ALY6010

R-PRACTICE MODULE 5

Name: Shreyansh Bhalodiya

NUID: 002664707

Instructor name: Mohsen Soltanifar

Dataset

A tibble: 6 × 8

id	time	treatment	smoker	hrt_months	wbc	age	outcome
<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
165	0	Gabapentin	1	3	5.7	61.4	14.6
165	1	Gabapentin	1	3	5.7	61.4	8.7
165	2	Gabapentin	1	3	5.7	61.4	8.3
165	3	Gabapentin	1	3	5.7	61.4	6.9
165	4	Gabapentin	1	3	5.7	61.4	6.4
166	0	Placebo	0	4	7.8	55.2	20.6

- We have a dataset with 8 columns. Looking at the columns we can see that it's about the treatment a person has gone through with different features like smoker, time, wbc, hrt_months age and outcome.
- Cleaning data from null values.

```
df <- na.omit(df)
sum(is.na(df$age))
### https://sparkbyexamples.com/r-programming/remove-rows-with-na-in-r/
```

```
0
```

PART 1

Let's first check the correlation with numerical columns after cleaning data from null values. We have smoker, hrt_months, wbc, age, outcome as major variables.

```
cor(df_1)
```

A matrix: 5 × 5 of type dbl

	smoker	hrt_months	wbc	age	outcome
smoker	1.000000000	-0.006440119	0.17798414	-0.12818466	0.01510919
hrt_months	-0.006440119	1.000000000	0.14778555	0.16682702	0.02223168
wbc	0.177984137	0.147785546	1.00000000	0.05343805	-0.03683491
age	-0.128184664	0.166827025	0.05343805	1.00000000	0.09041397
outcome	0.015109187	0.022231678	-0.03683491	0.09041397	1.00000000

ANALYSIS

- We can clearly see that data has no major correlation between variables selected above. We do have positive and negative correlation between few variables for example smoker$hrt_months, smoker$age and positive correlation between few variables like outcome$smoker, outcome$age and many more.
- We don't have any strong positive or negative correlation between any pair from the selected variable.
- Let's look at the regression analysis on the next page and see how it differs between pairs and groups of pairs. Do we get any strong R-value? Let's see.

PART 2

```
MODEL INFO:
Observations: 956
Dependent Variable: wbc
Type: OLS linear regression

MODEL FIT:
F(1,954) = 31.21, p = 0.00
R² = 0.03
Adj. R² = 0.03

Standard errors: OLS
-----------------------------------------------
                    Est.   S.E.   t val.      p
----------------- ------ ------ -------- ------
(Intercept)         6.36   0.06   113.32   0.00
smoker              0.96   0.17     5.59   0.00
-----------------------------------------------
```

```
MODEL INFO:
Observations: 956
Dependent Variable: outcome
Type: OLS linear regression

MODEL FIT:
F(1,954) = 7.86, p = 0.01
R² = 0.01
Adj. R² = 0.01

Standard errors: OLS
-----------------------------------------------
                    Est.   S.E.   t val.      p
----------------- ------ ------ -------- ------
(Intercept)         0.43   4.64     0.09   0.93
age                 0.24   0.09     2.80   0.01
-----------------------------------------------
```

ANALYSIS

- We tested correlation between wbc and smoker and regression value between the same pair. We found that correlation is 0.117 which is a positive correlation but not a strong one. On the same lines we got an R-value as 0.03 which is again not strong hence we don't have any strong relation between these two pairs.
- On the similar line if we check correlation between age and outcome is 0.09 which is again not a strong correlation and R-value we got is 0.01 which is again weak.
- We also tested the Regression fit for a group of variables like wbc, age and outcome with smoker as dependent variable but again the R-value we got is 0.05 which is not strong enough to hold the algorithm.

References

- N. (2022, July 20). *How to Remove Rows with NA in R*. Spark by {Examples}. https://sparkbyexamples.com/r-programming/remove-rows-with-na-in-r/
- How to check and create a regression line. https://cran.r-project.org/web/packages/jtools/vignettes/summ.html