

R Project

#wine quality analysis

과목: 지능형 시스템

교수님: 황규백 교수님

학부: 기계공학부

학번: 20141682

이름: 고동환

제출일자: 2018년 6월 11일

목차

1. 개요
2. 문제풀이 및 결과
3. 의의 및 평가
4. 출처

1. 개요

본 과제는 R프로그래밍을 이용하여 wine data set을 통하여 와인의 quality와 설명 변수들간에 어떤 관련이 있는지 분석해보는 것이다. 본 과제는 주어진 요구에 맞추어서 각각 red wine과 white와인에 대해서 문제를 풀이하고 해석하는 방식으로 진행한다.

2. 문제풀이

<1번 문제># 결측치를 확인하고 중간값으로 대체

사용된 라이브러리 1 library("DMwR")

결측치를 중간값을 바꾸어주는 DMwR::centralImputation 함수가 들어있는 패키지이다.

사용된 함수 1

redwine <- read.csv("winequality-red.csv",header=TRUE,na.string="NA",stringsAsFactors=default.stringsAsFactors(),sep=";")

whitewine <- read.csv("winequality-white.csv",header=TRUE,na.string="NA",stringsAsFactors=default.stringsAsFactors(),sep=";")

#이 함수를 통하여 엑셀로부터 데이터를 받을 수 있으며 sep=”;”사용하여서 세미콜론으로 구별된 데이터를 분리할 수 있다.

사용된 함수 2

sum(is.na(redwine))

sum(is.na(whitewine))

#이 함수를 통하여 데이터 셋에 전체 NA값이 몇 개 있는지를 확인해 볼 수 있으며 결과값은 0로 원본에 결측치가 존재하지 않는다고 볼 수 있다.

사용된 함수 3

DMwR::centralImputation(credwine)

DMwR::centralImputation(cwhitewine)

#DMwR패키지를 이용하여서 NA값을 중간값을 대체할 수 있다. 하지만 결측치는 존재하지 않는다

사용된 함수 4

max(credwine$quality) #최댓값 확인 8 하지만 whitewhine과 비교하기 위해 그래프에서는 9까지 그려줌

min(credwine$quality) #최솟값 확인 whitewine과 마찬가지로 3

max(whitewine$quality) #최댓값 9

min(whitewine$quality) #최소값 3

사용된 함수 5

plot(jitter(credwine$quality),xlab="number", ylab="quality",main="redwine",pch=20,cex=0.4, ylim=c(3,9))

plot(jitter(cwhitewine$quality),xlab="number", ylab="quality",main="whitewine",pch=20,cex=0.4, ylim=c(3,9))

#plot함수를 이용하여서 산점도를 그린다. 이 때 x축과 y축 main의 이름을 정해주고 있으며 범위는 ylim을 사용해서 [3,9]까지의 값을 갖도록한다. 이는 whitewine의 범위가 [3,9]이기 때문이다. 또한 jitter방식을 사용하여서 겹치는 데이터를 분산시켜 값에 따라 밀도를 보기 쉽게하였다.

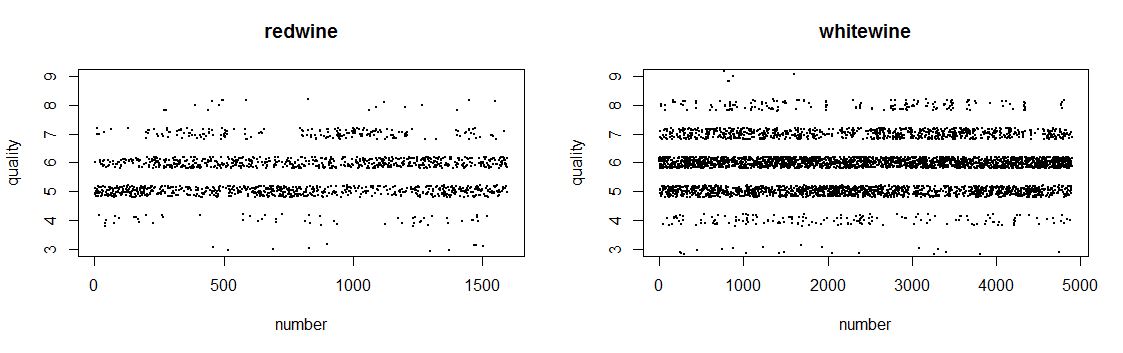
사용된 함수 6

opar <- par(mfrow=c(1,2))

par(opar)

#이 함수 사이에 redwine과 whitewine 그래프를 그려 넣으면 1행 2열의 그래프를 준비할 수 있다.

결과



redwine과 white와인의 반응 변수의 그래프를 동시에 나타내어 비교해 볼 수 있다. whitewine에 사용된

<2번 문제> #forward backward both방법을 사용하여 결측치가 제거된 데이터 품질 향상시키기

#문제 1번과 중복되는 라이브러리 library("DMwR") 및 read.csv함수는 생략합니다.

사용된 함수 1

lmredwine <- lm(quality ~ ., data = credwine)

lmwhitewine <- lm(quality ~ ., data = cwhitewine)

#선형회귀를 수행하는 함수로 lmredwine에 quality를 반응 변수로 선형회귀를 수행하고 있다.

사용된 함수 2

forwardRedwine <- step(lmredwine, direction = "forward")

forwardWhitewine <- step(lmwhitewine, direction = "forward")

#모델의 품질을 step의 AIC척도를 이용하여 평가하는 함수 중에서 forward방식을 이용한다.

#forward방식이란 유의미한 데이터를 집어넣어가면서 모델의 품질을 향상시키는 것이다.

#여기서 redwine의 forwardAIC값은 -1375.49를 의미하며 가장 기본적인 수치이다.

#여기서 whitewine의 fowardAIC값은 -2788.44를 의미하며 가장 기본적인 수치이다.

사용된 함수 3

backwardRedwine <- step(lmredwine, direction = "backward")

backwardWhitewine <- step(lmwhitewine, direction = "backward")

#모델의 품질을 step의 AIC척도를 이용하여 평가하는 함수 중에서 backward방식을 이용한다.

#backward방식이란 무의미한 데이터를 빼내가면서 모델의 품질을 향상시키는 것이다.

# 순서대로 rewine의 경우 density, fixed.acidity, residual.sugar, citric.acid를 제거하여 AIC값이

# redwine의 backwardAIC값이 -1380.8로 최적화가 되었다.

# 순서대로 whitewine의 경우 critic.acid, chlorides, total.sulfar.dioxide를 제거하여 AIC값이

# whitewine의 backwardAIC값이 -2793.6으로 나왔다.

사용된 함수 4

bothRedwine <- step(lmredwine, direction = "both")

bothWhitewine <- step(lmwhitewine, direction="both")

#모델의 품질을 step의 AIC척도를 이용하여 평가하는 함수 중에서 both방식을 이용한다.

#both방식이란 무의미한 데이터를 빼내가거나 유의미한 데이터를 넣어가면서 모델의 품질을 향상시키는 것이다.

#여기서는 결론적으로 각각의 와인이 backward와 같은 종류의 설명 변수를 빼내었을 때가 최적화가 되었으며 AIC값도 동일하게 유지되었다.

#따라서 backward와 both의 성능은 동일하며 앞으로 연산속도가 더 빠른 backward방식을 가장 우수하다고 판단하여 사용하도록 하겠다.

<3번 문제> 피어슨의 상관계수가 가장 높은 5쌍 찾기

#read.csv함수는 생략합니다.

사용된 라이브러리 1

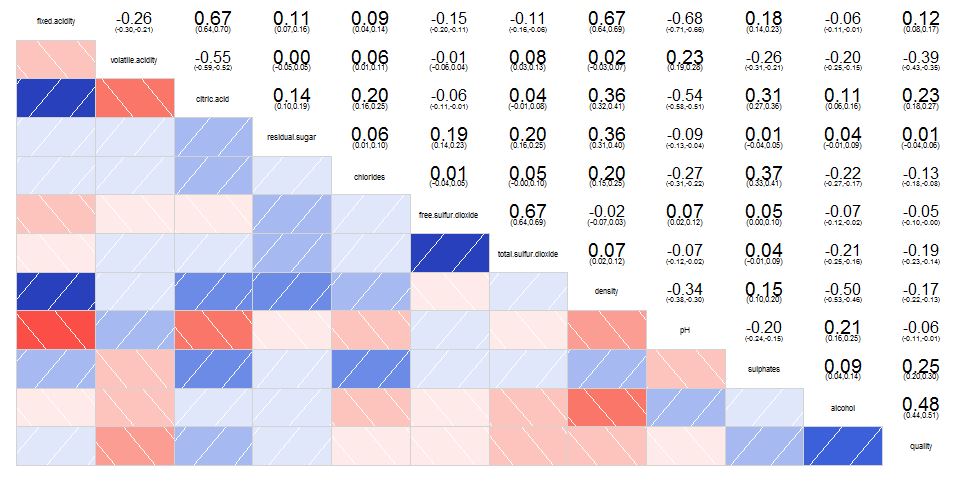
library("corrgram")

#피어슨의 상관계수를 그래프의 형태로 보기좋게 만들어 주는 함수이다.

