CR_Portfolio_4

Seunghun Lee

2020-11-16 09:38:45

Contents

Research Question 1
Descriptive statistics 1
Hypotheses 3
Critical test statistic 3
Sample test statistic results 3
Conclusion 5

Research Question

- A researcher was interested in investigating whether the pH level and alcohol level were significant predictors of the quality of Portuguese red wine. Both predictors were mean-centered. Using the sample of data for 1599 cases found in `winequality-red.csv`, test the researcher's hypothesis using α of 0.05.

```
# https://archive.ics.uci.edu/ml/datasets/Wine+Quality
wine <- read.csv("./data/winequality-red.csv", sep = ";") %>%
  select(quality, pH, alcohol)
# Checking the missing value
which(is.na(wine) == TRUE)
```

```
## integer(0)
```

Descriptive statistics

```
wine %>% str()
```

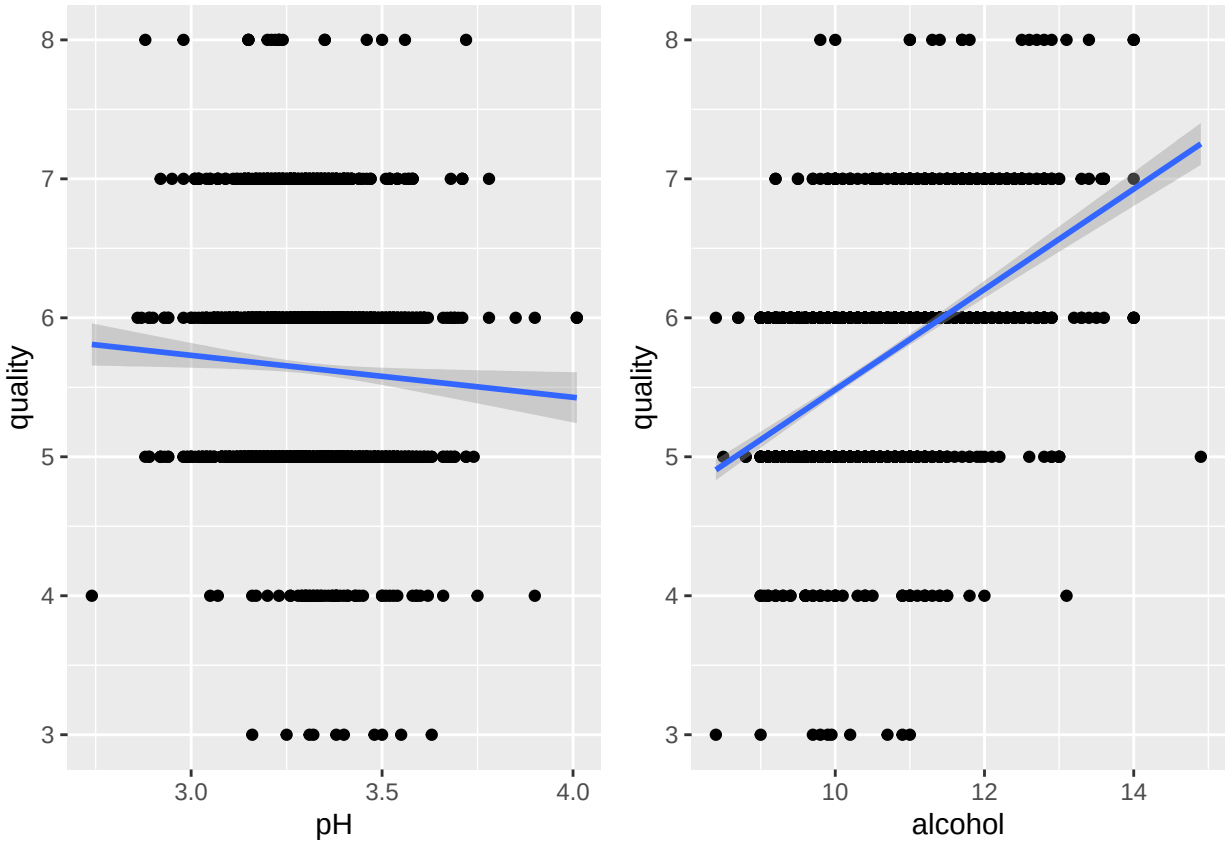
```
## 'data.frame':    1599 obs. of  3 variables:
##  $ quality: int  5 5 5 6 5 5 5 7 7 5 ...
##  $ pH     : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
##  $ alcohol: num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
```

```
wine %>% summary()
```

```
##     quality            pH           alcohol
##  Min.   :3.000   Min.   :2.740   Min.   : 8.40
##  1st Qu.:5.000   1st Qu.:3.210   1st Qu.: 9.50
##  Median :6.000   Median :3.310   Median :10.20
##  Mean   :5.636   Mean   :3.311   Mean   :10.42
##  3rd Qu.:6.000   3rd Qu.:3.400   3rd Qu.:11.10
##  Max.   :8.000   Max.   :4.010   Max.   :14.90
```

```
p1 <- wine %>%
  ggplot(aes(x = pH, y = quality)) +
  geom_point() +
  geom_smooth(method = "lm")
p2 <- wine %>%
  ggplot(aes(x = alcohol, y = quality))+
  geom_point() +
  geom_smooth(method = "lm")
grid.arrange(p1, p2,
             ncol = 2, nrow = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



Hypotheses

- H_{1A}: The predictors - pH level and alcohol level - explain variability in the quality of red wine.
- H_{0A}: The predictors - pH level and alcohol level - don't explain variability in the quality of red wine.

- H_{1B}: The pH level (while controlling for the alcohol level) is a negative predictor of the quality of red wine.
- H_{0B}: The pH level (while controlling for the alcohol level) is not a negative predictor of the quality of red wine.

- H_{1C}: The alcohol level (while controlling for the pH level) is a positive predictor of the quality of red wine.
- H_{0C}: The alcohol level (while controlling for the pH level) is not a positive predictor of the quality of red wine.

- H_{1D}: The quality of red wine with the mean level of pH and the mean level of alcohol is greater than zero.
- H_{0D}: The quality of red wine with the mean level of pH and the mean level of alcohol is not greater than zero.

Critical test statistic

