Mini-Project 1

STAT-340 Applied Regression Methods

Yerim Oh

Details

Due Date

This assignment is due at 11:59 PM on the 20th of September.

Grading

The problems will be graded for correctness. An emphasis will be placed on full explanations of your thought process. You usually won't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind your decisions is more important than making the "correct" decision.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Load Packages

The following R code loads packages needed in this assignment.

```
library(readxl)
library(dplyr)
library(ggplot2)
library(GGally)
```

Problem 1: White vinho verde from Portugal

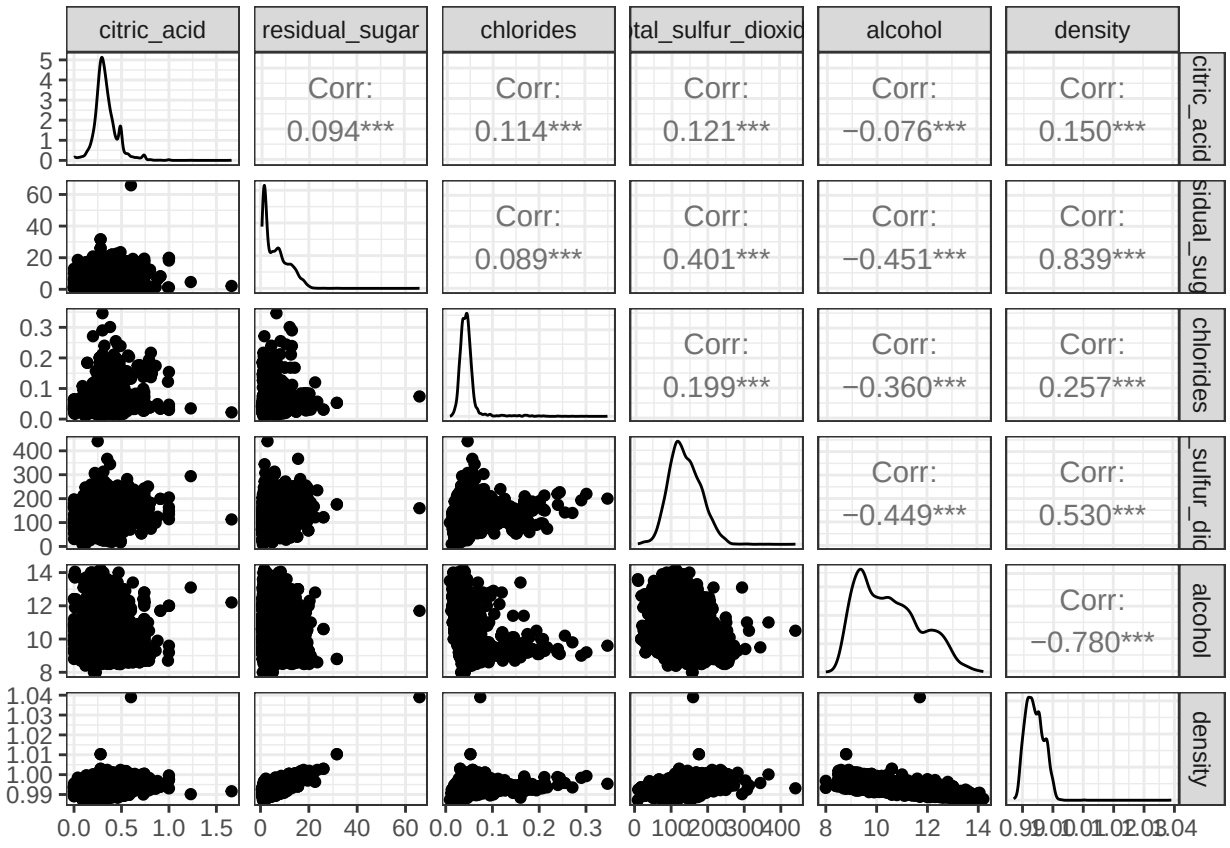
"The wine industry is investing in new technologies for both wine making and selling processes. Wine certification and quality assessment are key elements within this context. Certification prevents the illegal adulteration of wines (to safeguard human health) and assures quality for the wine market. Wine certification is generally assessed by physicochemical and sensory tests. Physicochemical laboratory tests routinely used to characterize wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts."

