AssignmentPartB

October 3, 2021

1 Thinking about Data / Statistics for Data Science

1.1 Assignment (Part B)

This Jupyter Notebook contains your assignment questions and a corresponding place for your answers. **All your answers for this assignment are to appear in this and only this notebook!**

Make sure that each cell of this notebook is appropriately set to either: 1. Markdown - for providing purely textual input 2. Code - for *R* code; note comments can be included in this cell, but they should be prefixed with a # symbol

When answering a mathematical question, show how you got the answer so at least partial credit can be given if the provided answer is wrong. When providing textual input for an answer: adopt the philosophy that **"less can be more"**. Note that you may be asked to explain your answers -- so make sure your answers are yours.

Submission deadline: 11:59pm on Sunday the 10th of October 2021 (**there will be no extension**)

Late submissions will receive a 10% reduction in marks for each day late.

If you submit late, it is your responsibility to download your assignment file from the Jupyter Server and e-mail the file to me (f.ubaudi@westernsydney.edu.au). The date of that e-mail will be taken as your submission date.

My advice is get started on this assignment immediately, that way you have time to seek assistance as needed.

When the deadline for this assignment arrives, a copy of this (your Jupyter Notebook) will be taken from your home directory. So if you do the assignment off-line, make sure you upload it back into your home directory before the deadline! Make sure you do not rename this file!

You are encouraged to backup your assignment periodically, by downloading a copy or making a copy in your home directory.

There are three questions and their worth is declared in the title. The question title also shows the break-down of marks according to the corresponding sub-questions. **This assignment is worth 20% of the final mark for this unit.** Altogether, your assignments are worth 40% of your final unit mark.

[ ]:
```r
# Adding comments to your code

# Any line that starts with a hash symbol is a comment
sample(0:7) # A comment can also end a line of code
```

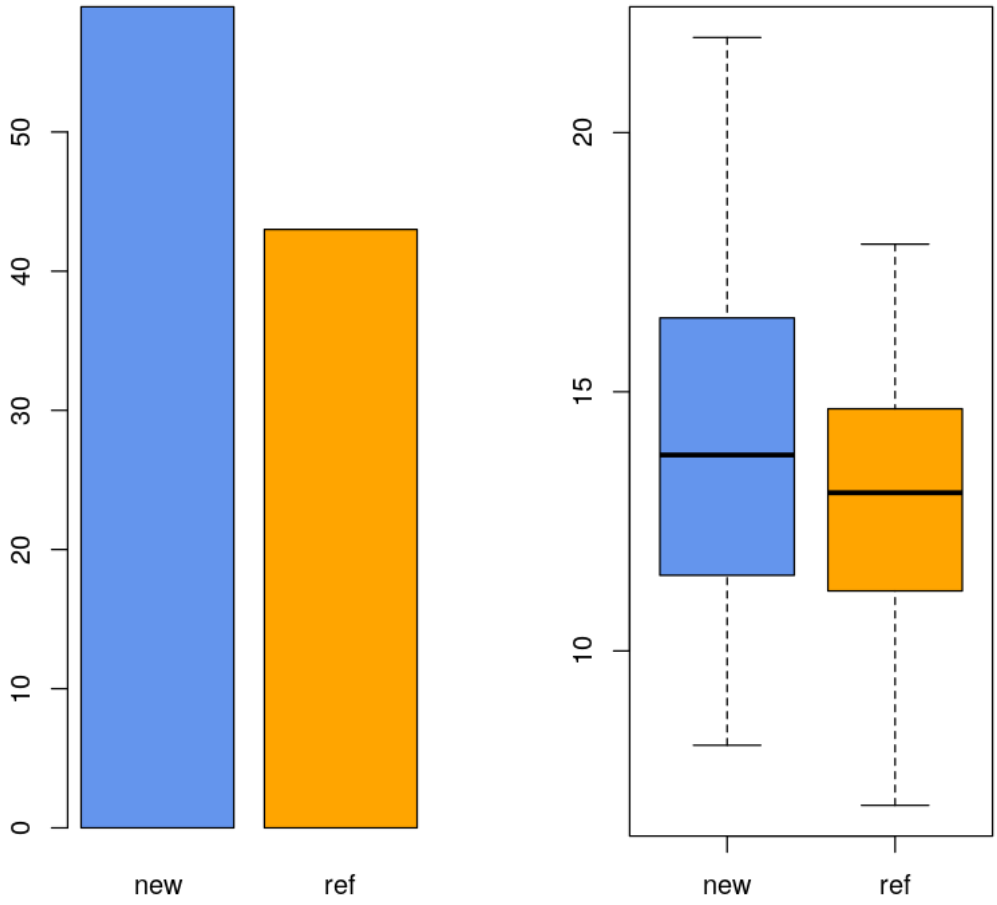
1.2 Question 1: (1 + 1 + 2 + 1 = 5 marks)