사용된 함수 1

cor(redwine)

corrgram(redwine,upper.panel=panel.conf)



Redwine의 corrgram이다. 여기서 그래프를 바탕으로 상관관계가 가장 높은 5쌍을 선별하면 다음과 같다.

#가장 상관관계가 높은 것은 절댓값을 기준으로 -0.68<fixed.acidity,pH>, 0.67<fixed.acidity,critic.acid>, 0.67<fixed.acidity,density>

#0.67<free.sulfur.dioxide,total.sulfur.dioxide>,-0.55<volatile.acidity,citric.acid>이다.

#첫번째로 -0.68<fixed.acidity,pH>는 타르타르산과 pH의 상관관계를 나타내는 데 pH는 산성도와는 반대이므로 이러한 결과가 도출되었다.

#타르타르산은 주석산이라고도 불리우며 흰색 결정을 가진 유기산의 일종이며 포도주에서 발견되는 주요한 산의 일종이다.

#두번째로 0.67<fixed.acidity,citric.acid>는 타르타르산 시트르산의 상관관계이다. 타르타스산이 높은 와인은 시트르산도 높은 것을 볼 수 있다.

#시트르산은 구연산이라고도 하며 귤속 과일에서 주로 발견되는 약한 유기산이다. 자연적인 보전제이며 산성 또는 신맛을 강화 한다.

#세번째로 0.67<fixed.acidity,density>는 타르타르산과 와인cm^3당 gram수의 관계를 의미한다. 타르타르산의 산성도가 높을 수록

#농후한 와인으로 만든다는 것을 생각해 볼 수 있다.

#네번째로 0.67<free.sulfur.dioxide,total.sulfur.dioxide> 무이산화황과 이산화항의 관계이다.

#둘다sulfar(황)을 포함하므로 이런 비례하는 상관관계가 발생하였음을 생각해 볼 수 있다.

#free.sulfur.dioxide는 wine에서 설탕과 과일의 껍데기, 잔여물 그리고 botrytis cinerea에 감염된 과일에서 생산된 이산화황을 제거한 이산화황 비율이다.

#total.dioxide는 전체 이산화황의 비율이므로 전체적으로 이산화황이 많으면 그 일부인 free.sulfur.dioxide의 양도 증가하며 반대로 free.sulfur.dioxide가 증가하면 전체 sulfur양도 증가한다.

#이산화황은 와인보존에 관련이 있다.

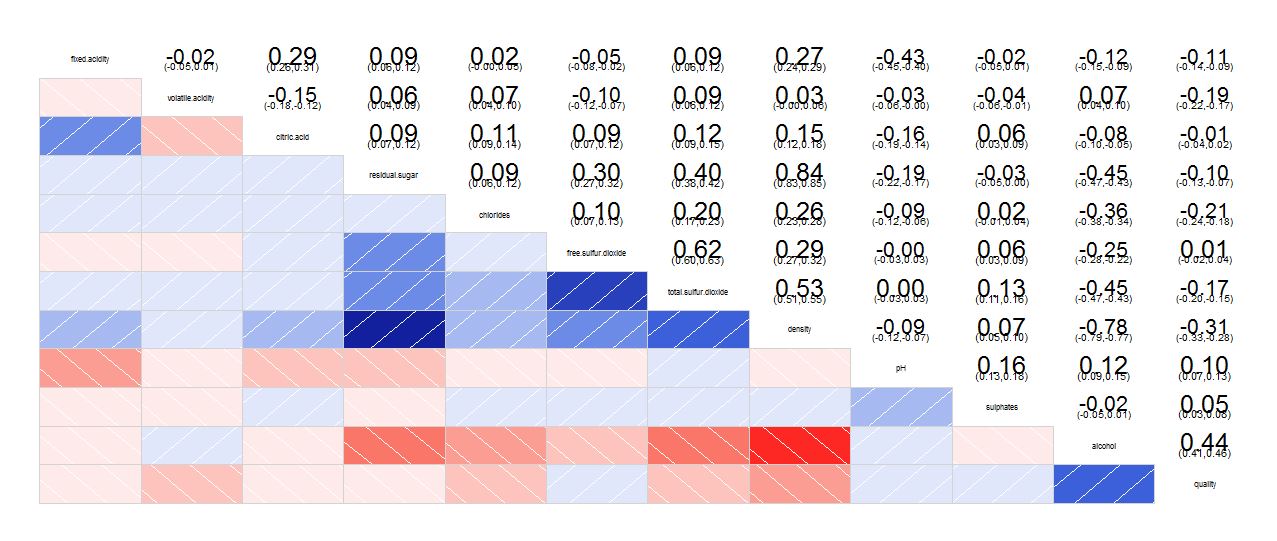
#다섯번째로 -0.55<volatil.acidty,critic.acid> 아세트산과 시트르산의 상관관계이다. 와인에서 아세트산이 증가하면 시트르산은 감소하였다는 것을 볼 수 있다.

#아세트산은 식초의 원료이며 에탄올을 산화시킨 아세트알데하이드를 산화시키므로써 생산된다.

사용된 함수 2 #와인 별 구별

cor(whitewine)

corrgram(whitewine,upper.panel=panel.conf)



여기서 whitewine의 corrgram을 바탕으로 가장 상관계수가 높은 5쌍을 선별하면 다음과 같다.

# 첫 번째 +0.84 residual.sugar,density 발효 후 남아있는 당량과 밀도가 가장 높은 상관관계를 가졌으며 비례함을 볼 수 있다.

# 두 번째 -0.78 density, alcohol 밀도와 알코올이 두 번째로 높은 상관관계를 가졌으며 반비례함을 볼 수 있다.

# 세 번째 +0.62 free.sulfur.dioxide,total.sulfur.dioxide 무이산화황과 이산화황은 redwine에서와 마찬가지로 서로 영향을 주며 비례하는 관계이다.

# 네 번째 +0.53 total.sulfur.dioxide,density 무이산화황과 밀도가 비례하는 관계를 지녔음을 볼 수 있다.

# 다섯 번째 -0.45 residual.sugar.alcohol (=-0.45 <- total.sulfur.dioxide,alcohol) 발효 후 남아있는 당량과 알코올이 반비례하는 관계에 놓여 있다.(전체 이산화황과 알코올의 관계도 동일한 정도의 반비례 관계를 지니고 있다.)

<4번 문제>

#read.csv함수는 생략합니다.

<redwine>

사용된 함수 1

lmRedwine <- lm(quality ~ ., data=redwine)

twointerRedwine <- lm(quality ~ (fixed.acidity+volatile.acidity+citric.acid+ residual.sugar+chlorides+ free.sulfur.dioxide + density + pH + sulphates + alcohol)^2, data = redwine)

#quality를 반응 변수로 선형 회귀를 수행하고 있다.

#두 번 째 함수의 경우 선형 회귀를 수행 시에 모든 경우에 대해 (A+B+C)^2의 형식으로 2차 interaction term을 추가하고 있다.

사용된 함수 2

anova(lmRedwine)

anova(twointerRedwine)

anova(lmRedwine,twointerRedwine)

#anova함수를 이용하여 일원 분산 분석을 수행한다.

첫 번째의 경우 제곱합이 666.41 제곱합 평균이 0.420으로 나온 것을 확인 해 볼 수 있다.

두 번째의 경우 제곱합이 616.12 제곱합 평균이 0.399으로 나온 것을 확인 해 볼 수 있다.

이를 통해 2차 interaction term을 추가한 모델이 오차가 더 적은 것으로 볼 수 있다.

그리고 두 모델 자체를 anova함수로 비교해보기도 하였는데 p-value가 0.05미만이 나왔으므로 두 그래프는 서로 다른 모델이라고 볼 수 있다.

사용된 함수 3

opar <- par(mfrow=c(2,2))

plot(lmRedwine,which = c(1:3,5))

par(opar)

opar <- par(mfrow=c(2,2))

plot(twointerRedwine,which = c(1:3,5))

par(opar)

# plot함수를 이용하여서 그래프의 모양새를 관찰해보기도 하였다.

#우선 그래프로 성능을 비교해봤을 때에 두 그래프 모드 Residuals vs Fitted그래프에서 기울기가 0에 가까움으로 두 모델 모두 정규 분포로 볼 수 있다.