In this mini-project, we are interested in learning about the variables that may be related to the wine `density`. We will use the dataset `wine` and focus on the following variables:

- `density` the wine density measured in grams per milliliter (g/mL).
- `citric_acidity`: amount of citric acid per liter (g/L)
- `residual_sugar`: amount of residual sugar per liter (g/L)
- `chlorides`: amount of sodium chloride per liter (g/L)
- `total_sulfur_dioxide`: total amount of sulfur dioxide per liter (mg/L)
- `alcohol`: percentage of alcohol in the wine (% vol)

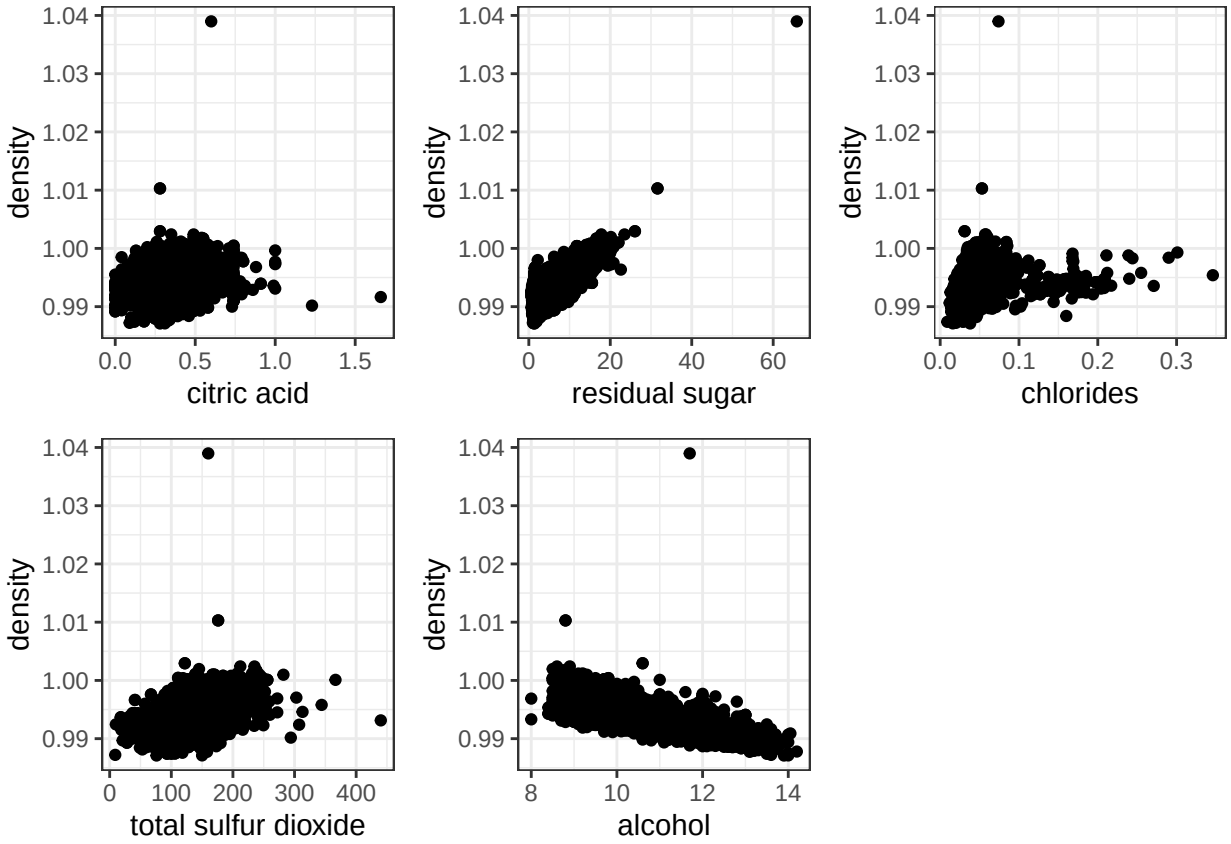
(a) Make a pair plot with all the variables. Based on this plot, which variable(s) appear(s) to be related to the wine density in g/mL (`density`)?

```
ggpairs(wine %>%
          select(citric_acid,
                 residual_sugar,
                 chlorides,
                 total_sulfur_dioxide,
                 alcohol,
                 density)) +
  theme_bw()
```




```
p1 <- ggplot(wine, aes(x = citric_acid, y = density)) +
  geom_point() +
  xlab("citric acid") +
  theme_bw()
p2 <- ggplot(wine, aes(x = residual_sugar, y = density)) +
  geom_point() +
  xlab("residual sugar") +
  theme_bw()
p3 <- ggplot(wine, aes(x = chlorides, y = density)) +
  geom_point() +
  xlab("chlorides") +
  theme_bw()
p4 <- ggplot(wine, aes(x = total_sulfur_dioxide, y = density)) +
  geom_point() +
  xlab("total sulfur dioxide") +
  theme_bw()
p5 <- ggplot(wine, aes(x = alcohol, y = density)) +
  geom_point() +
  xlab("alcohol") +
  theme_bw()

library(gridExtra) # allows to show multiplot plots on the same page
grid.arrange(p1, p2, p3, p4, p5, ncol = 3)
```