A study has been performed, in order to determine whether a *new* drug is in fact a statistically significant improvement over a *reference* drug (a known effective drug). The data file is called *"assignmentB_drugData.csv"*. The data file contains two variables: *response* and *drug*, which respectively correspond to, a measure of effectiveness and the specific drug in question. Consider effectiveness to report how effective the drug is, the larger the numeric value, the more effective.

(i) Write *R* code to load the data file and produce an appropriate visualisation. Make the visualisation appropriate for use in a report.

[ ]:
```r
# we can determine one is larger than the other for continous variables. in␣
↪this case it's response.
```

[58]:
```r
# Your R code to answer the question
drug_data <- read.csv("assignmentB_drugData.csv")
par(mfrow = c(1,2))
barplot(table(drug_data$drug), col = c("cornflowerblue","orange"))
boxplot(response~drug, data = drug_data, pch = 16,
        col = c("cornflowerblue","orange"))
```



(ii) Using the visualisation, very briefly report on what you see and predict whether the new drug is a statistically significant improvement over the old (ref)?

< Your answer >

(iii) Perform an appropriate hypothesis test in order to determine if a statistically significant *improvement* exists. Make sure you clearly state your hypotheses and any other relevant details. Show your working in clear logical steps. Use a simulation consisting of 5,000 replications and a critical value of 0.05.

[ ]:
```
# Your R code to answer the question
## null should be a statistically significant improvement doesn't exist
## H0 = U1 == U2
## H1 = U1 > U2 ## this is one sided hypothesis test
```

```
## alternative cal be !=, < or >.
## null hypothesis define the hypothesis testing
## p-value is determined by Alternate hypothesis.

## we expect simulated mean to be somewhere around 0
```

(iv) Briefly state the conclusion of the above hypothesis test and a briefly interpret it's meaning.