# Normal Q-Q그래프에서도 두 그래프모두 기울기가 1에 가까움으로 정규분포를 따르고 있으며

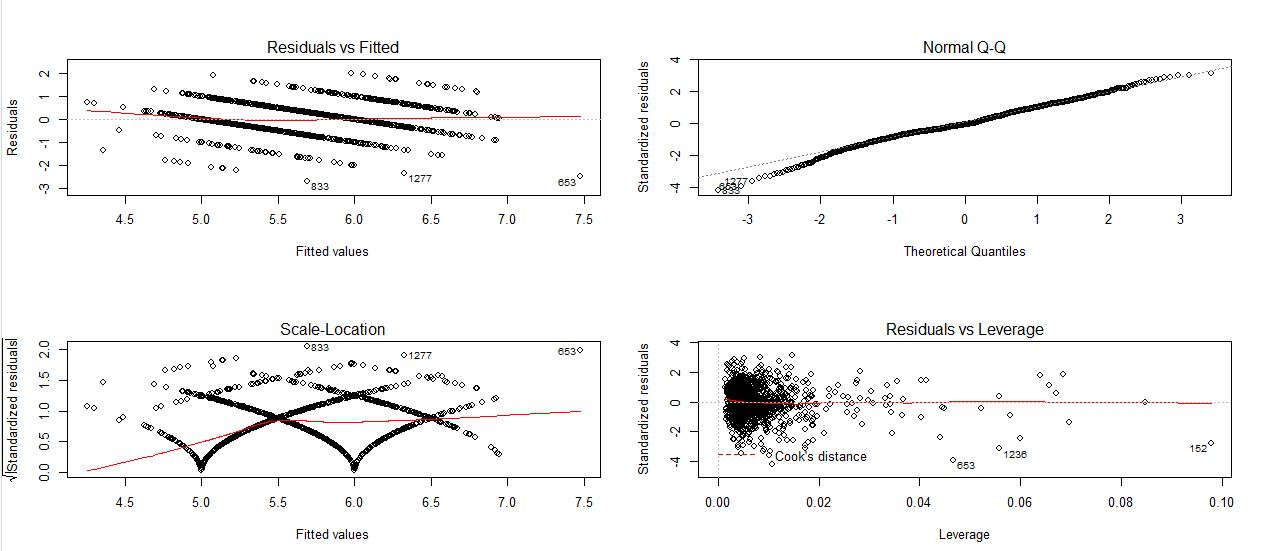
# Scale-Location에서는 두 그래프 모두 5.5전까지는 그래프가 불안정한데 이는 표준화 잔차가 크다는 의미이며 회귀 직선이 5.5까지는

# 잘 들어맞지 않는다는 의미이다. 그래프로 봤을 때는 2interaction term을 추가한 모델이 더 불안정하게 보인다.

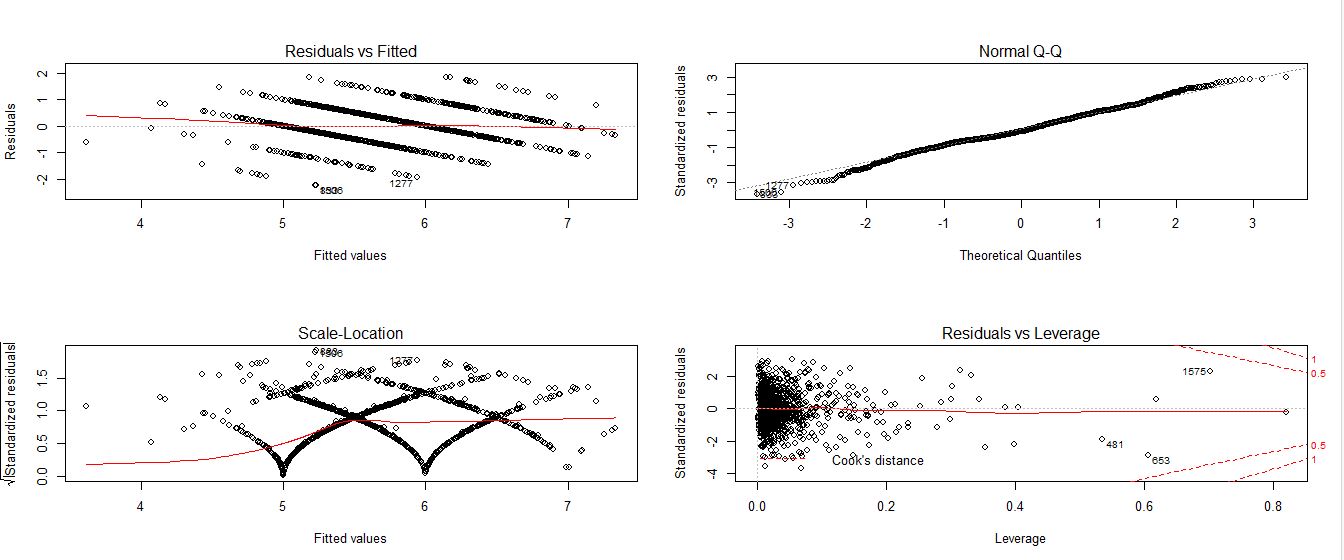
# 마지막 그래프인 Residuals vs Leverage에서는 두 그래프 모두 Cook's distance에 값이 놓여져있는 이상치들이 관측되는데

# 그래프상에서는 2interaction term을 적용하지 않은 첫번째 모델이 더 이상치들이 많은 것으로 보인다.

# 그래프를 통한 비교를 통해서는 두 그래프 모두 유사한 모형을 지녔으며 정규 분포에 가까운 그래프이고 낮은 값의 구역에서는 비교적 정규 분포를 따르지 않고 띄고 있다는 것이다.



<lm redwine plot 그래프>



<2차 interaction red wine plot 그래프>

사용된 함수 4

step(lmRedwine, direction = "backward") #AIC=-1380.8

step(twointerRedwine, direction="backward") # AIC=-1450.5

#step함수의 backward방식 사용하여서 두 모델의 AIC값을 비교해본 결과 2차 interaction term을 추가한 모델이 더 좋게 나왔다.

#plot그래프 anova함수와 step의 AICf를 이용하여서 비교를 수행하였다.

<white wine>

사용된 함수 1

lmWhitewine <- lm(quality~., data=whitewine)

twointerWhitewine <- lm(quality ~ (fixed.acidity+volatile.acidity+citric.acid+ residual.sugar+chlorides+ free.sulfur.dioxide + density + pH + sulphates + alcohol)^2, data = whitewine)

# quality를 반응 변수로 선형 회귀를 수행하고 있다.

# 두 번째 선형회귀에서는 2차 interaction term을 추가하고 있다.

사용된 함수 2

anova(lmWhitewine)

anova(twointerWhitewine)

anova(lmWhitewine,twointerWhitewine)

#anova함수를 이용하여 일원 분산 분석을 수행하고 있다.

#첫 번째 선형회귀에서는 제곱합이 2758.33 제곱합 평균이 0.56으로 나왔으며

#두 번째 2차 interaction을 적용한 선형회귀에서는 제곱합이 2554.89 제곱합 평균이 0.53으로 오차가 더 적게 나왔다.

#그리고 두 모델을 인자로 anova함수를 수행한 결과 두 모델은 p-value가 0.05미만이므로 다른 모델로 취급할 수 있다.

사용된 함수 3

opar <- par(mfrow=c(2,2))

plot(lmWhitewine,which = c(1:3,5))

par(opar)

opar <- par(mfrow=c(2,2))

plot(twointerWhitewine,which = c(1:3,5))

par(opar)

# plot함수를 이용하여서 그래프의 모양새를 관찰해보기도 하였다.

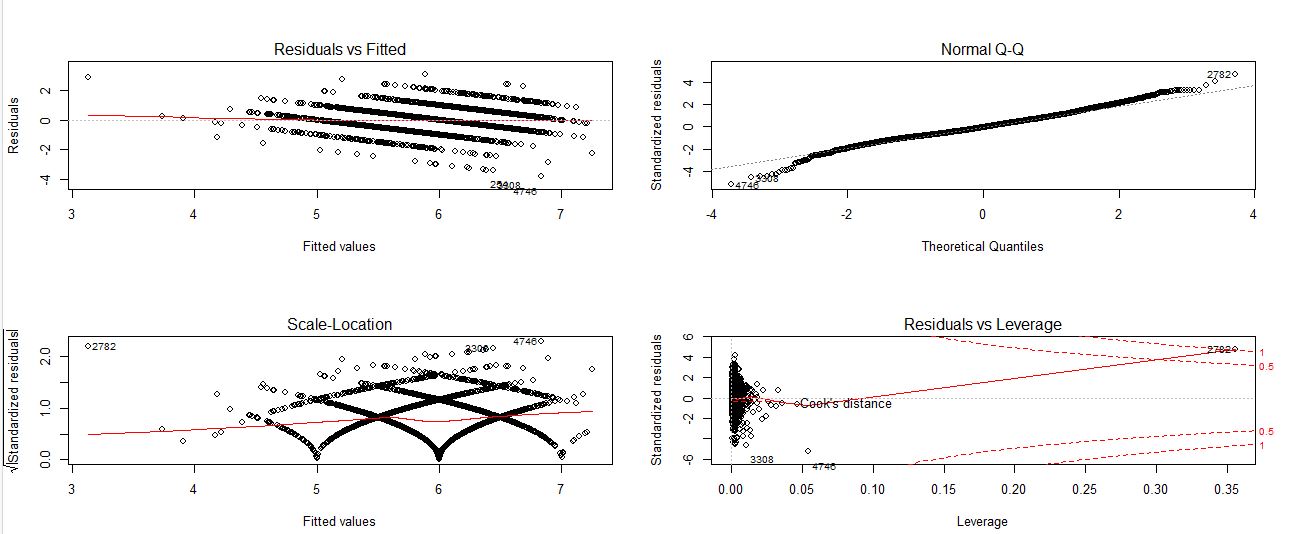
#두 그래프 모두 Residual vs Fitted그래프에서 기울기가 0에 가까움으로 정규분포를 따른다고 보인다.

