경과보고서_8조

1) 연구가설에 대한 자세한 소개

1. 와인의 quality는 density에 영향을 많이 받을 것이다.

density 와 quality는 모두 연속형 변수이므로 변수 간의 영향 관계를 알 수 있는 회귀분석을 이용하여 와인의 quality는 density에 영향을 많이 받을 것이라는 가설을 검정하였다.

2. 와인의 type별로 와인의 quality에 영향을 주는 변수가 달라질 것이다.

와인의 quality가 각 요소들의 linear model로 표현된다는 가정 하에, '와인의 type별로 와인의 quality에 영향을 주는 변수가 달라질 것이다.'라는 가설을 검정하였다. 먼저 index와 type을 제외한 나머지 변수가 모두 quality에 영향을 주는 linear model을 만들고, backward 변수선택법을 사용해 유의하지 않은 변수를 제거해주는 식으로 설명변수를 선택해 주었다. 이 과정을 red wine과 white wine에 대해서 수행한 후, 두 type간에 영향을 미치는 변수들이 다른지 여부를 살펴보았다.

3. 적절한(너무 낮지도, 높지도 않은) 산도와 당도를 가진 와인이 quality가 더 높을 것이다.

https://blog.naver.com/hitejinrovin/220923836914에 따르면, 와인의 품질은 균형(balance)에 따라 좌우된다. 구체적으로, 너무 많지도 부족하지도 않은 적당한 농도의 당도와 산도, 알코올 농도를 가진 와인이 품질이 좋다고 한다. 특정 와인의 balance를 정량적으로 계산하는 방법은 찾을 수 없었으나, 여기서는 당소, 산도의 값과 그 요소들의 평균까지의 거리가 balance를 나타낼 것이라는 가정 하에 blance라는 변수를 직접 만들어 이를 추가한 회귀분석을 진행하였다.

2) 가설에 대한 중간분석 결과

Basic Setting

```
df = read.csv("data/train.csv", header = TRUE)
df_red <- df %>% filter(type=='red')
df_white <- df %>% filter(type=="white")
```

1번 가설

```
lm<-lm(quality~density, data = df)
summary(lm)
```

```
##
## Call:
## lm(formula = quality ~ density, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1284 -0.6245  0.0171  0.4777  4.0171
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   91.939      3.697   24.87   <2e-16 ***
## density      -86.581      3.716  -23.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8303 on 5495 degrees of freedom
## Multiple R-squared:  0.0899, Adjusted R-squared:  0.08973
## F-statistic: 542.8 on 1 and 5495 DF,  p-value: < 2.2e-16
```

위의 결과를 보았을 때 우선 p 값이 0.05보다 작으므로 유의수준을 5%로 설정했다면 통계적으로 유의미한 결과가 나왔다고 볼 수 있다. R 제곱 값이 약 0.08이므로 density가 quality의 8%밖에 설명하지 못한다는 것을 알 수 있다. 만약 density와 quality가 선형관계를 띤다면 음의 상관관계를 가진다. R 제곱 값으로 미루어 보아 density는 quality에 큰 영향을 주지 않는다는 것을 알 수 있다.

2번 가설

```
df_red_lm <- df_red[, c(2:13)]  # Exclude 'index', 'type' column
lm_red_ori <- lm(quality~ ., data = df_red_lm)
coef(step(lm_red_ori, direction = "backward", trace = 0))
```

```
##          (Intercept)     volatile.acidity          citric.acid
##          4.620634786         -1.102465872         -0.226143145
##            chlorides  free.sulfur.dioxide total.sulfur.dioxide
##         -1.811982373          0.004885911         -0.003980351
##                   pH            sulphates              alcohol
##         -0.494730150          0.929490094          0.282697525
```

```
df_white_lm <- df_white[, c(2:13)]
lm_white_ori <- lm(quality~ ., data = df_white_lm)
coef(step(lm_white_ori, direction = "backward", trace = 0))
```

```
##         (Intercept)       fixed.acidity    volatile.acidity      residual.sugar
##        1.517077e+02        6.747169e-02       -1.919229e+00        8.379455e-02
## free.sulfur.dioxide             density                  pH           sulphates
##        3.415588e-03       -1.518877e+02        6.770859e-01        6.827911e-01
##             alcohol
##        1.966551e-01
```

분석 결과 red와 white 모두에서 유의하게 나온 변수는 volatile.acidity, free.sulfur.dioxide, pH, sulphates, alcohol이었고 red에만 유의하게 나온 변수는 citric.acid, chlorides, total.sulfur.dioxide, white에만 유의하게 나온 변수는 fixed.acidity, residual.sugar, density가 있다.

