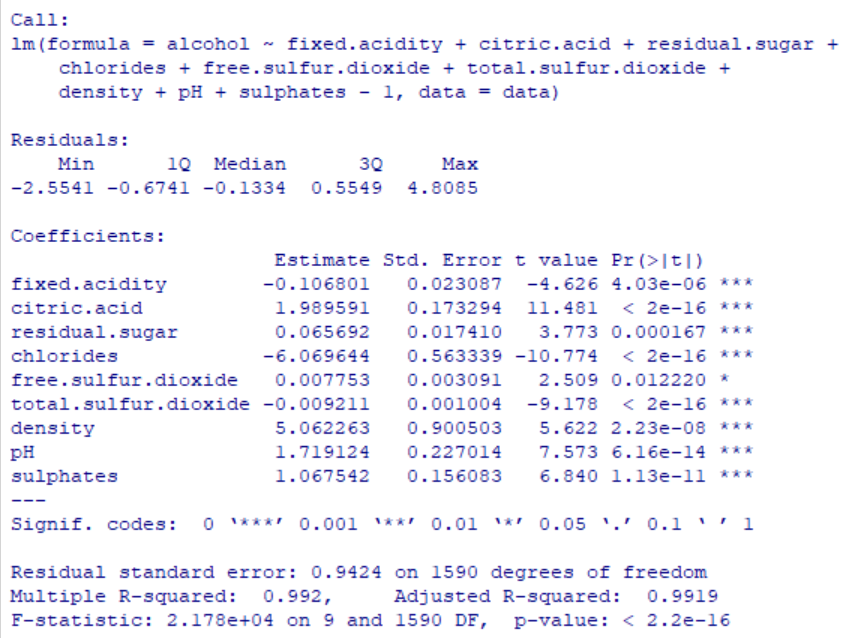


<회귀분석 결과>

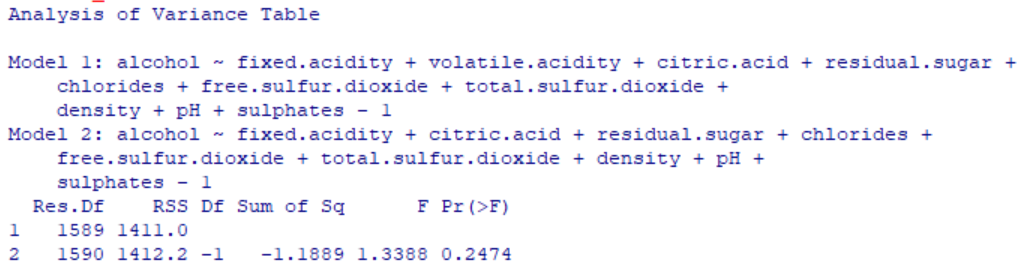
라는 회귀식을 만들 수 있었고,

잔차의 기술통계량은 평균 -2.5744, 중앙값은 -0.1391이 나오고, 결과 아랫쪽의 F-statistic 결과의p-value를 보면 <2.2e-16으로 유의수준 0.05보다 작아 이 모델은 유의하다는 판단을 내릴 수 있었고, 두 번째로는 coefficients 항목인데, 여러 변수들의 p-value를 확인한 결과, 휘발성 산(volatile.acidity)를 제외한 모든 변수들이 유의한 관계가 있는 것을 판단할 수 있었으며, 다중 회귀식이 선형을 만족한다는 것을 알 수 있었고. Coefficients 결과를 보았을 때, 각 변수가 통계적으로 유의한 경우 p-value 뒤에 \* 마크가 붙으며, \*의 개수가 늘어날수록 통계적으로 유의할 확률이 커지는데, 유리 이산화황(free sulfur dioxide)을 제외한 나머지 변수들은 모두 비슷한 정도로 유의하다는 것을 알 수 있었습니다. 그 결과로 이를 더 정확히 알아보기 위하여 변수 선택법 중 stepwise 방법을 이용하여 새로운 회귀 모델을 만들었는데, 휘발성 산을 제외한 다른 변수들로 구성되어있는 새로운 회귀 모델이 생성 되었습니다.