#두 그래프 모두 Normal Q-Q그래프에서 기울기가 1에 가까우나 낮은 구간에서는 불안정한 모습을 보인다.

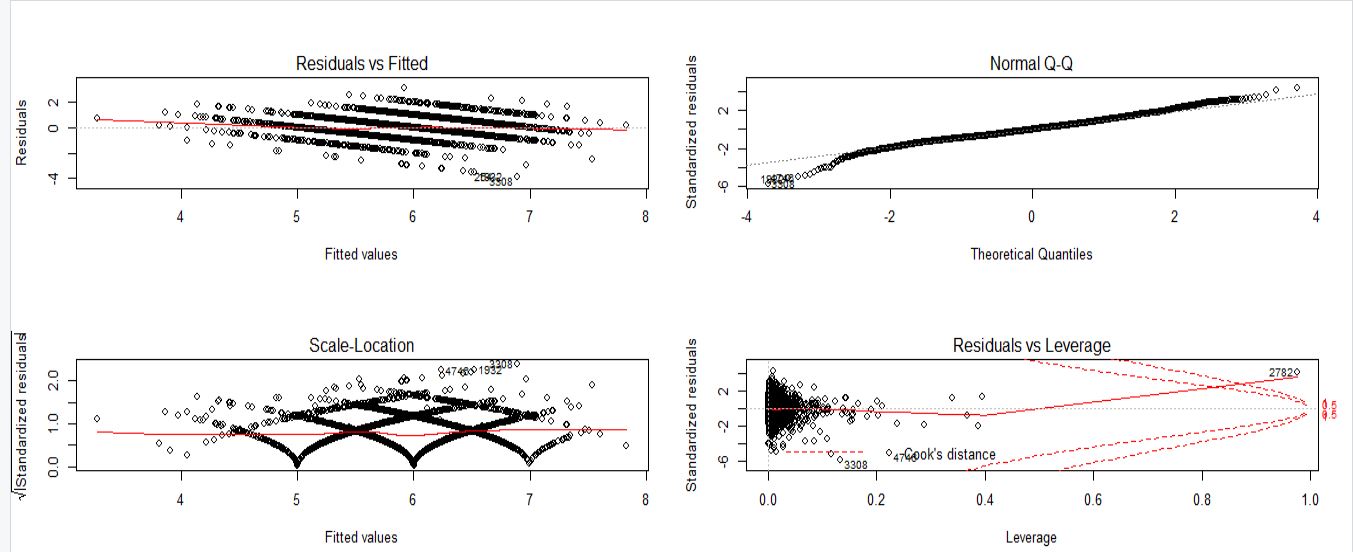
#Scale-Location그래프에서는 2차interaction을 적용한 모델이 그렇지 않은 모델보다 기울기가 0에 가까워서 안정된 모습을 보인다.

#Residual vs Leverage그래프에서는 Cook's distance 점근선이 2차 interaction그래프에서 세세하게 나오는데 이는 두 번째 그래프가 이상치에 더 민감하게 반응하는 것을 보인다.

#plot그래프를 통해 두 그래프가 정규 분포로 유사하며 2차 interaction그래프가 이상치가 적다.



<lm whitewine plot그래프>



<2차 interaction lm whitewine plot 그래프>

사용된 함수 4

step(lmWhitewine, direction="backward") #-2793.6

step(twointerWhitewine, direction="backward") #-3103.9

#step함수를 통해서 AIC를 계산해본 결과 2차 interaction을 적용한 모델이 AIC값이 -3103.9로 훨씬 우수하게 평가되었다.

#plot그래프 anova함수와 step의 AIC 이용하여서 비교를 수행하였다.

<5번 문제>

#2번문제에서 성능이 좋았던 backward모델과 그렇지 않은 모델 비교, 이상치를 제거한 모델과 그렇지 않은 모델 비교

사용한 라이브러리 1

library("car")

#이상치를 발견하는 outlierTest함수를 사용할 수 있다.

#read.csv함수는 생략합니다.

<red wine>

사용한 함수 1

lmRedwine <- lm(quality ~ ., data=redwine)

backwardRedwine <- step(lmRedwine, direction = "backward")

#quality를 반응변수로 선형회귀를 수행하였으며 step함수를 통해서 backward방식을 적용한 선형 회귀 모델 생성

사용한 함수 2

opar <- par(mfrow=c(2,2))

plot(lmRedwine,which = c(1:3,5))

par(opar)

opar <- par(mfrow=c(2,2))

plot(backwardRedwine,which = c(1:3,5))

par(opar)

#plot함수를 이용하여서 두 모델의 그래프를 비교

#두 그래프 각각이 Residuals vs Fitted그래프에서 기울기가 0에 가까우므로 정규분포를 따른다고 볼 수 있다

#두 그래프 각각이 Normal Q-Q그래프에서 기울기가 1인 양상을 보이나 낮은 값에 대해서는 정규분포를 따르지 않는 모양으로 보인다.

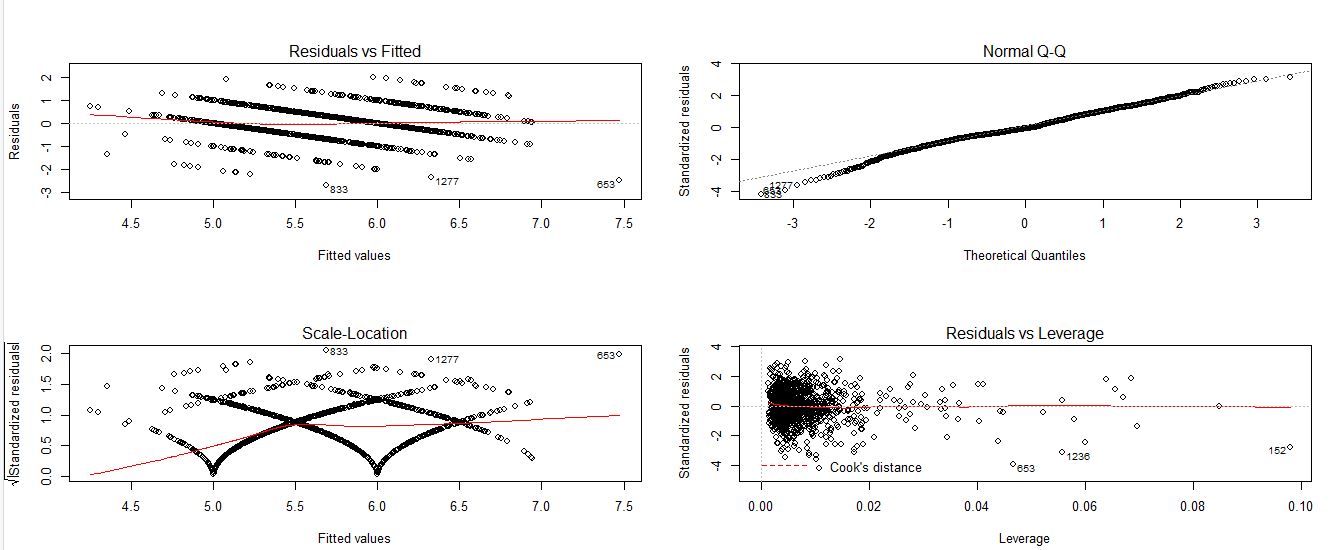
#Scale-Location 그래프에서 두 그래프 모두 5.5까지는 잔차가 증가하는 방향으로 이상치가 크다는 것으로 보이는데