결과를 분석해 본다면, 우선 white에만 residual.sugar가 품질에 영향을 미친다는 것은 화이트 와인을 마시는 소비자가 상대적으로 당도를 더 중요하게 고려한다는 것으로 설명할 수 있을 것 같다. https://www.asiae.co.kr/article/2018060806513888770 에 의하면, 화이트와인은 레드와인과 달리 껍질이나 씨에서 색소나 탄닌을 수출하는 과정이 없다. 따라서 화이트와인에 대한 소비자의 요구 또한 신선하고, 너무 떫거나 쓰지 않는 것으로, 가볍고 산뜻한 와인을 좋아한다고 한다.

3번 가설

```
df_1 <- df %>% mutate(balance = sqrt((residual.sugar - mean(residual.sugar))^2 + (citric.acid - mean(ci
df_1 <- df_1 %>% mutate(residual.sugar = scale(residual.sugar, center = TRUE, scale = FALSE), citric.ac

lm_1 <- lm(quality~ fixed.acidity + volatile.acidity + residual.sugar + free.sulfur.dioxide + density +
summary(lm_1)
```

```
##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol +
##     balance, data = df_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5276 -0.4707 -0.0312  0.4570  3.0077
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.801e+01  1.227e+01   3.097  0.00197 **
## fixed.acidity        5.768e-02  1.622e-02   3.556  0.00038 ***
## volatile.acidity    -1.247e+00  7.755e-02 -16.078  < 2e-16 ***
## residual.sugar       3.229e-02  5.484e-03   5.888 4.14e-09 ***
## free.sulfur.dioxide  2.026e-03  6.732e-04   3.010  0.00263 **
## density             -3.752e+01  1.251e+01  -3.000  0.00271 **
## pH                   4.611e-01  9.580e-02   4.813 1.53e-06 ***
## sulphates            7.678e-01  8.093e-02   9.487  < 2e-16 ***
## alcohol              3.014e-01  1.762e-02  17.103  < 2e-16 ***
## balance              3.082e-03  3.955e-03   0.779  0.43589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7392 on 5487 degrees of freedom
## Multiple R-squared:  0.2798, Adjusted R-squared:  0.2786
## F-statistic: 236.9 on 9 and 5487 DF,  p-value: < 2.2e-16
```

summary 결과 balance 값이 유의하지 않다고 나왔다. 사실 예상과는 매우 다른 결과가 나왔는데, balance를 sugar_balance, acid_balance, alcohol_balance 등으로 나누어서 검정해 보아도 모두 유의하지 않은 결과가 나타났다. 정확한 원인은 알 수 없으나 나름대로 추청한다면 다음과 같은 이유가 있을 수 있겠다.

사실 balance는 저런 간단한 식으로 구해지는 것이 아닐 수 있다. 출처로 제시한 사이트를 참고하면, 와인의 balance는 단순히 당도, 산도, 알코올 말고도 탄닌, 과일 풍미가 농축된 정도 같은 더 다양한 요인들에게 영향을 받는다. 이러한 점이 가설검정을 할 때 사용된 balance를 구하는 데에는 고려되지 않았고, 또한 각 요소별로 가중치가 다르다거나, 더 복잡한 식으로 balance가 결정될 가능성 역시 존재한다. 이러한 점에서 우리가 구한 balance라는 수치가 와인의 quality를 결정하는 데 적절한 요인이 아닐 수 있다.

3) 향후 분석 계획

위의 가설과 별개로, '과연 linear model이 와인의 quality를 평가하는 데(accuracy라는 기준을 가지고) 적절한 모형인가?' 라는 의문이 들었다. 이를 살펴보기 위해 linear model로 와인의 quality를 추정한 뒤 예측 정확도를 구해보는 작업을 해 본 결과, model의 전체적인 정확도가 50%~60% 사이로 그렇게 높지 않았다.

이는 와인의 quality가 linear model을 따르지 않을 수 있으며, OLS를 최소화하는 방식으로 제작된 linear model이 Acuuracy를 최대로 높이는 문제의 의도와 차이가 있어서 발생하는 문제라고 생각한다. 향후 최종 보고서 때는 model의 Accuracy를 더욱 개선시키기 위해 비선형적 모델(랜덤포레스트)를 적용한 뒤 linear model과의 성능 차이를 분석할 계획이다.