< Your answer >

1.3 Question 2: (1 + 2 + 2 = 5 marks)

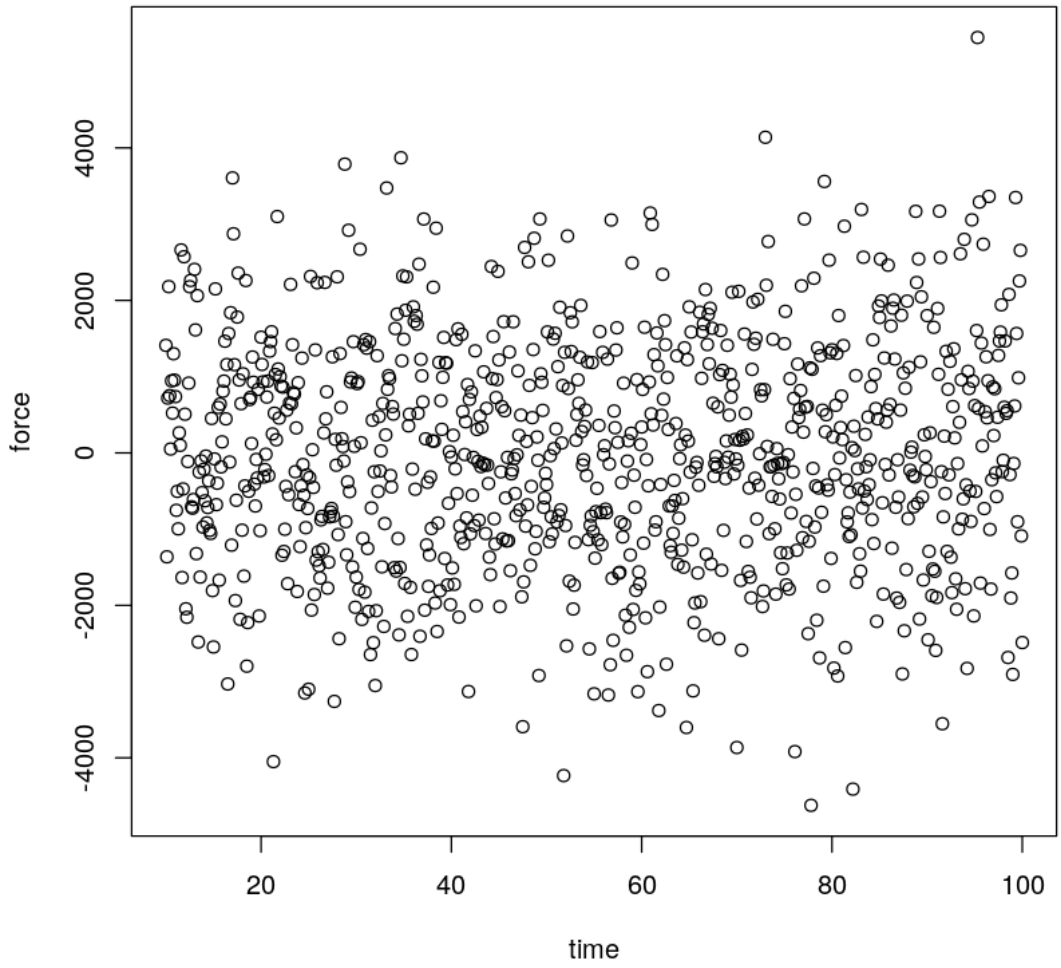
An experiment was performed to evaluate mechanical forces with respect to time. The data file is called "assignmentB_mechanics_1.csv". The data file contains two variables: *time* and *force*, the variable names respectively represent time and change in force. The question being asked is whether time and force may be connected?

(i) Write *R* code to load the data file and produce an appropriate visualisation. Make the visualisation appropriate for use in a report.

[61]:
```
# Your R code to answer the question
mechanics1 <- read.csv("assignmentB_mechanics_1.csv")
head(mechanics1)
plot(force~time, data = mechanics1)
```

A data.frame: 6 × 2

time <dbl>	force <dbl>
10.0	1410.45411
10.1	-1363.55945
10.2	718.85319
10.3	2180.75401
10.4	751.31454
10.5	50.49013