#첫 번째 그래프가 조금이나마 이상치가 적어서 기울기가 상대적으로 완만하게 보인다

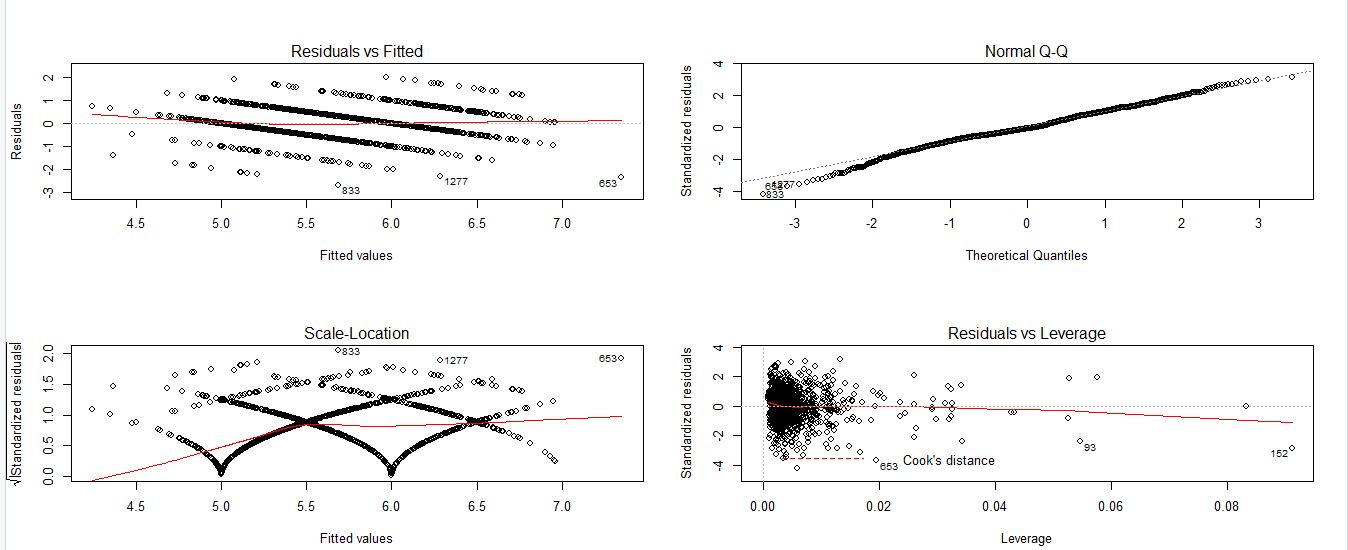
#Residuals vs Leverage에서는 첫 번째 그래프가 기울기가 0에 가까우므로 2차 interaction을 적용한 그래프보다 정규 분포를 잘 따르고 있으며

#이상치가 더 적다. 반면에 2차 interaction을 적용한 그래프에서는 Cook's distance가 더 세세하게 나와서 이상치에 대하여 민감한 것을 볼 수 있으며

#그래프 내에서도 이상치들이 더 많이 분포해 있기 때문에 그래프 기울기가 0에서 살짝 휘어있으며 첫 번째 그래프보다 정규분포를 따르지 않는 모양이다.



<lm redwine>



<backward lm redwine>

사용된 함수 3

outlierTest(lmRedwine)

#outlierTest를 통해서 redwine의 833번째 데이터가 이상치로 판명되었다.

사용된 함수 4

rmRedwine <- redwine[-c(833),] #833번째 데이터 제거

lmrmRedwine <- lm(quality ~ . ,data=rmRedwine)

step(lmRedwine, direction="backward") #AIC가 -1380.8이다

backwardrmRedwine <- step(lmrmRedwine, direction="backward") # AIC가 -1396.4가 되었다.

#단지 행하나를 제거하였는데 AIC가 큰 폭으로 감소하였다.(좋아졌다.)

사용된 함수 5

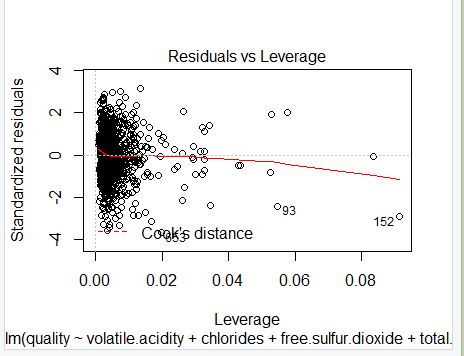
opar <- par(mfrow=c(2,2))

plot(backwardrmRedwine,which = c(5))

par(opar)

par(mfrow=c(1,1))

#단지 행 하나를 제거하였는데 Residuals vs Leverage 그래프가 크게 휘었다.



<이상치를 제거후 휘어진 그래프>

<white wine>

사용한 함수 1

lmWhitewine <- lm(quality ~ ., data=whitewine)

backwardWhitewine <- step(lmWhitewine, direction = "backward")

##quality를 반응변수로 선형회귀를 수행하였으며 step함수를 통해서 backward방식을 적용한 선형 회귀 모델 생성

사용한 함수 2

opar <- par(mfrow=c(2,2))

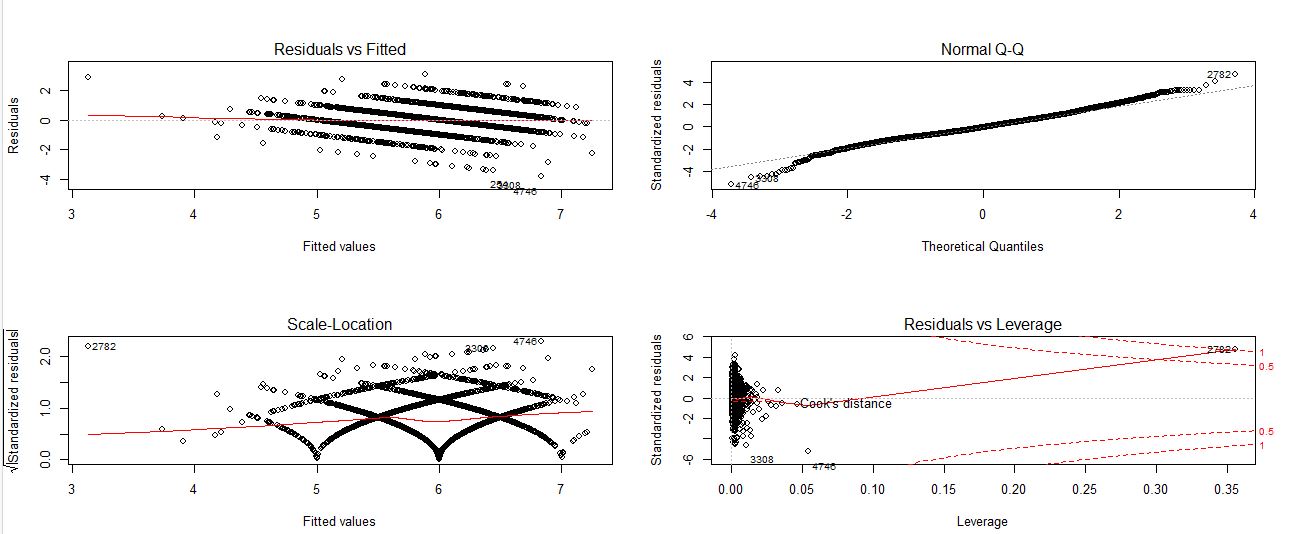
plot(lmWhitewine,which = c(1:3,5))

par(opar)

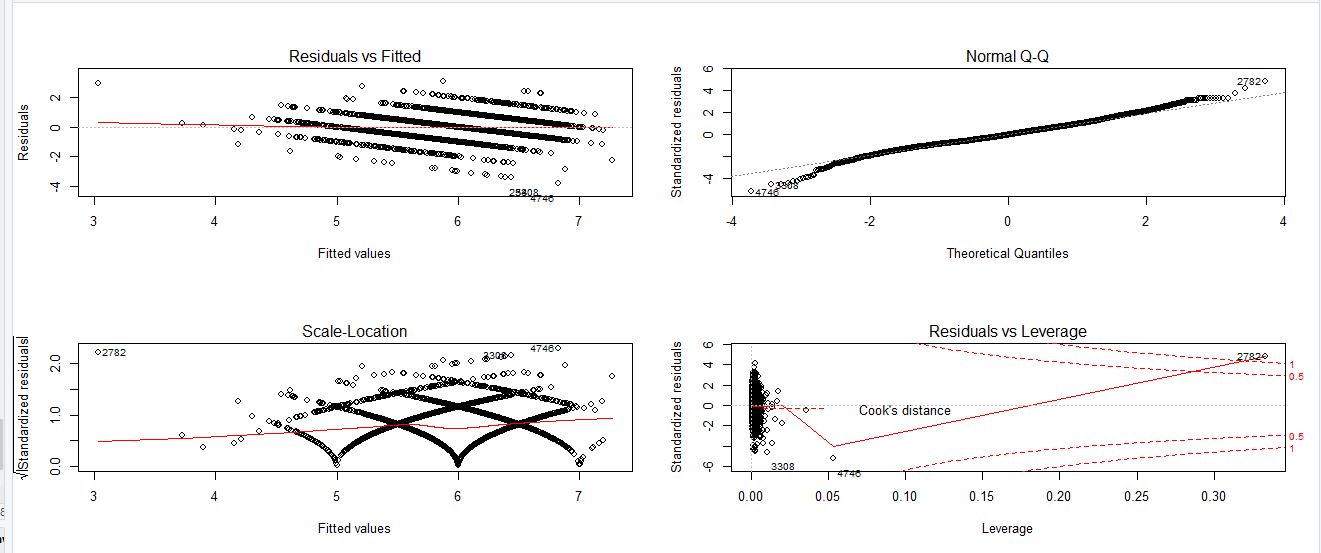
opar <- par(mfrow=c(2,2))

plot(backwardrmRedwine,which = c(5))

par(opar)



<lm white wine plot 그래프>



<backward lm whitewine plot 그래프>

#두 그래프 모드 Residuals vs Fitted 그래프에서는 기울기가 0에 가까우므로 표준 정규분표에 해당한다고 볼 수 있다.

