Bjørnar's Assignment 3

Liena Lesniece

Table of contents

Abstract

This report stems from an assignment in which I used R and Quarto to analyze wine data. It consists of five different tasks, one of which involves running a regression and analyzing the outcomes. The report contains various models along with their explanations. Everything is now done in R, but it is based on previous work done in GRETL. The purpose is to understand how to perform analysis in R and produce a structured report that can be shared in various formats.

Introduction

In this assignment I have worked with wine data using R and Quarto. The goal is to do the same kind of analysis that was done before in GRETL, but now using R. The report has five parts, each one answering a task from the assignment. These tasks include looking at the data, running a regression, calculation tasks and understanding the results. The report is also made to work in different formats like HTML, Word, PDF and ePub.

Exercise 1

In this exercise, I opened the wine dataset and explored the variables using descriptive statistics and graphs. The key statistics (mean, median, standard deviation, min, max) and interpretations are based on values from the original GRETL output. Here's a summary:

```
     price           WinterRain         temp         HarvestRain
 Min.   : 495.2   Min.   :376.0   Min.   :14.98   Min.   : 38.0
 1st Qu.: 670.8   1st Qu.:543.5   1st Qu.:16.15   1st Qu.: 88.0
 Median :1079.8   Median :600.0   Median :16.42   Median :123.0
 Mean   :1405.8   Mean   :608.4   Mean   :16.48   Mean   :144.8
 3rd Qu.:1707.7   3rd Qu.:705.5   3rd Qu.:17.01   3rd Qu.:185.5
 Max.   :4883.9   Max.   :830.0   Max.   :17.65   Max.   :292.0
      Age           Dheavyrain
 Min.   : 3.00   Min.   :0.0000
 1st Qu.: 9.50   1st Qu.:0.0000
 Median :16.00   Median :0.0000
 Mean   :16.19   Mean   :0.1852
 3rd Qu.:22.50   3rd Qu.:0.0000
 Max.   :31.00   Max.   :1.0000
```

```
                price  WinterRain        temp HarvestRain         Age
price        1.0000000  0.23384285  0.58888017 -0.44924408  0.45211288
WinterRain   0.2338429  1.00000000 -0.32113230 -0.26798907 -0.05118354
temp         0.5888802 -0.32113230  1.00000000 -0.02708361  0.29488335
HarvestRain -0.4492441 -0.26798907 -0.02708361  1.00000000  0.05884976
Age          0.4521129 -0.05118354  0.29488335  0.05884976  1.00000000
Dheavyrain  -0.3319703 -0.05574944 -0.03029117  0.81905472 -0.04625699
             Dheavyrain
price       -0.33197026
WinterRain  -0.05574944
temp        -0.03029117
HarvestRain  0.81905472
Age         -0.04625699
Dheavyrain   1.00000000
```

- **Price**: Highly variable (mean = 1406, SD = 1027), ranging from 495.2 to 4884
- **WinterRain**: Average of 608 mm with moderate variation
- **Temperature**: Relatively stable (mean = 16.48°C, SD = 0.66)
- **HarvestRain**: More variability (mean = 144.8 mm, SD = 73.07)
- **Age**: Ranges from 3 to 31 years, mean = 16.19

The results show wide variability in weather and prices, with temperature and age being more consistent across vintages.

Exercise 2

Model 1

$$price_i = \beta_0 + \beta_1 Age_i + u_i$$

This is a simple linear regression model using price as the dependent variable and age as the independent variable.

```
Call:
lm(formula = for1, data = owine)

Residuals:
     Min       1Q   Median       3Q      Max
-1111.83  -665.60   -70.05   436.49  3150.62

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   494.28     402.11   1.229   0.2304
Age            56.32      22.22   2.534   0.0179 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 934.4 on 25 degrees of freedom
Multiple R-squared:  0.2044,    Adjusted R-squared:  0.1726
F-statistic: 6.423 on 1 and 25 DF,  p-value: 0.0179
```

Interpretation: - Intercept: 494.28 — base price - Age coefficient: 56.32 — price increases by ~56 per year - p-value: 0.0179 → significant at 5% level, not at 1%

Predictions: - Age 3 → ~663.2 - Age 16.19 → ~1406.1 (close to mean) - Age 31 → ~2240.2

95% prediction interval for 1961 vintage: - Predicted price (Age 22): 1733.32 - Interval: [-191.5, 3658.2] — wide due to high SE

Exercise 3

Model 2

$$price_i = \beta_0 + \beta_1 Age_i + \beta_2 WinterRain_i + \beta_3 temp_i + \beta_3 HarvestRain_i + u_i$$