Among the variables, `residual_sugar` and `total_sulfur_dioxide` appears to have positive linear correlations, and `alcohol` have a strong negative linear relationship. However, other variables might also have a statistically significant relationship.

(b) Fit a linear regression model that includes all the variables. Describe the interpretation of the estimated slope for the variable `residual_sugar`.

```
model_fit <- lm(density ~ citric_acid+residual_sugar+chlorides+total_sulfur_dioxide+alcohol, wine)
summary(model_fit)
```

```
##
## Call:
## lm(formula = density ~ citric_acid + residual_sugar + chlorides +
##     total_sulfur_dioxide + alcohol, data = wine)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0021263 -0.0005689 -0.0001093  0.0004314  0.0251132
##
## Coefficients:
##                         Estimate Std. Error  t value Pr(>|t|)
## (Intercept)            1.003e+00  1.719e-04 5832.375  < 2e-16 ***
## citric_acid            1.177e-03  1.052e-04   11.191  < 2e-16 ***
## residual_sugar         3.488e-04  2.884e-06  120.939  < 2e-16 ***
## chlorides              2.038e-03  6.229e-04    3.271  0.00108 **
```

```
## total_sulfur_dioxide  5.040e-06  3.434e-07   14.676  < 2e-16 ***
## alcohol              -1.148e-03  1.280e-05  -89.742  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0008789 on 4892 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9137
## F-statistic: 1.036e+04 on 5 and 4892 DF,  p-value: < 2.2e-16
```

Estimated slope for `residual_sugar` = 3.488e-04: For every additional amount of residual sugar per liter, the estimated wine density measured in grams per milliliter increases by 3.488e-04 **on average** while the other variables are constant.

(c) State the model you have fit in matrix form. Make sure you clearly specify the dimension of the matrices and vectors in the model.

$$\begin{bmatrix} 1.0010 \\ 0.9940 \\ \vdots \\ y_{4898} \end{bmatrix} = \begin{bmatrix} 1 & 0.36 & 20.70 & 0.045 & 170.0 & 8.8 \\ 1 & 0.34 & 1.60 & 0.049 & 132.0 & 9.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{4898\cdot 1} & x_{4898\cdot 2} & x_{4898\cdot 3} & x_{4898\cdot 4} & x_{4898\cdot 5} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{4898} \end{bmatrix}$$

- $\boldsymbol{Y}$: 4898 x 1
  - `density`
- $\boldsymbol{X}$: 4898 x 6
  - `citric_acid`, `residual_sugar`, `chlorides`, `total_sulfur_dioxide`, `alcohol`
- β: 6 x 1
- ε: 4898 x 1

(d) Extract the design matrix from your model fit and use it to find the fitted values for the regression. Make a scatter plot of the residuals versus the fitted value, both calculated "by hand".