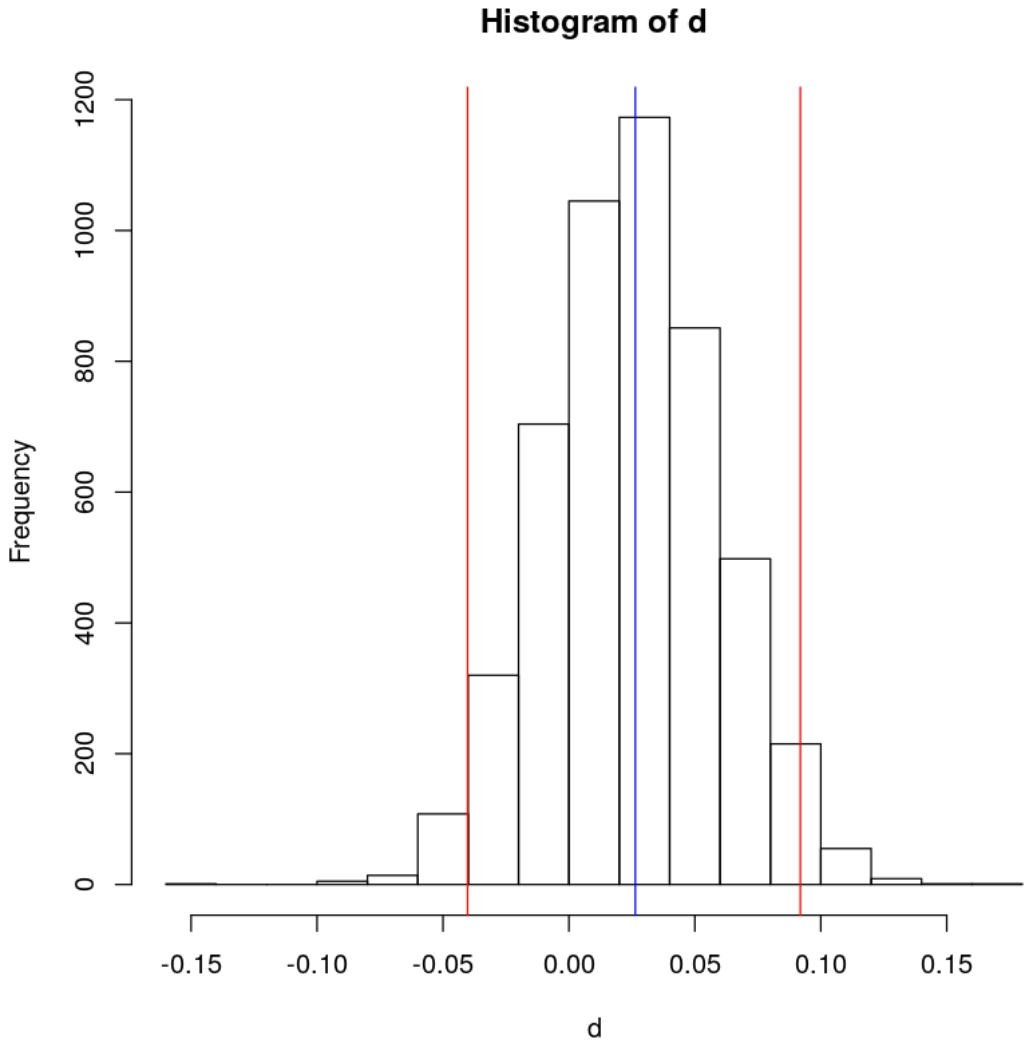
(ii) Using a confidence interval, determine if there is evidence that a relationship does not exist between time and force? Show your working in clear logical steps. Use a simulation consisting of 5,000 replications.

[63]:
```
# Your R code to answer the question
cs <- cor(mechanics1$force, mechanics1$time, method = "spearman")
cp <- cor(mechanics1$force, mechanics1$time)
cs
cp
```