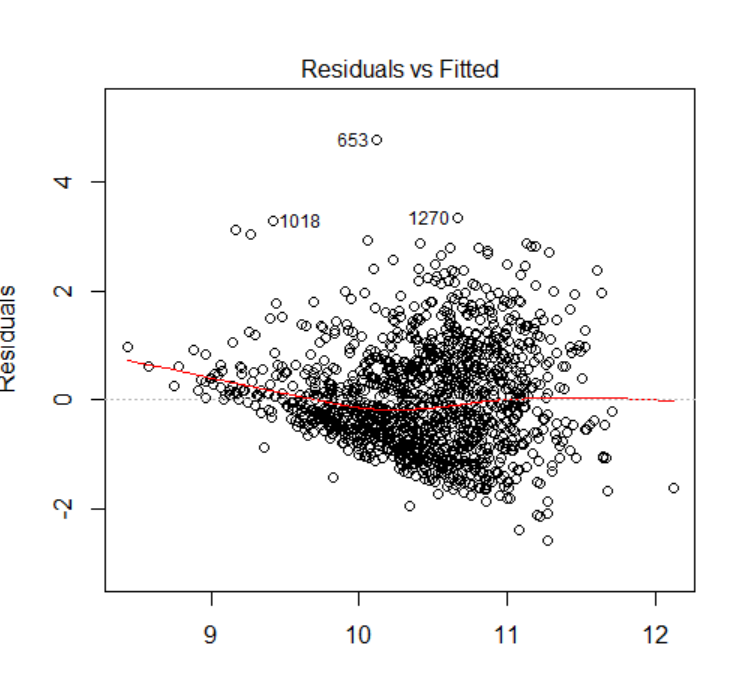


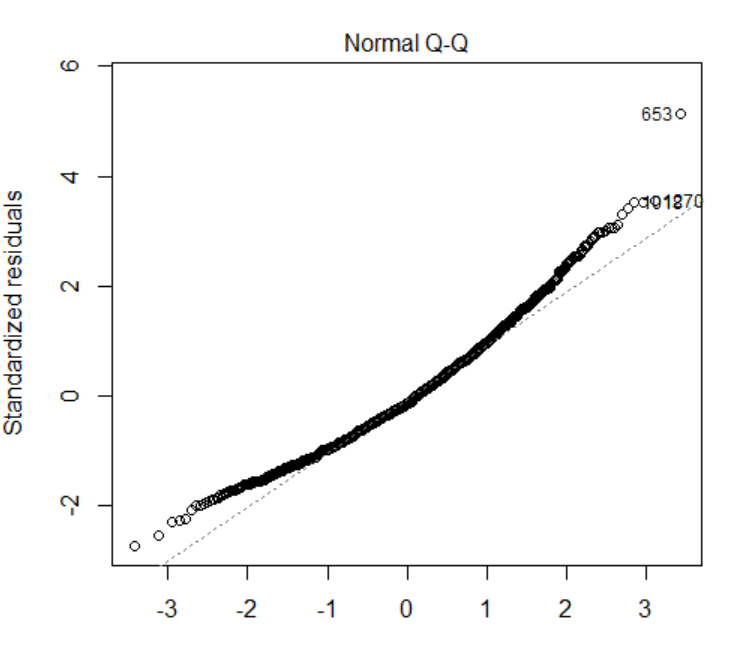


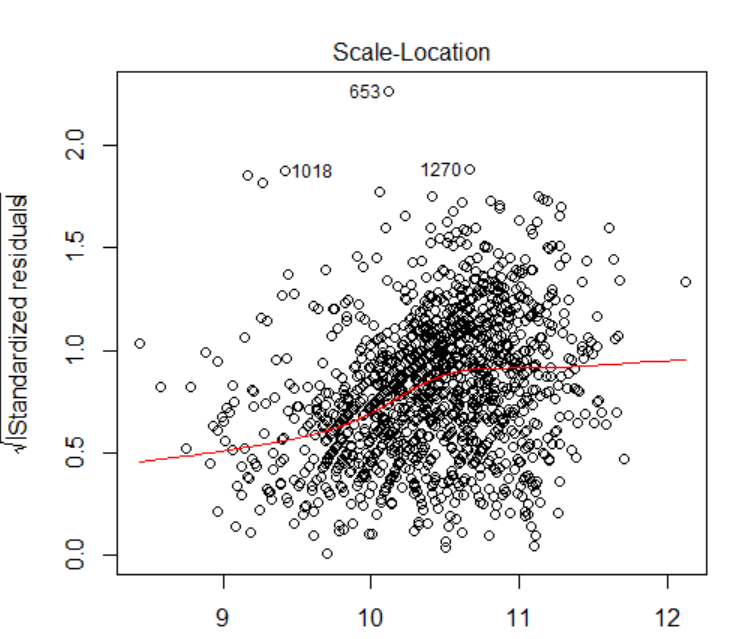
이 때, 이 모델이 기존의 회귀모델과 통계적으로 어떠한 차이점이 있는지 알아보기 위하여 Partitial F-test를 해보았고, 이에는 anova 함수를 사용하였습니다.



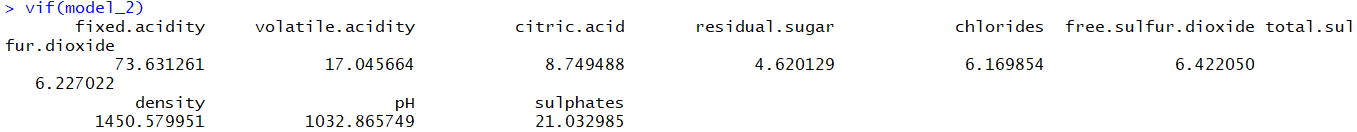
이 결과로 보았을 때, 두 모델의 alcohol에 대한 설명력의 차이는 없다고 볼 수 있었고, 왜 그런지 문제를 보았을 때, 두 모델 다 rsquare와 adjusted rsquared의 값이 0.99이상으로 과적합을 의심해 보아야 한다는 생각이 들었습니다.



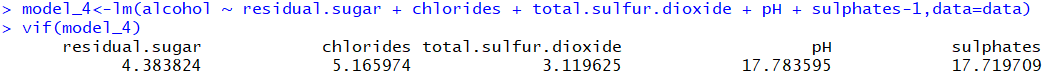




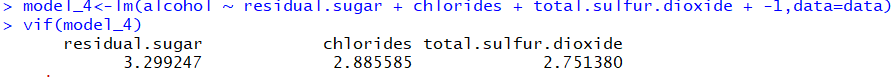
위와 같은 그래프들을 보았을 때, 정규성은 만족하는 듯이 보였지만, 독립성을 만족하지는 못하는것으로 보여서, vif 함수를 통하여 다중 공선성을 확인해 보았습니다. 회귀 모델의 설명 변수들 사이에 상관관계가 있는 경우가 있는데, 보통 VIF 값이 10이상 넘으면 다중공선성을 의심해 보아야 하는데, 한 예측 변수와 다른 예측 변수 사이에 높은 상관성이 있다는 것을 깨닫고, 다중공선성 해결에 관하여 알아보았습니다.