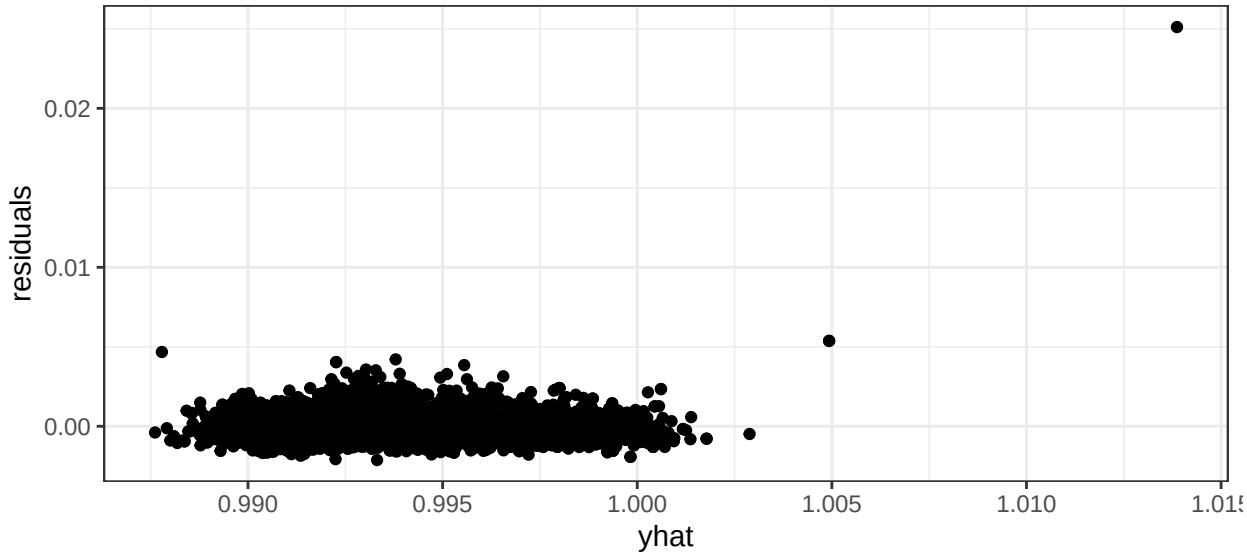
```r
# extract the design matrix from the model fit
X <- model.matrix(model_fit)
y <- matrix(wine$density)
beta_hat <- solve( t(X) %*% X) %*% t(X) %*% y # beta_hat = (X'X)^-1 X'y
beta_hat
```

```
##                              [,1]
## (Intercept)           1.002688e+00
## citric_acid           1.176889e-03
## residual_sugar        3.487977e-04
## chlorides             2.037504e-03
## total_sulfur_dioxide  5.039731e-06
## alcohol              -1.148310e-03
```

```r
# add the fitted values and residuals in the dataframe
wine <- wine %>%
  mutate(yhat = beta_hat[1] + beta_hat[2] * wine$citric_acid + beta_hat[3] * wine$residual_sugar + beta
```

```
         ei_manual = wine$density - yhat)

# scatter plot: fitted value vs. residuals
ggplot(wine, aes(x = yhat, y = ei_manual)) +
  geom_point() +
  ylab("residuals") +
  theme_bw()
```



(e) Find and interpret a 95% confidence interval for the effect of the alcohol content on the wine density in g/mL (`density`).

```
confint(model_fit)
```

```
##                              2.5 %       97.5 %
## (Intercept)           1.002351e+00  1.003025e+00
## citric_acid           9.707285e-04  1.383050e-03
## residual_sugar        3.431436e-04  3.544518e-04
## chlorides             8.162637e-04  3.258745e-03
## total_sulfur_dioxide  4.366536e-06  5.712926e-06
## alcohol              -1.173396e-03 -1.123225e-03
```

We are 95% confident that one additional percentage of alcohol in the wine will decrease the average wine density measured in grams per milliliter by an amount between $-1.173 \cdot 10^{-3}$ and $-1.123 \cdot 10^{-3}$, while the other variables are constant. By 95% confident, we mean that for 95% of the samples, a confidence interval calculated in this way will include the true effect of `alcohol` on `density`.

(f) Conduct a hypothesis test of the claim that there is no association between any of the variables included in the model and the wine density in g/mL (`density`). Make sure you state: (i) The null and alternative hypotheses, (ii) the full and reduced model, if appropriate, (iii) The value of the test statistic and p-value, (iv) your conclusion in a complete sentence. Remark. You don't need to "reject" or "fail to reject" the null hypothesis; instead, interpret what the p-value for the test means in terms of strength of evidence against the null hypothesis.