#두 그래프 모드 Normal Q-Q그래프에서 중간에서는 기울기가 1에 가까워 정규분포를 따르지만 –Standarized residuals구간에서는 비교적 기울기에서 벗어나 있어서 정규분포를 따르지 않고 있습니다.

#두 그래프 모두 Scale-Location에서 기울기가 0에 가깝거나 낮은 Fitted values 구간에서는 조금 양의 방향으로 상승하는 경향이 있어서 대부분 정규분포를 따르고 있으며 낮은 Fitted values의 구간에서는 오차가 다소 발생할 여지가 있다.

#Residuals vs Leverage 그래프에서는 Cook’s distance는 비슷한 모양이나 기울기의 그래프의 모양이 다르다는 것을 볼 수 있다. 그래프 상에서는 backward lm whitewine 그래프가 잔차들이 더 분산되어 있는 것으로 보인다.

#두 그래프 모두 대부분 정규분포를 따르고 있으며 낮은 구간에서는 정규분포를 따르지 않고 분산이 상대적으로 심한 경향을 보인다. Cook’s distance가 이제껏 보아왔던 그래프들과 다르게 정중앙에 위치하는 것으로 보아 이상치에 민감하게 반응할 것으로 예상된다.

사용한 함수 3

outlierTest(lmWhitewine)

#outlierTest를 통해서 4786, 2782, 3308, 254, 446번째 구간이 이상치로 판명되었다.

사용한 함수 4

rmWhitewine <- whitewine[-c(4786, 2782, 3308,254,446),]

lmrmWhitewine <- lm(quality ~ . ,data=rmWhitewine)

step(lmWhitewine, direction="backward") #원본backward AIC=-2793.6

backwardrmWhitewine <- step(lmrmWhitewine, direction="backward") #이상치 제거 후 AIC -2874.6

#whitewine에서 4786, 2782, 3308, 254, 446번째 행을 제거한다.

#그 이후에 이상치가 제거된 데이터 셋으로 선형회귀를 다시 수행한다.

#step함수의 backward방식으로 AIC값이 얼마나 좋아졌는지 확인해본다.

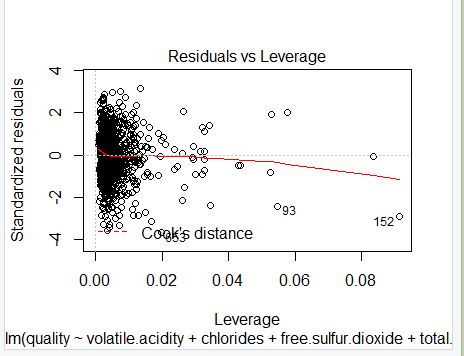
#AIC값이 80정도 더 감소하여 좋아졌다.

사용한 함수 5

par(mfrow=c(1,1))

plot(backwardrmWhitewine,which = c(5))

이상치가 제거된 후 Residuals vs Leverage에서 이상치들과 Cook’s distance가 어떻게 변화하였는지 볼 수 있다. 그래프로 본 결과 그래프에서 휘어졌던 기울기가 상대적으로 더 0을 향하도록 수정되었으며 Cook’s distance가 정중앙에 위치하였던 기존과 다르게 밑으로 위치하여 이상치에 덜 민감하게 반응하며 안정화 되었다는 것을 볼 수 있다.



<이상치 제거 후 안정화 된 Cook’s distance>

<6번 문제> #결측치의 영향력 보기

#read.csv함수는 생략합니다.

사용한 라이브러리 1

library("DMwR")

#결측치를 평균으로 채워주는 함수 DMwR::centralImputation를 사용할 수 있다.

library("dplyr")

#각 컬럼별 평균을 구해주는 mapply(mean,)을 사용한다.

library("foreach")

#구문을 반복하는 foreach를 사용한다.

library("doParallel")

#foreach와 병행하는 %dopar%을 이용한다.

사용한 함수 1

sum(is.na(dataset))

#데이터 셋의 결측치를 확인한다.

사용한 함수 2

lm(f반응변수~., data=dataset)

#선형 회귀를 수행한다.

사용한 함수 3

step(선형회귀모델, direction=”backward”)

#AIC값으로 모델의 품질을 평가할 수 있다.

사용한 함수 4

anova(선형회귀모델)

#Mean sq(전체 분산 평균)과 각각의 칼럼의 분산 평균에 대해서 볼 수 있다.

사용한 함수 5

mapply(mean, dataset)

#데이터셋의 평균을 구할 수 있다.

결과

<red wine>

결측치 0.01

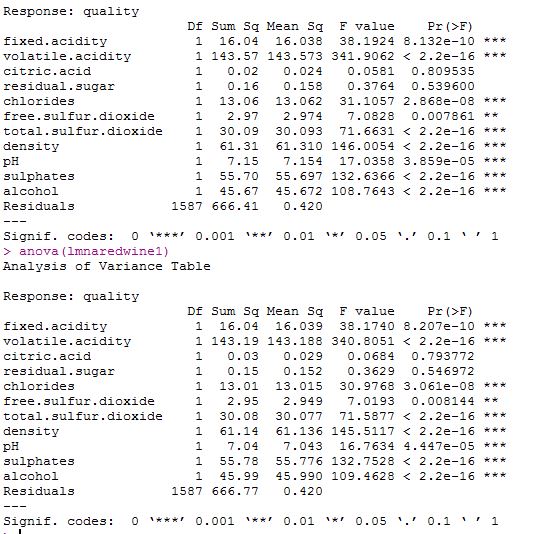
#step으로 분석해봤을 때 AIC가 -1379.9ㄱ 나오며 이것은 기존의 backward AIC값보다 안좋게 나왔다.

#anova로 분석해봤을 때 Mean sq가 0.420나왔으므로 루트를 씌워주면

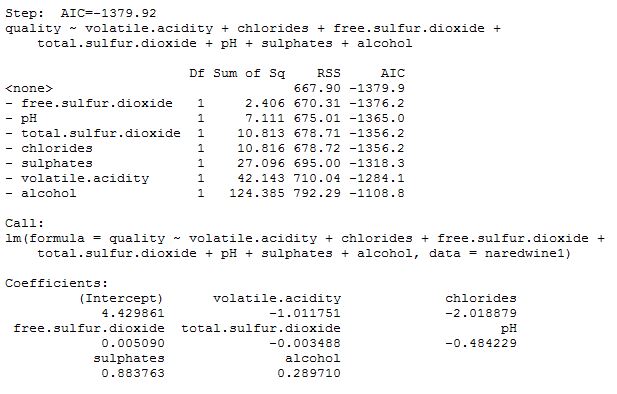
#z <- 0.420^0.5

#표준 편차는 0.648074이 나오며 기존의 그래프와 동일하였다.

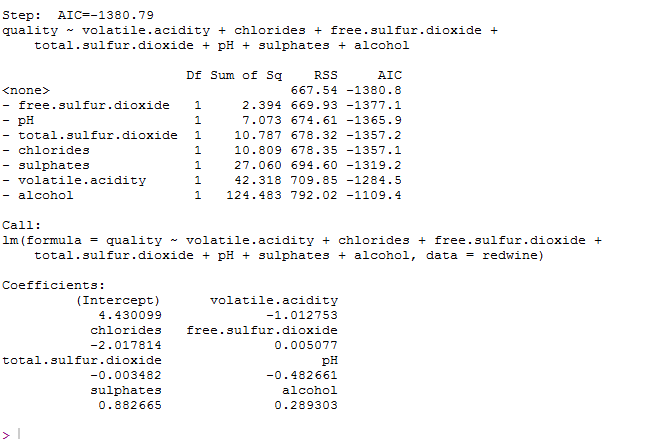
#평균은 mapply(median,)로 비교해봤을 때 동일하였다.(변화가 없다.)



<결측치 중간값으로 대체한 데이터 셋으로 선형 회귀한 lmnaredwine1 anova 일원 분산 분석>



<결측치 중간값 대체한 데이터 셋으로 선형 회귀한 lmnaredwine1 step을 이용한 분석>

<기존의 데이터 셋으로 선형 회귀한 lmredwine을 step을 이용한 분석>

결측치 0.05

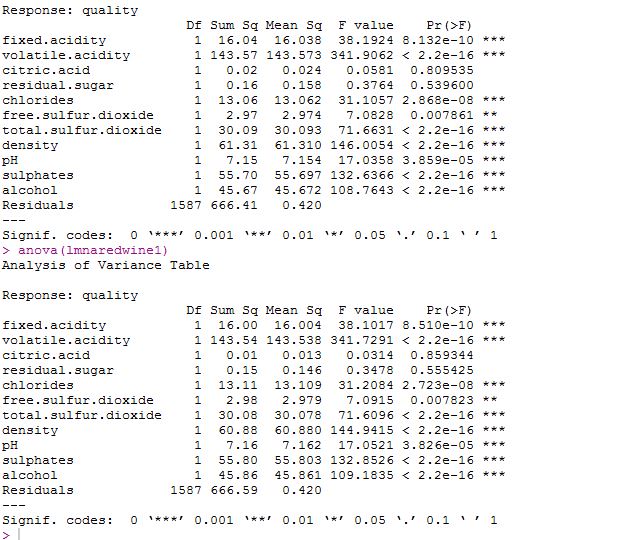
#step으로 분석해봤을 때 AIC가 -1380.5오며 이것은 기존의 backwar AIC값보다 안좋게 나왔다.