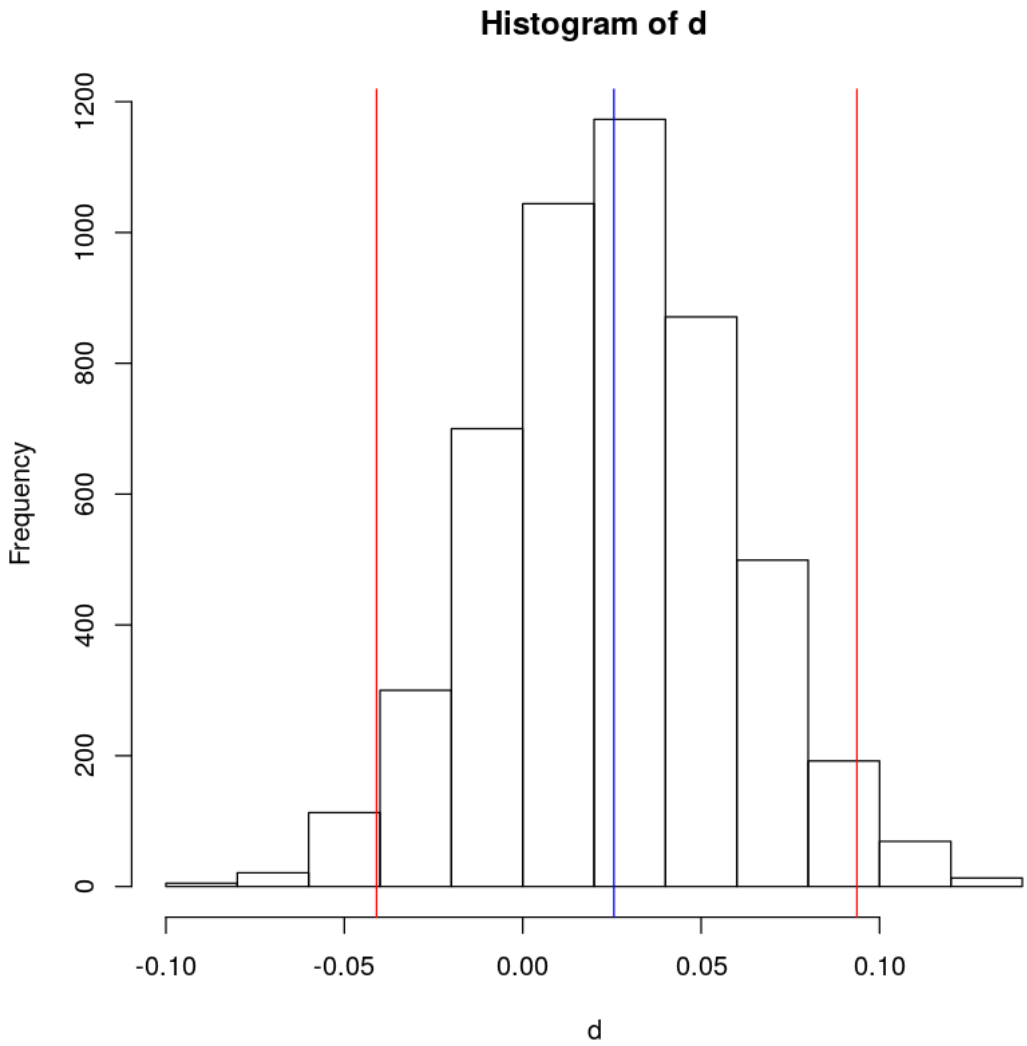
0.0255603817061276

0.0263594598131447

[66]:
```r
res <- cor(mechanics1$force, mechanics1$time)
n = nrow(mechanics1)
d = replicate(5000, {
    ind = sample(1:n, replace = TRUE, size = n)
    cor(mechanics1$force[ind], mechanics1$time[ind])
})
val <- quantile(d, c(0.025, 0.975))

hist(d)
abline(v = res, col = "blue")
abline(v = val, col = "red")
```



[67]:
```
res <- cor(mechanics1$force, mechanics1$time, method = "spearman")
n = nrow(mechanics1)
d = replicate(5000, {
    ind = sample(1:n, replace = TRUE, size = n)
    cor(mechanics1$force[ind], mechanics1$time[ind], method = "spearman")
})
val <- quantile(d, c(0.025, 0.975))

hist(d)
abline(v = res, col = "blue")
abline(v = val, col = "red")
```



(iii) Briefly state the conclusion of the above hypothesis test and a briefly interpret it's meaning.