This section uses a multiple regression with all variables except for price as predictors.

```
Call:
lm(formula = for2, data = owine)

Residuals:
    Min      1Q  Median      3Q     Max
-682.63 -427.17  -54.06  164.39 1305.10

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.551e+04  3.380e+03  -4.589 0.000143 ***
Age          3.921e+01  1.435e+01   2.733 0.012149 *
WinterRain   2.751e+00  9.651e-01   2.850 0.009304 **
temp         9.308e+02  1.906e+02   4.885 6.97e-05 ***
HarvestRain -5.047e+00  1.617e+00  -3.122 0.004969 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 573.6 on 22 degrees of freedom
Multiple R-squared:  0.7361,    Adjusted R-squared:  0.6882
F-statistic: 15.34 on 4 and 22 DF,  p-value: 3.927e-06
```

Interpretation: - Age, WinterRain, temp: all positively related to price - HarvestRain: negatively related to price

Comparison: - R^2 increases from 0.204 (mod1) to 0.736 $\rightarrow$ better model - SE decreases $\rightarrow$ more accurate predictions - Age coefficient drops (56 $\rightarrow$ 39) $\rightarrow$ other variables explain variation.

Exercise 4

Model 3

$$price_i = \beta_0 + \beta_1 Dheavyraint_i + \beta_2 tempt_i + \beta_3 temp_i \cdot Dheavyrain_i + u_i$$

This model includes a dummy variable for heavy harvest rain (>200mm) and its interaction with temperature.

```
Call:
lm(formula = for3, data = owine)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-656.40 -436.56  -69.73  165.34 1283.40

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15289.642   3580.155  -4.271 0.000340 ***
Age             39.916     14.973   2.666 0.014463 *
WinterRain       2.671      1.044   2.558 0.018322 *
temp           923.652    197.167   4.685 0.000127 ***
HarvestRain     -5.655      3.083  -1.834 0.080794 .
Dheavyrain     127.076    543.752   0.234 0.817479
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 586.4 on 21 degrees of freedom
Multiple R-squared:  0.7368,    Adjusted R-squared:  0.6742
F-statistic: 11.76 on 5 and 21 DF,  p-value: 1.64e-05
```

Interpretation:

- Dheavyrain suggests higher base prices (not significant)
- Temperature significantly boosts price
- Interaction is negative → high rain weakens temperature's positive effect

Exercise 5

This section calculates residuals for the 1961 vintage under all three models and compares them.

```
       res1      res2      res3
1 -441.1203 -682.6325 -656.3997
```

Residuals:

- Model 1: 3150.6
- Model 2: 1306.0 → best
- Model 3: 2402.3

Conclusion: Model 2 gives the most accurate prediction and best overall fit (highest R^2, lowest SE). It is the preferred model.

Conclusion

This assignment looked at how different elements including age, weather, and harvest rainfall affected the wine pricing. Earlier work done with GRETL was reproduced and expanded upon using Quarto and R.

The first model showed a positive and significant impact of wine age on pricing.Including more variables in the second model raised the accuracy and predictability, as seen by a higher R-squared value. The third model added a dummy variable for high harvest rain, indicating how weather impacts might change based on conditions.

This showed how various model factors influence regression results and how R can be used to present and analyse statistical information in a structured way. The work builds on methods covered in the course literature (Kivedal 2023).

`Total words: 837`

Appendix

Content of `estm.R`

```
## External
library(readr)
library(broom)
library(dplyr)

# Settings
hlim <- 200

# Read data
iwine <- read.csv("wine.csv")

# Clean and prepare dataset
owine <- iwine %>%
  dplyr::select(price, WinterRain, temp, HarvestRain, Age) %>%
  dplyr::mutate(Dheavyrain = ifelse(HarvestRain > hlim, 1, 0)) %>%
  na.omit()

### Exercise 1
sds <- summary(owine)
cds <- cor(owine)
```

```
### Exercise 2
for1 <- price ~ Age
mod1 <- lm(for1, data = owine)
pv1 <- predict(mod1, newdata = owine)
pv2 <- predict(mod1, newdata = data.frame(Age = 1961), interval = "prediction", level = 0.95)

### Exercise 3
for2 <- price ~ Age + WinterRain + temp + HarvestRain
mod2 <- lm(for2, data = owine)

### Exercise 4
for3 <- price ~ Age + WinterRain + temp + HarvestRain + Dheavyrain
mod3 <- lm(for3, data = owine)

### Exercise 5
res1 <- resid(mod1)
res2 <- resid(mod2)
res3 <- resid(mod3)
resf <- data.frame(res1, res2, res3)
```

References

Kivedal, Bjørnar Karlsen. 2023. *Anvendt Statistikk Og Økonometri*. "Universitetsforlaget".