Software Analytics - Assignment 4

Due date: 11:59 pm Sep 24, 2023.

General information

You can use all available resources: notes, AI apps, Google search... But you need to do all the work on your own. You cannot discuss or share your work with your friends or classmates.

The instructor and TA will not provide any clarification or suggestions. You should work with the best of your knowledge and understanding.

You need to submit a document containing your answers and the R code you use to produce the analysis and answers. You should copy the analysis results and figures produced to your answer document. Please ensure that you submit both the Word/PDF solution file and the "Rhistory" file, which is necessary for a complete evaluation of your work.

Note: If you don't submit the Rhistory file, you will be deducted 50% of the total points.

For each question, before writing the code, you should explain your ideas on how to solve the problem: Why did you choose a particular method, function, or solution? What was your thought process in approaching the problem? Your descriptions will account for half of the points for each question. Providing clear and thoughtful explanations shows that you understand the topic and will be crucial for achieving full marks.

Question

Load the dataset auto.csv (given in the shared Google folder) into R. This dataset has some columns:

name: name of the car model (e.g., bmw 2002)
origin: where the car is produced. 1 = US, 2 = EU, 3 = Asia.
mpg: miles per gallon (the higher the better).
weight: weight in lbs
model_year: the model year
horsepower: the engine power (measured in horsepower). The higher the stronger.
cylinders: the number of cylinders in the car engine. The higher the stronger.
acceleration: the time (in seconds) for the car to speed up from 0 to 60 mph.
displacement: the volume of air the car engine can take in to burn.

Q1 (2 pt). Compare **mpg** of cars: US vs Asia; EU vs Asia, US vs EU using t.test and wilcox.test. Describe the results in non-technical language.

To compare the miles per gallon (mpg) of cars from different regions (US, EU, and Asia), we need to follow a systematic approach. First, we'll load the dataset "auto.csv" into R. Next, we'll filter the data into three subsets based on the "origin" column, creating groups for US, EU, and Asia. Then, for each pair of regions (US vs. Asia, EU vs. Asia, and US vs. EU), we'll perform two statistical tests: the t-test and the Wilcoxon rank-sum test. The t-test assumes normally distributed data and equal variances between groups, making it suitable for comparing means. On the other hand, the Wilcoxon test is non-parametric and does not require normality assumptions, making it useful when these assumptions may not hold. The p-values obtained from these tests will help us assess whether there are statistically significant differences in mpg between the regions. Interpreting the results, we'll look for p-values less than 0.05, indicating significant differences.

```
cars = read.csv("auto.csv")
mpgUS = cars$mpg[mpg$origin == 1]
mpgUS = cars$mpg[cars$origin == 1]
mpgEU = cars$mpg[cars$origin == 2]
mpgAS = cars$mpg[cars$origin == 3]
testUSAS = t.test(mpgUS, mpgAS)
```

testEUAS = t.test(mpgEU, mpgAS)
testUSEU = t.test(mpgUS, mpgEU)
wilcoxUSAS = wilcox.test(mpgUS, mpgAS)
wilcoxEUAS = wilcox.test(mpgEU, mpgAS)
wilcoxUSEU = wilcox.test(mpgUS, mpgEU)

```
> print(testUSAS)

        Welch Two Sample t-test

data:  mpgUS and mpgAS
t = -13.034, df = 138.64, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.997430  -8.836897
sample estimates:
mean of x mean of y
 20.03347  30.45063

> print(testEUAS)

        Welch Two Sample t-test

data:  mpgEU and mpgAS
t = -2.7075, df = 137.85, p-value = 0.007637
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.9273838 -0.7679996
sample estimates:
mean of x mean of y
 27.60294  30.45063

> print(testUSEU)

        Welch Two Sample t-test

data:  mpgUS and mpgEU
t = -8.4311, df = 105.32, p-value = 1.93e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.349583 -5.789361
sample estimates:
mean of x mean of y
 20.03347  27.60294
```

```
> print(wilcoxUSAS)

        Wilcoxon rank sum test with continuity correction

data:  mpgUS and mpgAS
W = 2456.5, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> print(wilcoxEUAS)

        Wilcoxon rank sum test with continuity correction

data:  mpgEU and mpgAS
W = 1889, p-value = 0.001962
alternative hypothesis: true location shift is not equal to 0

> print(wilcoxUSEU)

        Wilcoxon rank sum test with continuity correction

data:  mpgUS and mpgEU
W = 3279, p-value = 1.957e-14
alternative hypothesis: true location shift is not equal to 0

>
```

Both the t-tests and the Wilcoxon tests confirm that US cars have significantly lower mpg than cars from the other regions. The Wilcoxon test also indicates that there is a statistically significant difference in mpg between EU and Asian cars, but the difference is less pronounced than the difference between US and Asian cars. In simple terms, cars in the US tend to be less fuel-efficient compared to both Asian and European cars. European cars are a bit better than US cars in terms of fuel efficiency, but Asian cars come out on top with the best fuel economy.