Hypothesis

- H_0: $\beta_1 = \cdots = \beta_5 = 0$; There is no relationship between any of the variables included in the model and the wine density
- H_A: At least one of $\beta_1, \ldots, \beta_5$ is not equal to 0

Model

- full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$
- reduced model: $y = \beta_0 + \varepsilon$

```
m2_fit <- lm(density ~ 1, wine) # model with only an intercept
anova(m2_fit, model_fit)
```

```
## Analysis of Variance Table
##
## Model 1: density ~ 1
## Model 2: density ~ citric_acid + residual_sugar + chlorides + total_sulfur_dioxide +
##     alcohol
##   Res.Df      RSS Df Sum of Sq     F    Pr(>F)
## 1   4897 0.043806
## 2   4892 0.003779  5  0.040027 10364 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- F-stat = 10364
- p-value = less than 2.2e-16

We have a very strong evidence against the hypothesis that `citric_acid`, `residual_sugar`, `chlorides`, `total_sulfur_dioxide`, and `alcohol` are not related to `density`.

(g) Conduct a hypothesis test of the claim that there is no association between the amount of sodium chloride per liter (g/L) (`chlorides`) or the total amount of sulfur dioxide per liter (mg/L) (`total_sulfur_dioxide`) and the wine density in g/mL (`density`). Make sure you state: (i) The null and alternative hypotheses, (ii) the full and reduced model, if appropriate, (iii) The value of the test statistic and p-value, (iv) your conclusion in a complete sentence.

Hypothesis

- H_0: $\beta_3 = \beta_4 = 0$; There is no relationship between the amount of sodium chloride per liter or the total amount of sulfur dioxide per liter and the wine density
- H_A: $\beta_3 \neq 0$ or $\beta_4 \neq 0$; At least one of β_3 or β_4 is not equal to 0

Model

- full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$
- reduced model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 x_5 + \varepsilon$

```r
m3_fit <- lm(density ~ citric_acid+residual_sugar+alcohol, wine)
anova(m3_fit, model_fit)
```

```
## Analysis of Variance Table
##
## Model 1: density ~ citric_acid + residual_sugar + alcohol
## Model 2: density ~ citric_acid + residual_sugar + chlorides + total_sulfur_dioxide +
##     alcohol
##   Res.Df       RSS Df  Sum of Sq      F    Pr(>F)
## 1   4894 0.0039587
## 2   4892 0.0037788  2 0.00017987 116.43 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- F-stat = 116.43
- p-value = less than 2.2e-16

We have a very strong evidence against the hypothesis that `chlorides` and `total_sulfur_dioxide` are not related to `density`.

(h) Conduct a hypothesis test of the claim that there is no association between the amount of sodium chloride per liter (`chlorides`) and the wine density in g/mL (`density`). Make sure you state: (i) The null and alternative hypotheses, (ii) the full and reduced model, if appropriate, (iii) The value of the test statistic and p-value, (iv) your conclusion in a complete sentence.

Hypothesis

- H_0: $\beta_3 = 0$; There is no relationship between the amount of sodium chloride per liter and the wine density
- H_A: $\beta_3 \neq 0$; There is a relationship between the amount of sodium chloride per liter and the wine density

```r
summary(model_fit)
```

```
##
## Call:
## lm(formula = density ~ citric_acid + residual_sugar + chlorides +
##     total_sulfur_dioxide + alcohol, data = wine)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0021263 -0.0005689 -0.0001093  0.0004314  0.0251132
##
## Coefficients:
##                        Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           1.003e+00  1.719e-04 5832.375  < 2e-16 ***
## citric_acid           1.177e-03  1.052e-04   11.191  < 2e-16 ***
## residual_sugar        3.488e-04  2.884e-06  120.939  < 2e-16 ***
## chlorides             2.038e-03  6.229e-04    3.271  0.00108 **
## total_sulfur_dioxide  5.040e-06  3.434e-07   14.676  < 2e-16 ***
## alcohol              -1.148e-03  1.280e-05  -89.742  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.0008789 on 4892 degrees of freedom
## Multiple R-squared:  0.9137, Adjusted R-squared:  0.9137
## F-statistic: 1.036e+04 on 5 and 4892 DF,  p-value: < 2.2e-16
```

- t-stat = 3.271
- p-value = 0.00108

We have a strong evidence against the hypothesis that `chlorides` is not related to `density`.

Data Source and Description: Cortez, Paulo, et al. "Modeling wine preferences by data mining from physicochemical properties." Decision support systems 47.4 (2009): 547-553.