< Place your answer here >

1.4 Question 3: (1 + 4 + 2 + 1.5 + 1.5 = 10 marks)

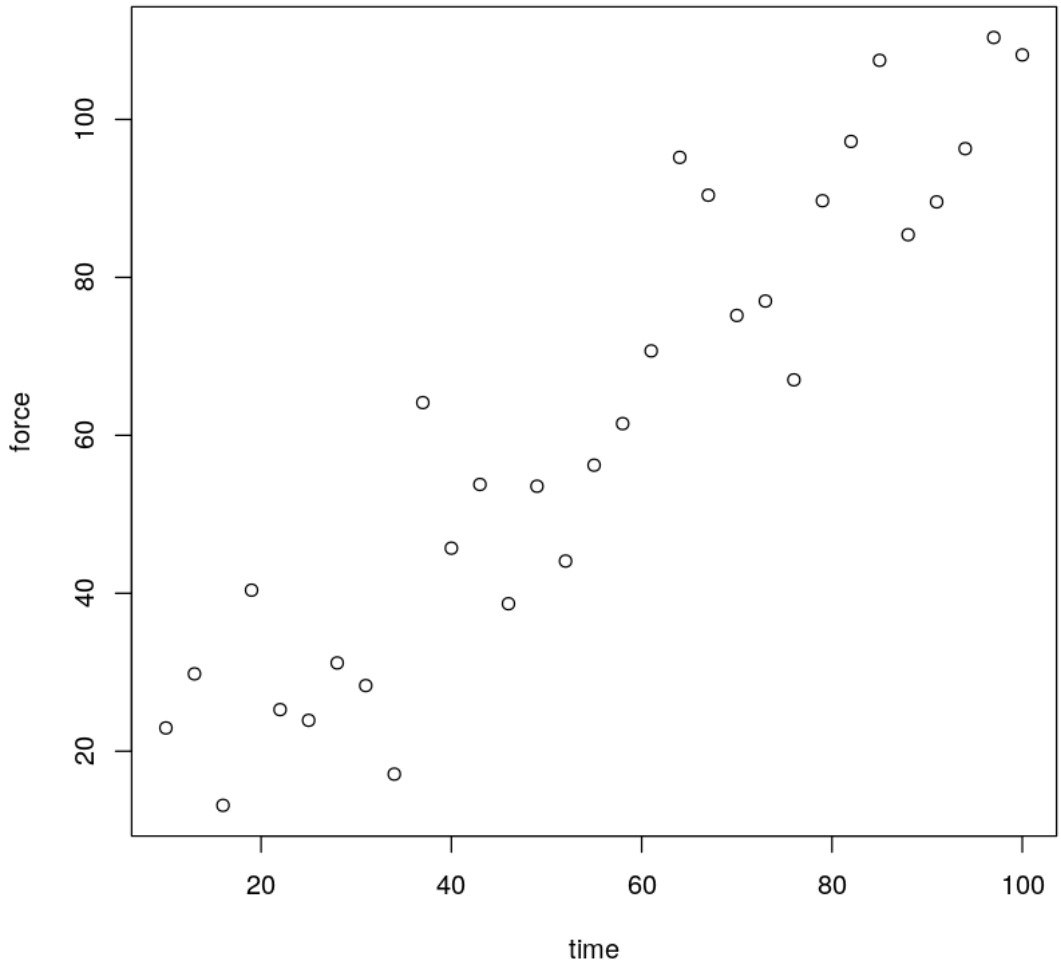
A refinement was performed on the experiment described in question 2. The data file is called "assignmentB_mechanics_2.csv". The data file contains two variables: *time* and *force*, the variable names respectively represent time and force. The question being asked is how *force* changes with respect to *time*? Note that a different force axis was considered.

(i) Write *R* code to load the data file and produce an appropriate visualisation. Add a linear model to the visualisation. Make the visualisation appropriate for use in a report.

[18]:
```
# Your R code to answer the question
mechanics2 <- read.csv("assignmentB_mechanics_2.csv")
head(mechanics2)
plot(mechanics2)
```

A data.frame: 6 × 2

time <int>	force <dbl>
10	22.94948
13	29.79796
16	13.14718
19	40.39359
22	25.27050
25	23.91431