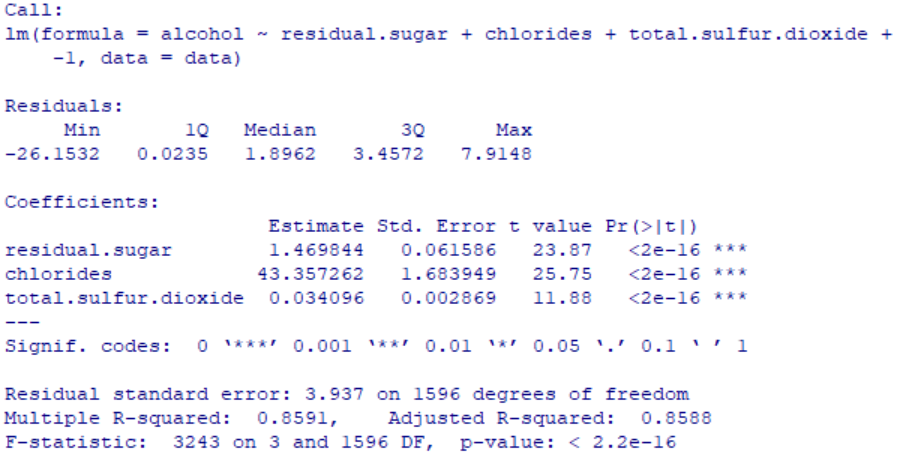


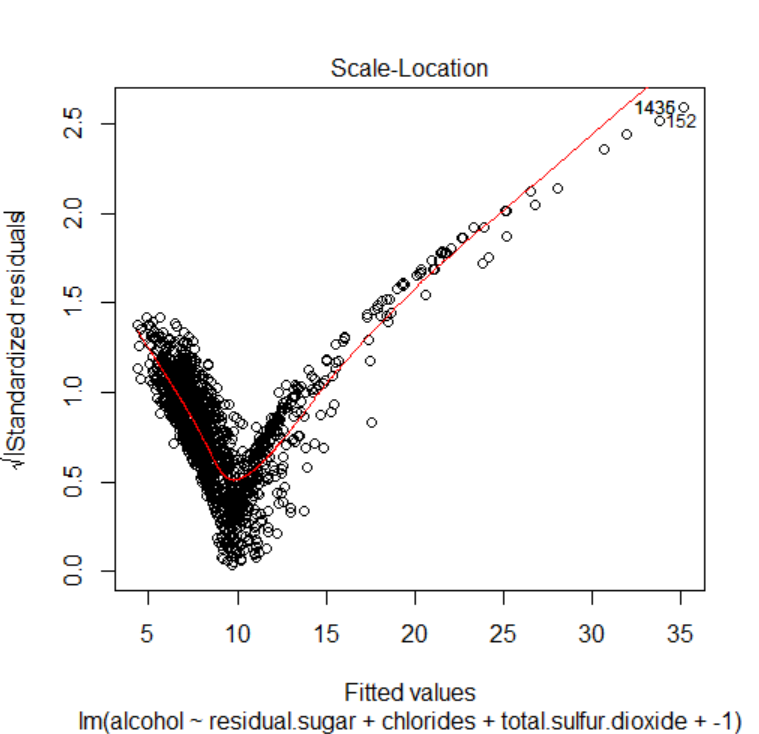
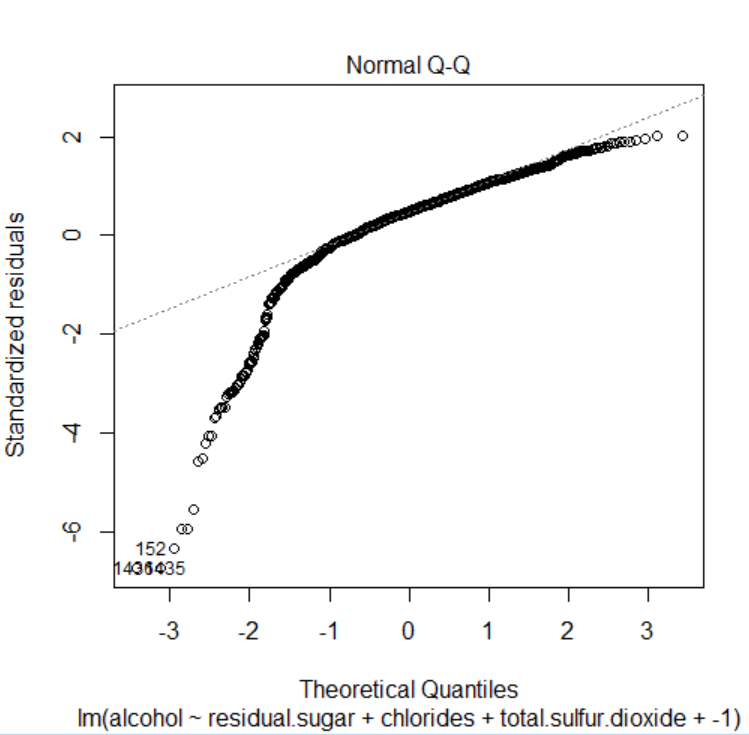
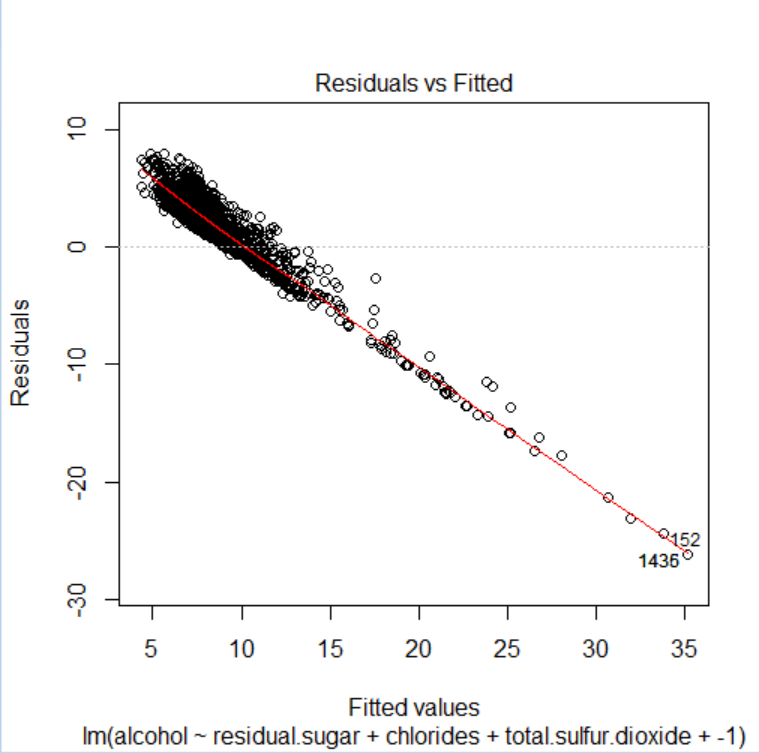
첫 번째 방법으로는 직관적으로 알코올에 미치는 영향이 큰 변수와 포함하는 변수를 설정하여서 이를 통해 변수를 선택하는 방법으로 해보았습니다.



하지만, 다른 부분에서는 어느 정도 미약하게 나마 개선이 되었지만, 이 때에도 PH와 황산염에서 다중공선성 문제는 해결되지 않아서 처음 vif함수를 썼을 때, 높은 것을 제외하고, 회귀분석을 시행해 보았습니다.







이 때, 다중공선성 문제는 해결 되었으나, 정규성 문제나 기본적인 회귀분석에 있어서 문제가 발생하여서, 다중공선성 해결하는 방법에 대해서 더 알아보았습니다.

3. 다중공선성과 다중공선성에 대한 대책(<http://contents.kocw.net/document/ch8_7.pdf>)

- 다중공선성을 내포하고 있는 회귀모형의 경우 회귀분석의 목적이 개별개수에 대한 추정이 아니라 종속변수 전체에 대한 예측에 있다면 다중공선성은 문제가 되지 않습니다

- 만일 회귀분석 목적이 사용되는 자료의 특성상 개별 설명변수의 종속변수에 대한 효과나 영향을 분리하는 추정일 경우, 이러한 개별변수에 대한 분리 추정을 할 수 없게 되는 문제점이 있으므로 이를 완화 시키거나 해결할 필요가 있습니다

다중공선성의 기본적 해결방법으로는

1. 관측값을 늘려 표본크기를 증가시킴

2. 원자료에 대해 차분 혹은 로그변환을 하거나 명목변수는 실질변수를 사용함.

3. 사전정보를 이용하여 변수를 상관관계가 높은 다른 변수로 대체함

◈ VIF 에 대한 설명   
  
 VIF란 데이터의 다중공선성을 진단하는 방법 중 가장 많이 쓰이는 것으로, 'Variation Inflation Factor' 의 줄임말입니다. 한글로는 '분산 확대 인자' 또는 '분산 팽창 계수' 또는 '분산 팽창 요인' 등으로 불립니다.  
  
VIFk는 k 번째 변수의 다중공선성의 정도를 의미한다. 즉, VIF 는 입력 데이터 전체의 다중공선성 정도를 하나의 수치로 제시하는 것이 아니라, 각각의 입력 변수들을 대상으로 계산됩니다.   
  