#anova로 분석해봤을 때 Mean sq가 0.420나왔으므로 루트를 씌워주면

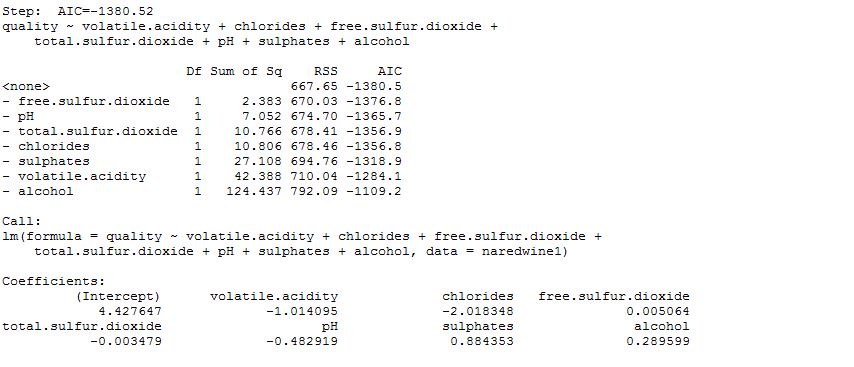
#z <- 0.420^0.5

#표준 편차는 0.648074이 나오며 기존의 그래프와 동일하였다.

#제곱합에서 0.07정도의 차이가 있는데 너무나 작아서 제곱합 평균이 동일하였다.

#평균은 mapply(median,)로 비교해봤을 때 동일하였다.(변화가 없다.)

<결측치 중간값으로 대체한 데이터 셋으로 선형 회귀한 lmnaredwine1 anova 일원 분산 분석>



<결측치 중간값 대체한 데이터 셋으로 선형 회귀한 lmnaredwine1 step을 이용한 분석>

결측치 0.1

#step으로 분석해봤을 때 AIC가 -1380.0이 나오며 이것은 기존의 backwar AIC값보다 안좋게 나왔다.

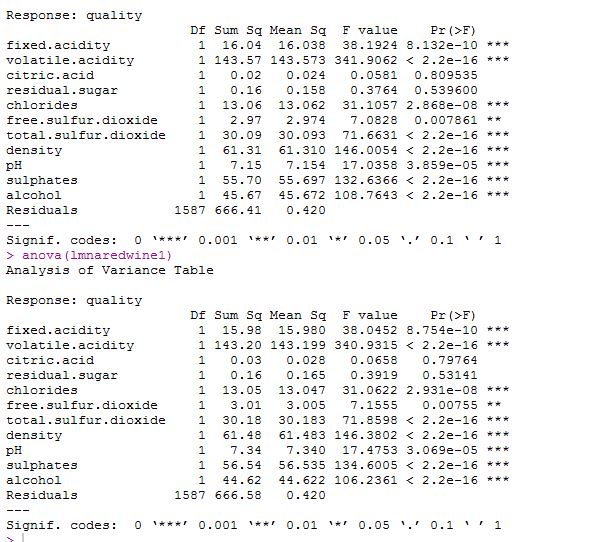
#anova로 분석해봤을 때 Mean sq가 0.420나왔으므로 루트를 씌워주면

#z <- 0.420^0.5

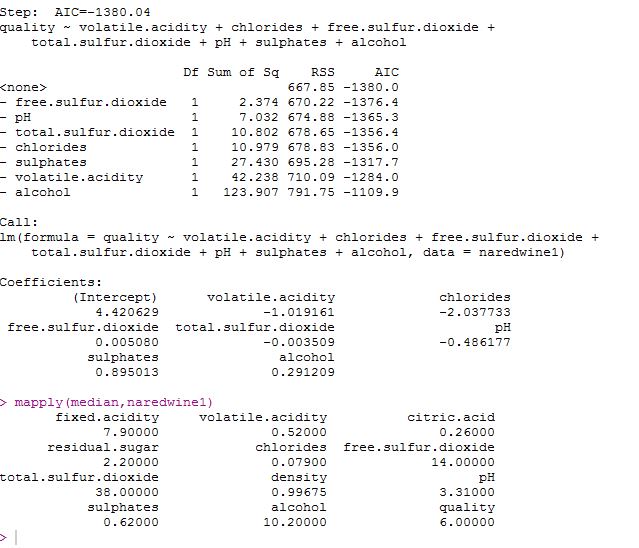
#표준 편차는 0.648074이 나오며 기존의 그래프와 동일하였다.

#제곱합에서 0.17정도의 차이가 있는데 너무나 작아서 제곱합 평균이 동일하였다.

#평균은 mapply(median,)로 비교해봤을 때 동일하였다.(변화가 없다.)



<결측치 중간값으로 대체한 데이터 셋으로 선형 회귀한 lmnaredwine1 anova 일원 분산 분석>



<결측치 중간값 대체한 데이터 셋으로 선형 회귀한 lmnaredwine1 step을 이용한 분석>

<white wine>

결측치 0.01

#step으로 분석해봤을 때 AIC가 -2792.5이 나오며 이것은 기존의 backwar AIC값보다 안좋게 나왔다.

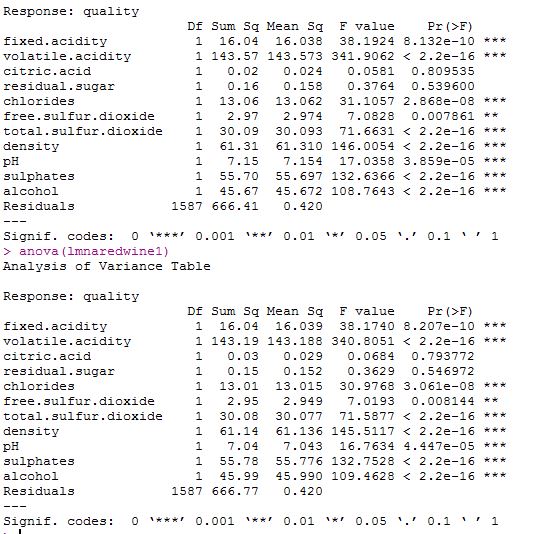
#anova로 분석해봤을 때 Mean sq가 0.56나왔으므로 루트를 씌워주면

#z <- 0.56^0.5

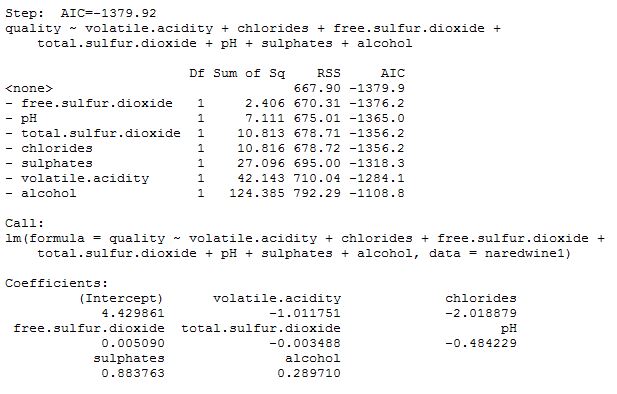
#표준 편차는 0.7483315가 나오며 기존의 그래프와 동일하였다.

#제곱합에서 0.66정도의 차이가 있는데 너무나 작아서 제곱합 평균이 동일하였다.

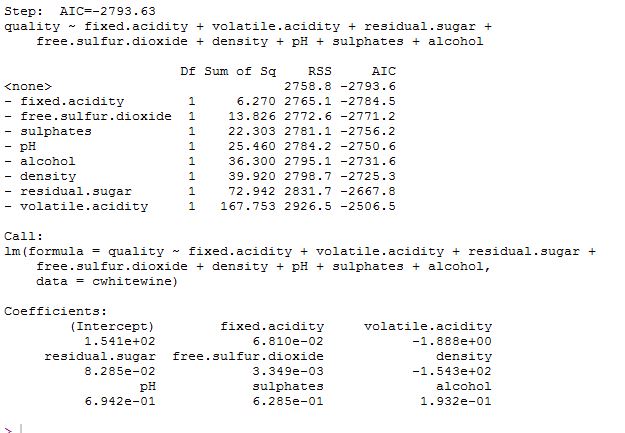
#평균은 mapply(median,)로 비교해봤을 때 동일하였다.(변화가 없다.)