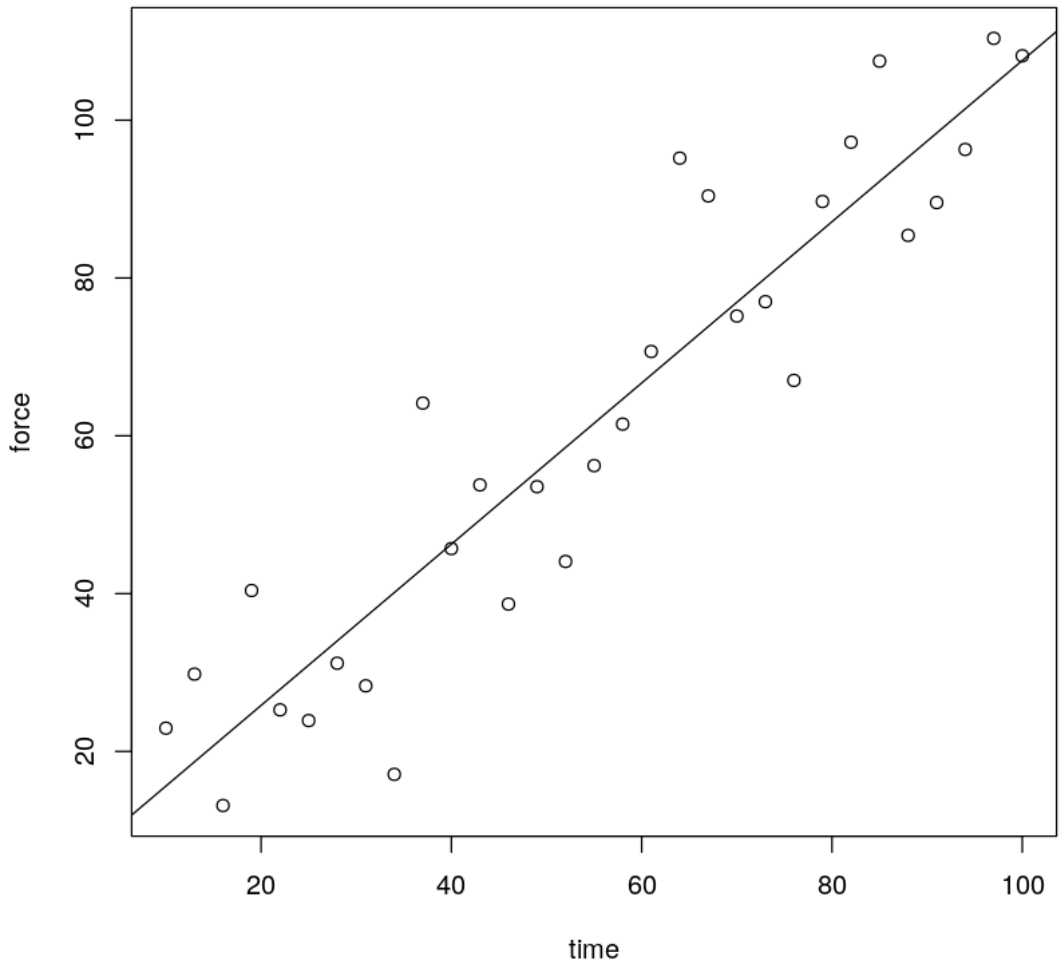


[17]:
```
# creating a linear model
model <- lm(force~time, mechanics2)
summary(model)
coeffs <- coef(model)
coeffs
plot(mechanics2)
abline(coef = coeffs)
```

```
Call:
lm(formula = force ~ time, data = mechanics2)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-23.037  -7.814  -1.909   6.600  24.403

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.40085    4.66252   1.158    0.256
time         1.02165    0.07619  13.409 5.83e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.38 on 29 degrees of freedom
Multiple R-squared:  0.8611,        Adjusted R-squared:  0.8563
F-statistic: 179.8 on 1 and 29 DF,  p-value: 5.829e-14
```