VIFk의 계산 방식은 분석에 사용될 데이터가 입력변수 n개와 종속변수 1개로 구성된다고 할 때, 총 변수는 n+1 이고, 이 데이터의 변수들은 모두 수치형 변수이어야 합니다.   
  
그럼, VIFk는 변수 k를 종속변수로 지정하고, 나머지 n-1 개의 입력변수를 입력변수로 지정하여 회귀분석을 수행한다. 이 회귀분석에서 종속변수 Y는 제외 (다중공선성은 입력변수들 간의 상관관계를 측정하는 것이기 때문)  
이 회귀분석에서의 결정계수 Rj2을 구한 후 아래의 수식으로 VIFk를 구한다. 

VIFk= 1 / (1 - Rj2)

결정계수 Rj2 값은 0에서 1의 값을 가지므로 VIF 값은 1에서 무한대의 범위를 갖습니다.

(1) 결정계수 Rj2값이 0에 가깝다는 것은 입력변수 k가 다른 입력변수들과 상관성이 거의 없다는

것을 의미하고,

(2) 결정계수 Rj2값이 1에 가깝다는 것은 입력변수 k가 다른 입력변수들과 상관성, 즉 다중공선성

이 크다는 것을 의미합니다. 

→ VIF 값이 1 에 가까울수록 다중공선성의 정도가 작은 것이며, 반대로 값이 커질수록 다중공선성의 정도가 큰 것을 의미

회귁분석의 결정계수가 종속변수 Y와 독립변수 X1, X2, X3,..등과 얼마나 가까운지, 즉 얼마나 밀착해있는지의 개념을 이용하여, 독립변수들간의 이런 밀착성(다중공선성)에 대해서 보고자 하는 것으로, 어떤 특정 독립변수 의 VIF는 와 를 포함하지 않는 나머지 독립변수간의 결정계수를 보는 개념입니다.

예를 들어 A, B, C 세 개의 독립변수가 있고, B의 VIF를 구하려면 일단 B를 종속변수로 놓고, A와 C를 독립변수로 하고 그런 다음 회귀분석을 하여 결정계수를 구하는 개념.

VIFB= 1 / (1 – RB2)

◈ 다중공선성 유발변수의 탐색 및 제거

- 심각한 다중공선성을 유발시키는 설명변수를 확인한 다음 이들 변수중 추정에 해로운 변수를 제거하는 방법

a) 다중공선성 유발변수의 탐색

i) 각 설명변수를 다른 모든 설명변수 들에 대하여 회귀분석 한 다음 R2 값을 서로 비교하여 가장 큰 R2 값을 갖는 회귀방정식상의 선형 함수 관계가 높은 변수가 다중공선성 문제를 유발시키는 것으로 판정 할 수 있습니다.

ii) 특정 설명변수를 제외시켰을 때의 R2 값이 제외시키기 이전의 R2 값에 비 해 감소하지 않을 경우 이 설명변수에 의해 다중공선성이 높아진 것으로 볼 수 있습니다.

b) 다중공선성 유발변수의 제거

i) 선형상관관계가 높거나 높은 다중공선성을 유발시키는 것으로 확인된 두 설명변수와 종속변수의 상관계수를 산출, 비교하여 상관계수 값이 작은 설명변수를 제거합니다. 이는 종속변수에 대한 추가적인 설명능력이 상대적으로 작다는 것을 의미하기 때문

ii) β 계수(βcoefficient)나 탄력성(elasticity)를 이용한다 β 계수(βcoefficient)는 최소자승 추정치와 β 계수나 탄력성 값을 비교하여 절대치가 작은 변수를 제외 시킵니다.

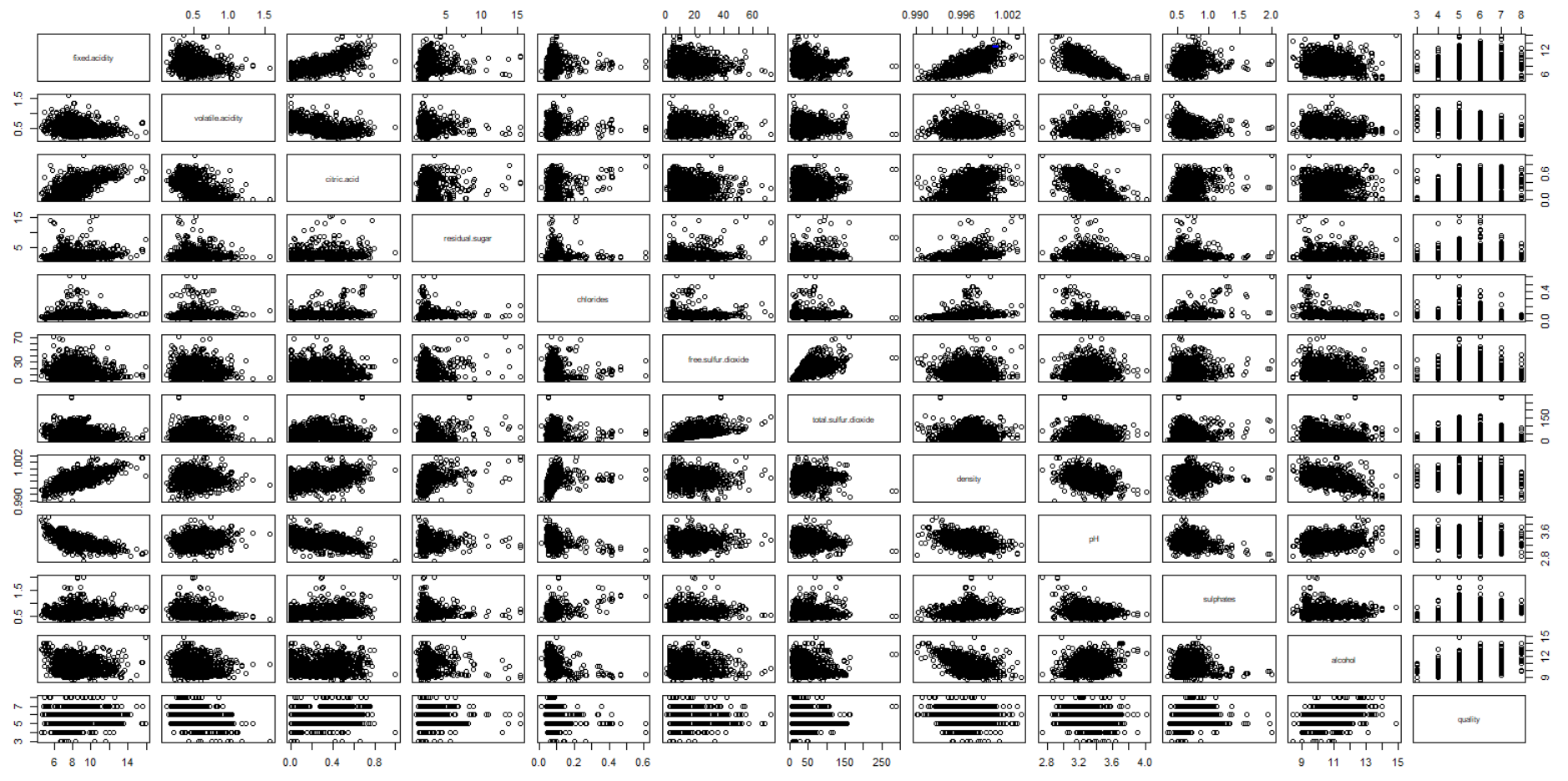
- 일단 회귀모형 내에 도입된 변수는 합당한 이론이나 가설에 입각하여 채택된 것이므로 무작정 변수를 제거하면 다중공선성 문제는 완화될 수 있으나 때로는 더 심각한 “설정모형 오류”를 유발할 수 있기 때문에 주의.