Q2 (2 pt). Compare **mpg** of cars by cylinders (engine size) of 4, 6, and 8. We exclude cars with cylinders of 3 and 5 because of small sample size. You can use both t.test and wilcox.test. Describe the results in non-technical language.

To compare the miles per gallon (mpg) of cars with different numbers of cylinders (4, 6, and 8), we'll follow a straightforward approach. First, we'll filter the dataset to create three groups based on cylinder count. Then, for each pair of cylinder groups (4 vs. 6, 4 vs. 8, and 6 vs. 8), we'll conduct two statistical tests: a t-test and a Wilcoxon rank-sum test. The t-test will help us assess if there are significant differences in the mean mpg between the two groups, assuming certain statistical assumptions hold. The Wilcoxon test, on the other hand, is non-parametric and does not rely on these assumptions, making it suitable when data distributions are not perfectly normal. We'll interpret the results by examining the p-values from both tests for each comparison, considering p-values less than 0.05 as indicating a statistically significant difference in fuel efficiency. This approach allows us to determine whether the number of cylinders in a car's engine affects its mpg in a simple and straightforward manner.

```
mpg4 = cars$mpg[cars$cylinders == 4]
mpg6 = cars$mpg[cars$cylinders == 6]
mpg8 = cars$mpg[cars$cylinders == 8]
test46 = t.test(mpg4, mpg6)
test48 = t.test(mpg4, mpg8)
```

test68 = t.test(mpg6, mpg8)
wilcox46 = wilcox.test(mpg4, mpg6)
wilcox48 = wilcox.test(mpg4, mpg8)
wilcox68 = wilcox.test(mpg6, mpg8)

```
> test46

        Welch Two Sample t-test

data:  mpg4 and mpg6
t = 16.01, df = 223.27, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  8.164385 10.456466
sample estimates:
mean of x mean of y
 29.28392  19.97349

> test48

        Welch Two Sample t-test

data:  mpg4 and mpg8
t = 29.251, df = 299.73, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 13.35737 15.28426
sample estimates:
mean of x mean of y
 29.28392  14.96311

> test68

        Welch Two Sample t-test

data:  mpg6 and mpg8
t = 9.9274, df = 147.39, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.012997 6.007778
sample estimates:
mean of x mean of y
 19.97349  14.96311
```

```
> wilcox46

        Wilcoxon rank sum test with continuity correction

data:  mpg4 and mpg6
W = 15347, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> wilcox48

        Wilcoxon rank sum test with continuity correction

data:  mpg4 and mpg8
W = 20338, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0

> wilcox68

        Wilcoxon rank sum test with continuity correction

data:  mpg6 and mpg8
W = 7656.5, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

In summary, regardless of whether we used t-tests or Wilcoxon tests, the results consistently show significant differences in fuel efficiency between cars with different numbers of cylinders. Specifically, 4-cylinder cars are the most fuel-efficient, followed by 6-cylinder cars, and 8-cylinder cars are the least fuel-efficient.

Q3 (2 pt). You want to compare the performance of two machine learning algorithms. You do the following experiments:
- Collect 10 datasets
- Run each algorithm on those datasets. For each dataset, you conduct a 10-fold cross validation and measure the average accuracy and running time.

How the result table look like (generate an example table)? How can you use t.test on that result table? Can you use a paired t.test?

Dataset	Algorithm 1 - Accuracy	Algorithm 1 - Time	Algorithm 2 - accuracy	Algorithm 2 - time
1	.97	20 sec	.85	17 sec
2	.67	35 sec	.76	13 sec
…	…	…	…	…
10	.82	12 sec	.90	22 sec

We can use a single t.test to analyze the performance of each algorithm using different data sets or equal data sets if you are using a paired t.test, which you can use. For the later, the paired t-tests will help you determine whether there is a statistically significant difference in performance between Algorithm A and Algorithm B. If the p-values from these tests are less than your chosen significance level (e.g., 0.05), you can conclude that there is a statistically significant difference.