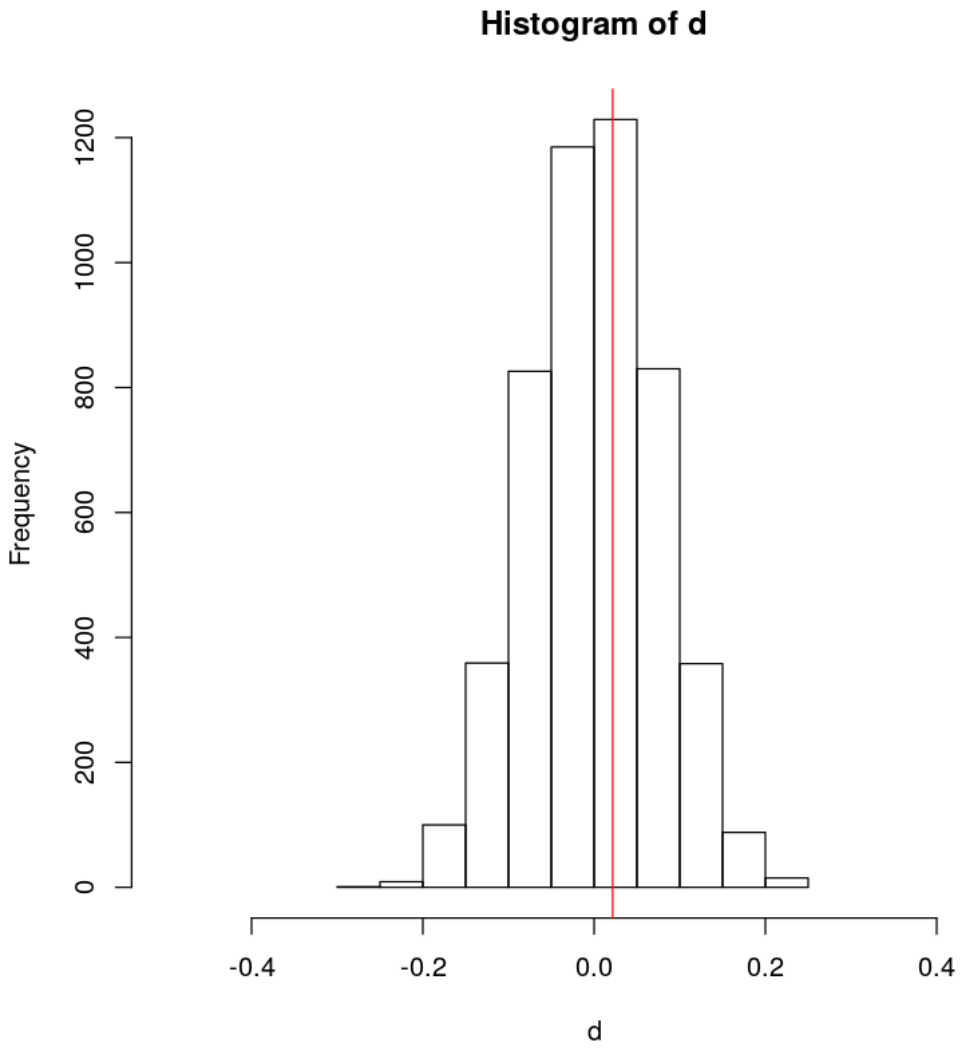
(Intercept) 5.40084886379199 **time** 1.02164869928855



(ii) Using a hypothesis test, determine whether the slope of the linear model constructed in (i) could be 1.0 in general (hence wrt population)? Make sure you clearly state your hypotheses and any other relevant details. Show your working in clear logical steps. Use a simulation consisting of 5,000 replications and a critical value of 0.05.

[55]:
```r
# Your R code to answer the questionn m ,.
m <- lm(force - time ~ time, mechanics2)
b <- coef(m)[2]
cat("Slope is :",b)
d <- replicate(5000, {
    t <- sample(mechanics2$time)
    m <- lm(force - time ~ t, mechanics2)
    coef(m)[2]
})

hist(d, xlim = c(-1,1) * 0.5)
abline(v = b, col = "red")
```

```
Slope is : 0.0216487
```



[57]:
```
pVal <- mean(abs(d) > abs(b))
pVal
```

0.7812

(iii) Using a simulation consisting of 5,000 replications, determine a 95% confidence interval for the slope. Show your working in clear logical steps.

[23]:
```
# Your R code to answer the question

rowCount <- nrow(mechanics2)

d1 <- replicate(5000,
                {
```

```
                ind <- sample(1:rowCount,
                              replace = TRUE)
                s <- mechanics2[ind, ]
                m1 <- lm(force ~ time, s)
                coef(m1)[2]
              })

quantile(d1, c(0.025, 0.975))
```

2.5\% 0.897194128427255 **97.5\%** 1.16376179581904

(iv) Regarding sub-question (ii), state your conclusion to the test and how you arrived at that conclusion. Make sure you state what the conclusion means with respect to the slope.

< Place your answer here >

(v) Regarding sub-question (iii), interpret the implications of the confidence interval. Also relate the implications to the findings of sub-question (ii).

< Place your answer here >