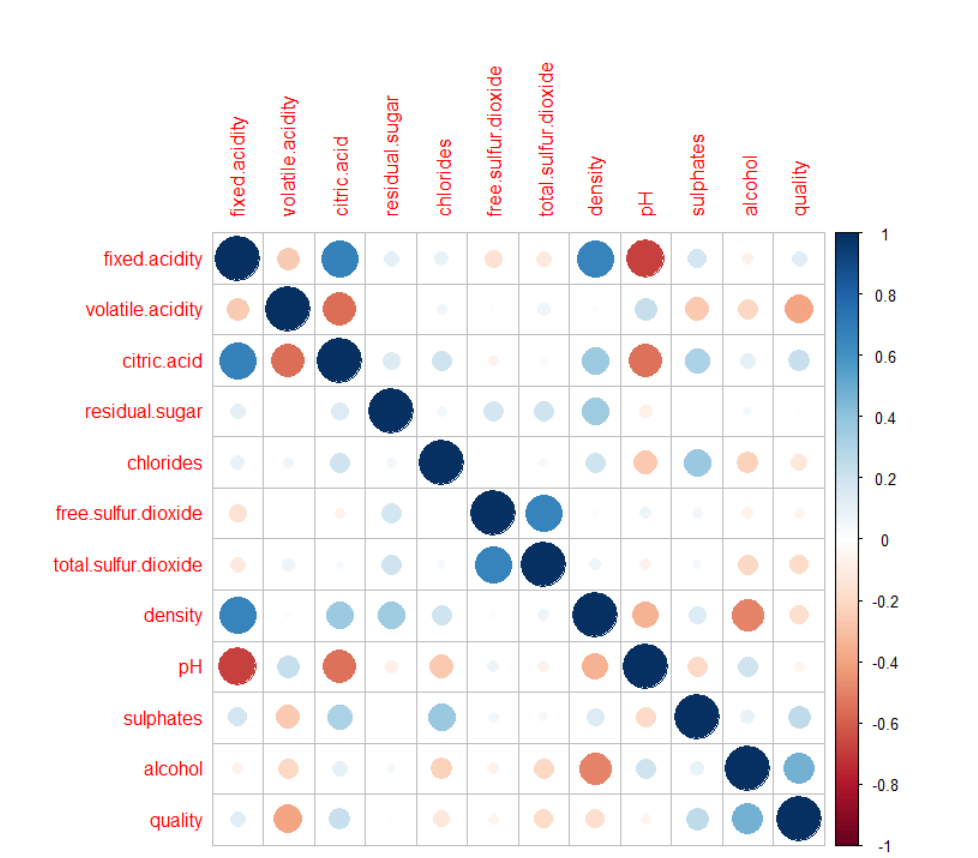
2) 자료와 모형의 보완 및 변형 - 다중공선성문제를 해소하는 궁극적이고 최선이 되는 방책은 표본관측치를 추가적으로 확보하거나 모형을 변형시켜 기존의 표본관측치 값에 의해 야기된 다 중공선성을 완화시키는 것입니다

a) 일차분차(first difference) 또는 비율(ratio)을 사용하는 방법. 기본회귀모형,

대신에, 일차분차식, 비율식 을 이용하여 추정함으로써 다중공선성을 완화시킬 수 있습니다

b) 추가적인 표본관측치를 확보하여 활용하는 방법 Æ 추가로 확보하는 자료들은 가능한 기존의 설명변수의 평균값과 상이한 값을 많이 포함시켜야 합니다.





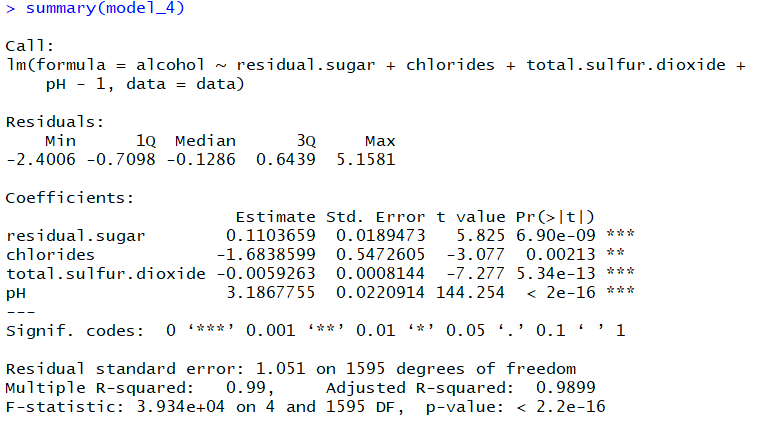
→ pairs를 통하여 선형관계를 가지고 있는 변수와 corrplot을 이용하여 상관성이 높은 변수를

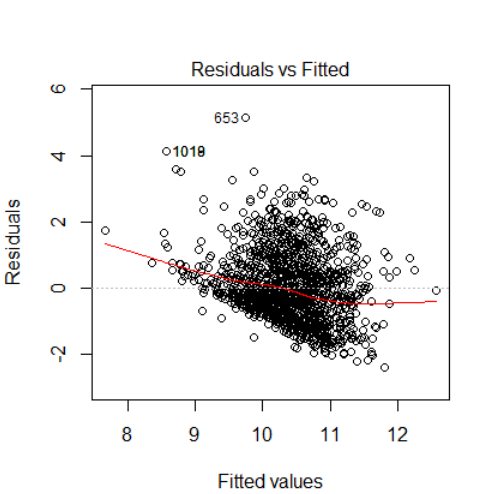
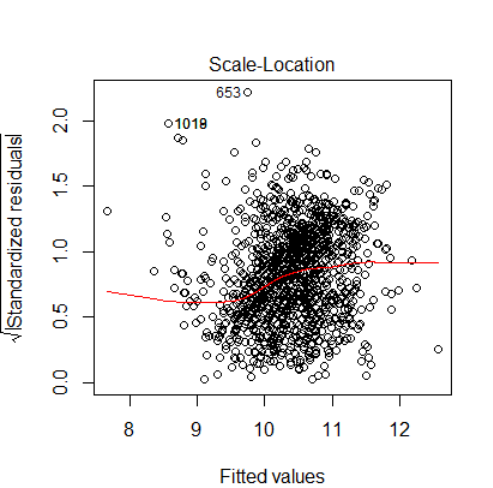
추출하였습니다.