<결측치 0.01 평균 대체한 데이터 셋으로 선형회귀한 lmnaredwine1의 anova(일원 분산 분석)>



<결측치 0.01 평균 대체한 데이터 셋으로 선형회귀한 lmnaredwine1의 step>



<whitewine 그래프의 기본 step>

결측치 0.05

#step으로 분석해봤을 때 AIC가 -2793.32이 나오며 이것은 기존의 backwar AIC값보다 안좋게 나왔다.

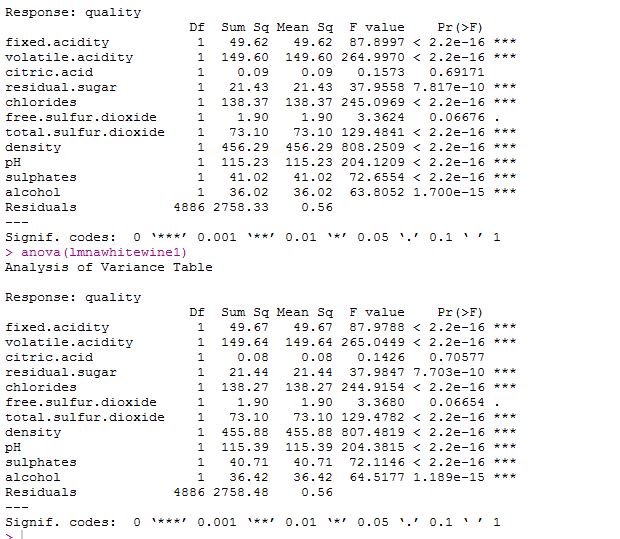
#anova로 분석해봤을 때 Mean sq가 0.56나왔으므로 루트를 씌워주면

#z <- 0.56^0.5

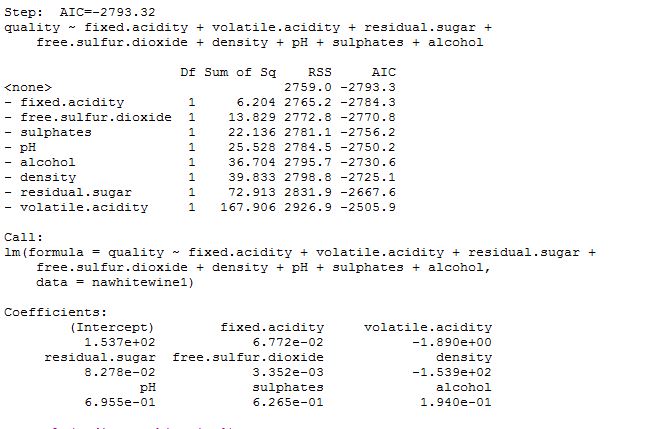
#표준 편차는 0.7483315가 나오며 기존의 그래프와 동일하였다.

#제곱합에서 0.15정도의 차이가 있는데 너무나 작아서 제곱합 평균이 동일하였다.

#평균은 mapply(median,)로 비교해봤을 때 동일하였다.(변화가 없다.)



<결측치 0.05 평균 대체한 데이터 셋으로 선형회귀한 lmnaredwine1의 anova(일원 분산 분석)>



<결측치 0.05 평균 대체한 데이터 셋으로 선형회귀한 lmnaredwine1의 step>

결측치 0.1

#step으로 분석해봤을 때 AIC가 -2790.9가 나오며 이것은 기존의 backwar AIC값보다 안좋게 나왔다.

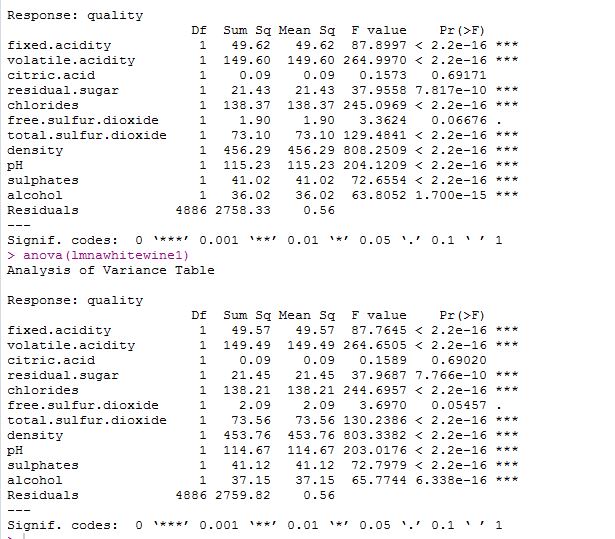
#anova로 분석해봤을 때 Mean sq가 0.56나왔으므로 루트를 씌워주면

#z <- 0.56^0.5

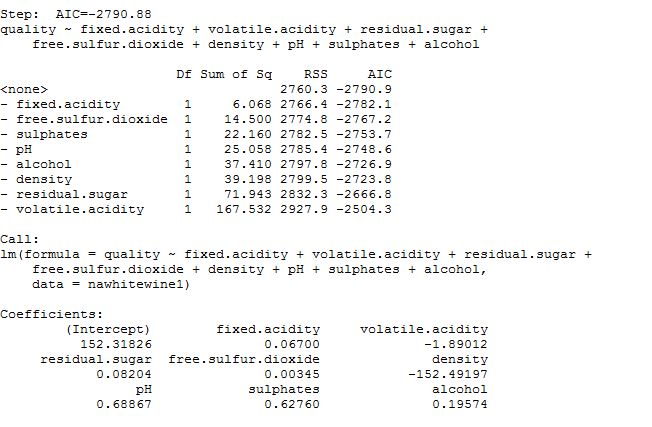
#표준 편차는 0.7483315가 나오며 기존의 그래프와 동일하였다.

#제곱합에서 1.49정도의 차이가 있는데 너무나 작아서 제곱합 평균이 동일하였다.

#평균은 mapply(median,)로 비교해봤을 때 동일하였다.(변화가 없다.)



<결측치 0.1 평균 대체한 데이터 셋으로 선형회귀한 lmnaredwine1의 anova(일원 분산 분석)>



<결측치 0.1 평균 대체한 데이터 셋으로 선형회귀한 lmnaredwine1의 step>

#결론적으로 결측치를 평균으로 대체하면 평균에 가까워지므로 각 컬럼별 평균은 유사하고

#표준 편차도 동일하게 나오며

#AIC값을 살짝 낮게 나오는 경향이 있다.

#그러므로 결측치를 평균으로 대체하는 것은 기존의 원본 데이터 셋에 비추어봤을 때 좋은 기법인 것 같다.

3. 의의 및 평가

본 프로젝트를 수행함으로써 빅데이터 분석 및 시각화를 해주는 R프로그래밍에 대해서 본격적으로 사용해볼 수 있었다. 기존의 수업에서나 평소에 연습해볼 때는 길게 코드를 작성하지는 않았고 예제를 시험해 보는 형식으로 해보았는데 이렇게 한 번 분석해 봄으로써 R프로그래밍에 대해서 더 깊이 이해하는 계기가 되었다. 분석 및 통계에 관한 여러가지 용어들에 대해서도 더 자세히 찾고 배울 수 있었다.

본 과제에 대해서 수행에 대해서 자체적으로 평가해보면 R을 사용함에 있어서 더 좋은 기법이 많이 있지는 않을까하는 생각과 동시에 이 정도의 과제를 책과 수업을 참고하여서 빅데이터 분석 해볼 수 있었다는 것이 큰 의미를 가진다고 생각한다. R외의 내가 아는 다른 어떠한 언어로도 본 과제를 수행하는 데는 엄청난 수고를 했을 것이다. 따라서 본 과제에서 R프로그래밍의 빅데이터 분석 및 시각화를 하는 방법을 사용해서 풀어 볼 수 있었다는 점에서 수업에서 배우고 책에서 본 내용들을 잘 사용해본 것 같다.

출처-R을 이용한 데이터 처리&분석 및 실무/저 서민구/출판사 길벗

R데이터 결측치 정제하기 - <https://rpubs.com/jmhome/R_data_processing>

R분석과 프로그래밍(dplyr패키지 사용) - <http://rfriend.tistory.com/tag/%EB%AC%B4%EC%9E%91%EC%9C%84%20%ED%91%9C%EB%B3%B8%20%EB%8D%B0%EC%9D%B4%ED%84%B0%20%EC%B6%94%EC%B6%9C%20sample_n%28%29%20sample_frac%28%29>

평균과 표준 편차 - <https://m.blog.naver.com/PostView.nhn?blogId=jindog2929&logNo=10162226641&proxyReferer=http%3A%2F%2Fwww.google.co.kr%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3D%26esrc%3Ds%26source%3Dweb%26cd%3D9%26ved%3D0ahUKEwj09NmzxsjbAhVBErwKHdxaCREQFghcMAg%26url%3Dhttp%253A%252F%252Fm.blog.naver.com%252Fjindog2929%252F10162226641%26usg%3DAOvVaw3RA6PYzQk-ZdvP46MzH9i6>