```
# Critical t
qt(0.95, df = 1596)
```

```
## [1] 1.645809
```

```
qt(0.05, df = 1596)
```

```
## [1] -1.645809
```

```
# Critical F
qf(0.95, df1 = 2, df2 = 1596)
```

```
## [1] 3.001362
```

- **The critical F statistic = +3.001**
- **The critical t statistic = +-1.646**
- For testing H_{0A}: $\alpha = 0.05$, $df_{reg} = 2$, $dfError = 1596$; critical F**(2,1596) = 3.001**
- For testing H_{0B}: $\alpha = 0.05$, $df = 1596$; critical t**(1596) = -1.645**
- For testing H_{0C} and H_{0D}: $\alpha = 0.05$, $df = 1596$; critical t**(1596) = 1.645**

Sample test statistic results

```r
# Mean-centered data
wine$pH.ctrd <- scale(wine$pH,
                      center = TRUE,
                      scale = FALSE)
wine$alcohol.ctrd <- scale(wine$alcohol,
                           center = TRUE,
                           scale = FALSE)
# Models
mod1 <- lm(quality ~ pH.ctrd + alcohol.ctrd, data = wine)
mod2 <- lm(scale(quality) ~  scale(pH.ctrd) + scale(alcohol.ctrd), data = wine)
summary(mod1)
```

```
##
## Call:
## lm(formula = quality ~ pH.ctrd + alcohol.ctrd, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7153 -0.4066 -0.1105  0.5076  2.4584
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.63602    0.01748 322.480  < 2e-16 ***
## pH.ctrd      -0.85011    0.11571  -7.347 3.23e-13 ***
## alcohol.ctrd  0.38617    0.01676  23.036  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6989 on 1596 degrees of freedom
## Multiple R-squared:  0.252,  Adjusted R-squared:  0.2511
## F-statistic: 268.9 on 2 and 1596 DF,  p-value: < 2.2e-16
```

```r
summary(mod2)
```

```
##
## Call:
## lm(formula = scale(quality) ~ scale(pH.ctrd) + scale(alcohol.ctrd),
##     data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3623 -0.5034 -0.1369  0.6285  3.0442
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.399e-15  2.164e-02   0.000        1
## scale(pH.ctrd)      -1.625e-01  2.212e-02  -7.347 3.23e-13 ***
## scale(alcohol.ctrd)  5.096e-01  2.212e-02  23.036  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8654 on 1596 degrees of freedom
```

```
## Multiple R-squared:  0.252,  Adjusted R-squared:  0.2511
## F-statistic: 268.9 on 2 and 1596 DF,  p-value: < 2.2e-16
```

Conclusion

- **For testing** H_{0A}: Reject H_{0A} and conclude that the predictors - pH level and alcohol level - explain a significant amount of variability in the quality of red wine.
- [$R^2 = 0.252$, $R^2_{adj} = 0.2511$, $F(2, 1596) = 268.9$, $p < 0.05$].
- Together, the pH level and alcohol level explain about 25.1% of the variability in the quality of red wine.

- **For testing** H_{0B}: Reject H_{0B} and infer that the pH level (while controlling for the alcohol level) is a significant negative predictor of the quality of red wine.
- [B = -0.850, β = -0.163, $t(1596)$ = -7.347, $p < 0.05$].
- The results indicate that, controlling for the alcohol level, the higher pH level, the lower wine quality will be. Specifically, for two wines with the same alcohol level, for the wine with one higher pH level, that wine is predicted to be 0.850 points lower on the quality.

- **For testing** H_{0C}: Reject H_{0C} and infer that the alcohol level (while controlling for the pH level) is a significant positive predictor of the quality of red wine.
- [$B = 0.386$, $\beta = 0.510$, $t(1596) = 23.036$, $p < 0.05$].
- The results indicate that, controlling for the pH level, the higher alcohol level, the higher wine quality will be. Specifically, for two wines with the same pH level, for the wine with one higher alcohol level, that wine is predicted to be 0.386 points higher on the quality.

- **For testing** H_{0D}: Reject H_{0D} and infer that the quality of red wine with the average of pH level and average level of alcohol level is significantly greater than zero.
- [$B = 5.636$, $t(1596) = 322.48$, $p < 0.05$].
- The intercept is interpreted as the predicted quality (estimated to be 5.636) for a wine at the mean on pH level and at the mean on alcohol level.