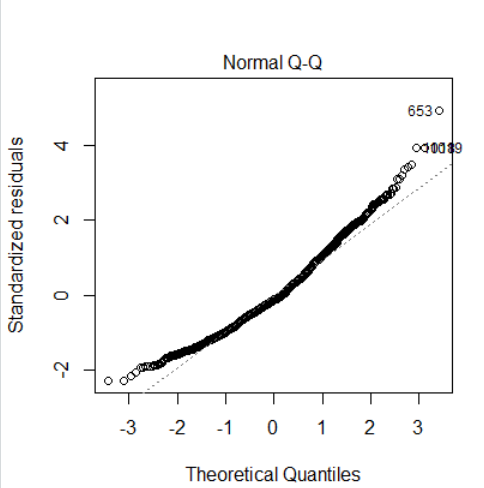
|  |  |
| --- | --- |
| **변수명** | **변수 설명** |
| fixed acidity | 비휘발성 산 |
| volatile acidity | 휘발성 산 |
| citric acid | 구연산(시트르산) |
| residual sugar | 잔당 |
| chlorides | 염화물 |
| free sulfur dioxide | 유리 이산화황 |
| total sulfur dioxide | 총 이산화황 |
| density | 밀도 |
| Ph | 산도(수소 이온 농도) |
| Sulphates | 황산염 |
| Alcohol | 알코올 |
| Quality | 와인 품질(나쁨 : 0 ~ 좋음 : 10) |



4가지 변수를 추출하여 회귀 모델을 생성하였고, 이를 통해서 다중 공선성 문제를 해결할 수 있었고, 다음과 같은 회귀분석을 할 수 있었습니다.





4. 결과

기존의 다중공선성 문제는 변수 선택을 통하여, 원하던 바로 해결하였지만, R-square가 너무 높게 나오게 되어서, 과적합을 의심하였는데, 이를 해결하지는 못하고, 과적합에 대해서 공부를 더 해보았습니다.

5. 과적합

◈ 과적합이란

과적합이라는 용어는 통계학이나 기계학습에서 주로 쓰이며, 과거데이터를 통해 모델을 세운다면 자주 마주하게 되는 문제입니다. 다른 중요한 통계적 개념들과도 밀접하게 연관되어 있는 것들이 많아 매우 중요한 개념입니다.

과적합은 모델이 실제 변수들 간의 관계보다는 과거 학습데이터 (training data)의 노이즈를 설명하게 되는 경우를 표현할 때 쓰입니다. 즉, 모델이 과거의 데이터를 너무 과하게 설명한 나머지 실제 변수들 간의 관계를 잘못 설명하게 되는 경우입니다. 이러한 모델은 이미 학습한 데이터는 잘 설명하는 것처럼 보이지만 아직 학습하지 않은 데이터에 대해서는 제대로 설명하지 못합니다.

◈ 과적합

* 모델이 학습 데이터셋 안에서는 일정 수준 이상의 예측 정확도를 보이지만, 새로운 데이터에 적용하면 잘 맞지 않는 것
* 층이 너무 많거나 변수가 복잡해서 발생하기도 하고 테스트셋과 학습셋이 중복될 때 생기기도 한다

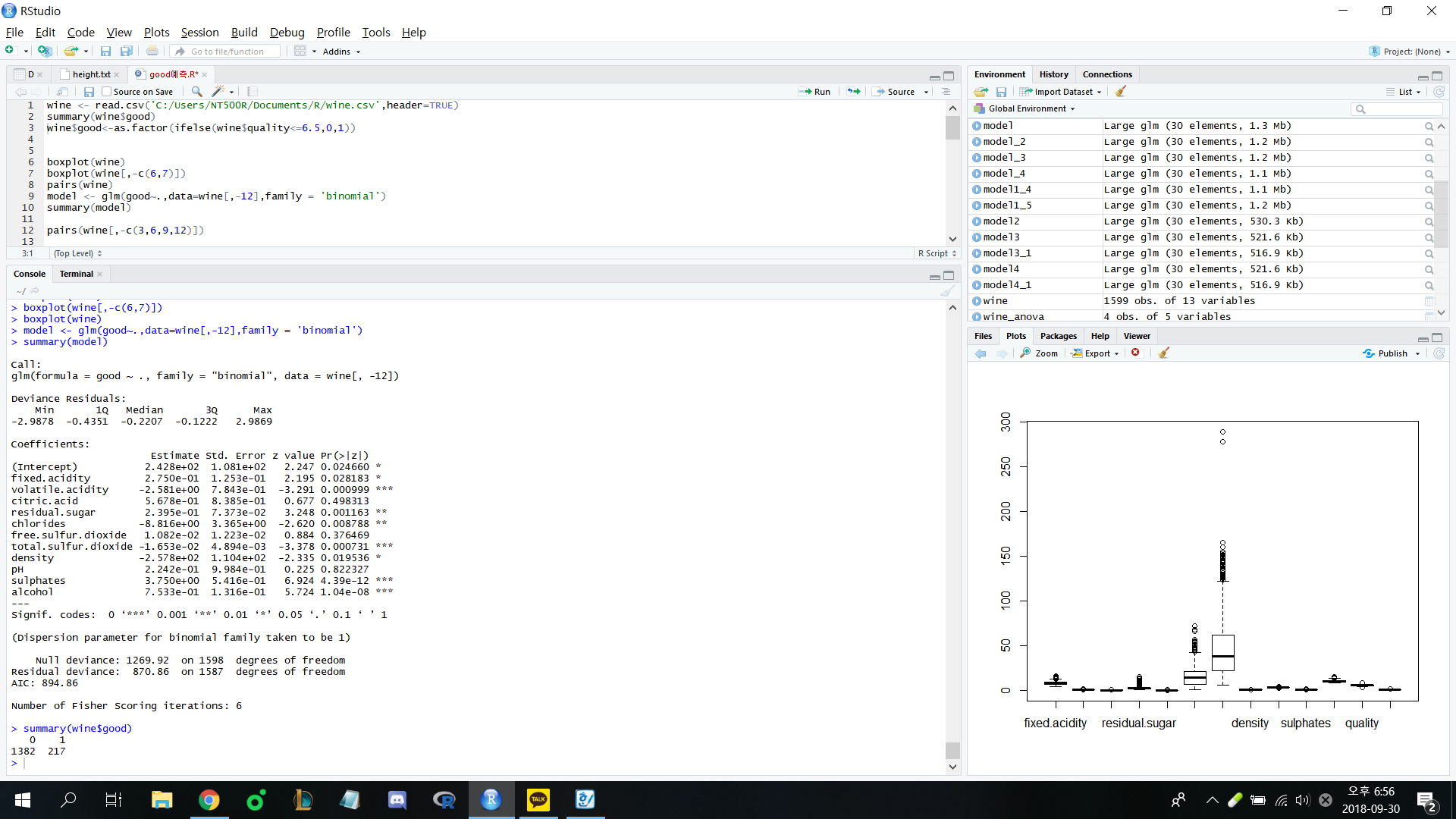
◈ 드롭아웃

* 드롭아웃은 단순한 기법을 통해 신경망의 일반화 성능을 향상시킬 수 있는데, 학습시킬 때 무작위로 뉴런을 ‘드롭아웃(제외)’ 시킨다
* 드롭아웃을 적용하면 실질적으로 많은 모델을 생성, 학습하고 그 안에서 예측을 실행하기 때문에 성능이 향상된다
* 하나의 모델로 학습하면 어쩔 수 없이 오버피팅이 생길 가능성이 있지만 다수의 모델로 학습을 하고 각 모델로 예측을 한다면 소위 ‘집단지성'을 얻게 되어 오버피팅의 위험을 줄일 수 있다
* 드롭아웃하는 신경망 모델을 본래 한 개지만 여러 개의 모델을 생성해서 학습을 실행하는 것을 ‘앙상블 학습(ensemble learning)’ 이라고 한다
* keep\_proba : tensorflow 에서 드롭아웃하지 않을 확률(학습할 때는 0.5 테스트할 때는 1.0 으로 하기 때문에 placeholder를 이용해서 정의한다)

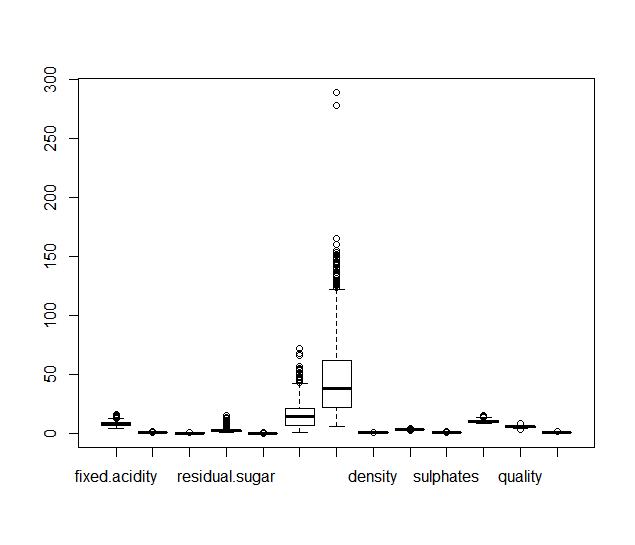
■ 로지스틱 회귀분석

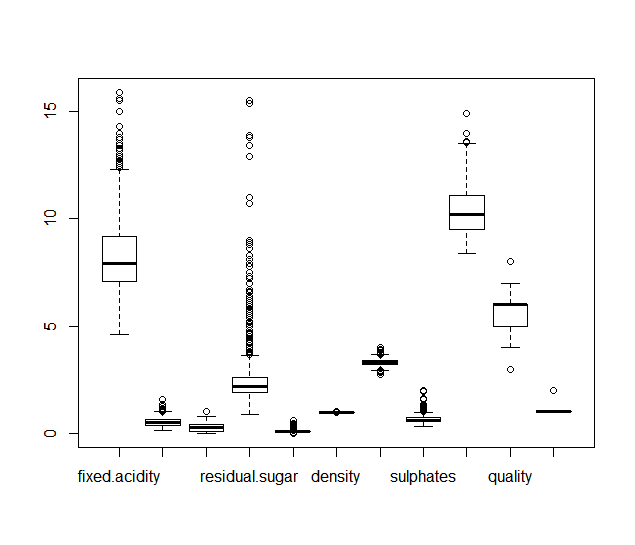
품질이 좋은 와인과 그렇지 않은 와인 예측

1.quality값을 6.5기준으로 binomial 데이터로 변환

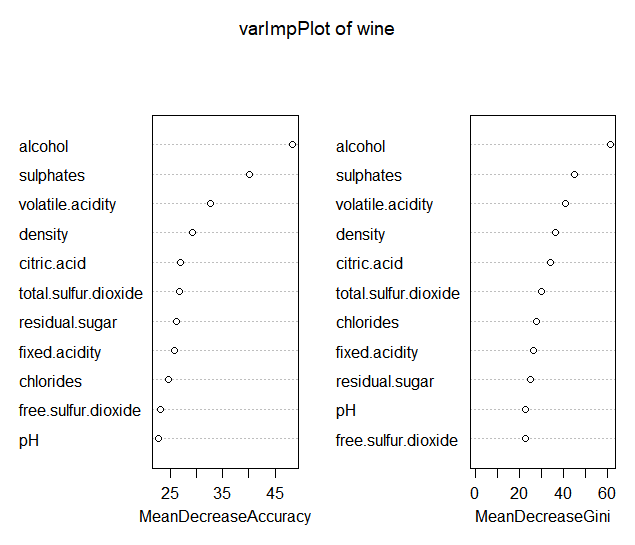


2.boxplot 확인작업



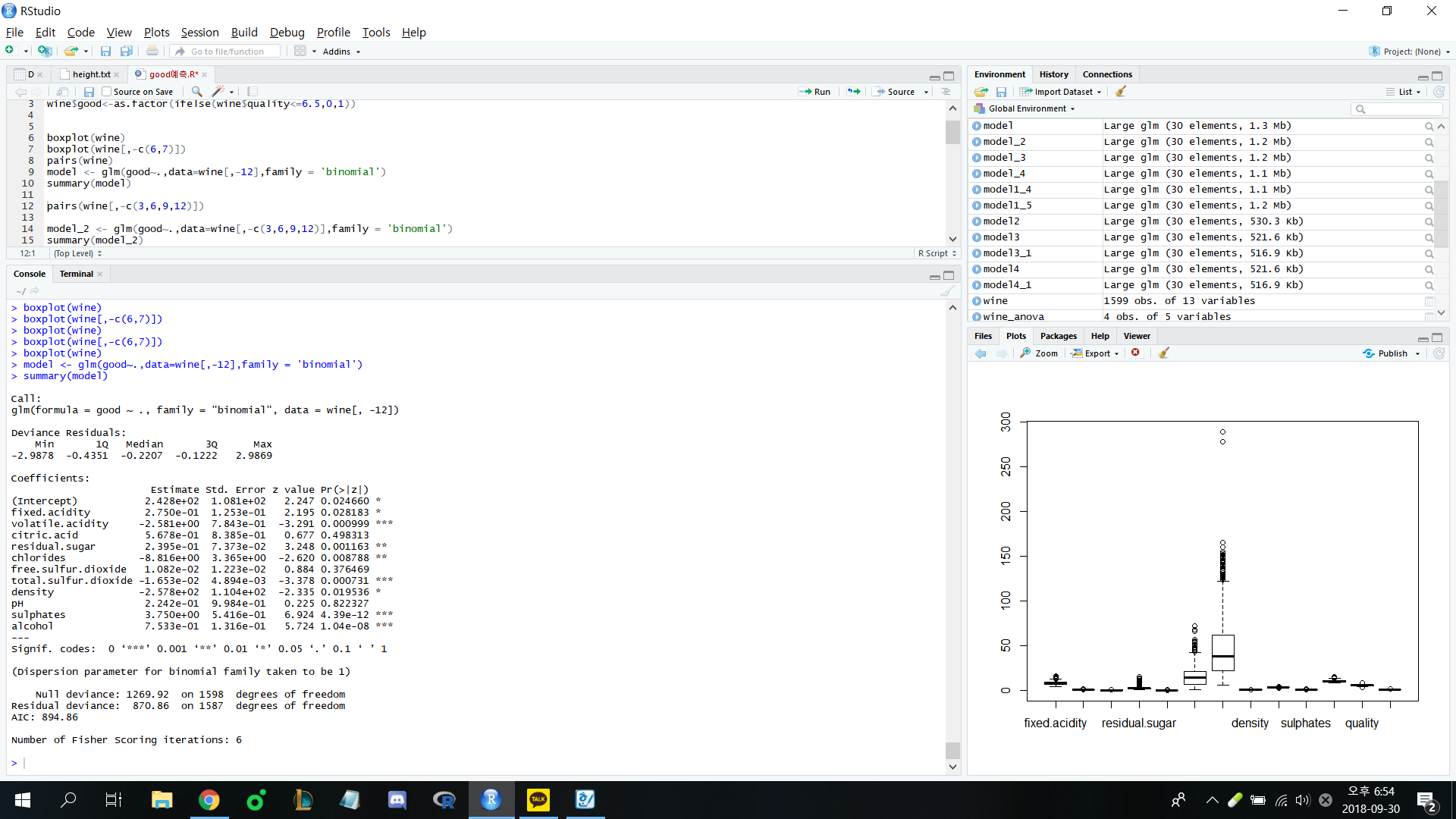


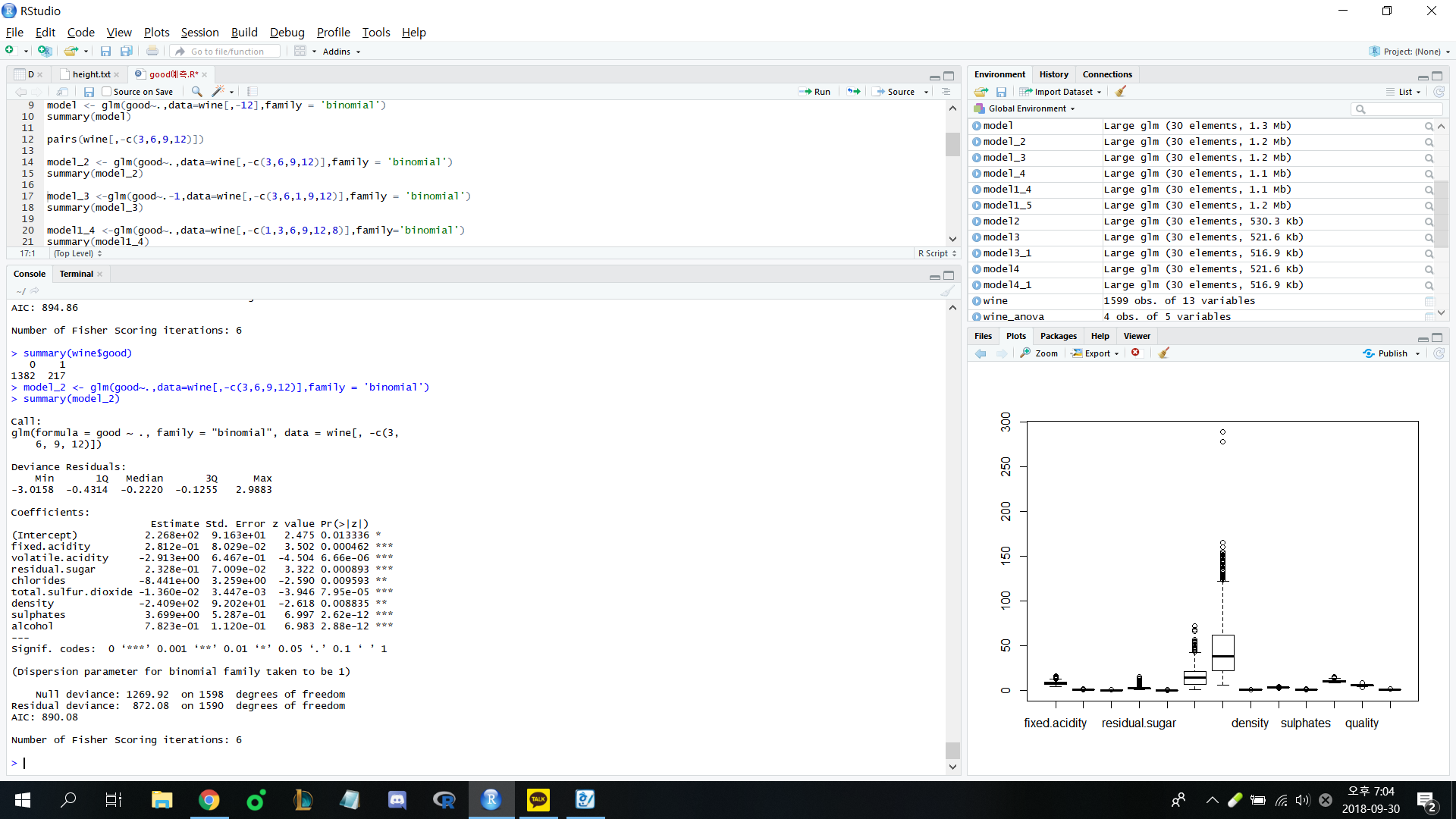
3.변수 중요도 체크



4.모델링

1)good~all



2) 1)에서의 p값을 기준으로 0.05보다 큰 변수들 제거(step사용했을 때 가장 좋은 모델)

4. 모델 평가(predict ftc사용)

