Homework 2 Report

Name

Annika Godines

Setup

Read in the data with `read_csv()` and store the data as an R object named `dataset`. Check the data to make sure all of the expected observations and variables are there.

```r
## Load the data and any necessary packages here.

dataset <- read.csv("maacs.csv.gz")
#Compared to Canvas HW2 Overview
names(dataset) # Correctly Matches
```

```
##  [1] "id"             "age"            "gender"         "fev1"          
##  [5] "eNO"            "cough_days"     "pm25"           "no2"           
##  [9] "mouse"          "mouse_allergic"
```

```r
head(dataset, 10) # Age & pm25 & no2 are different (not rounded?)
```

```
##          id   age gender fev1 eNO cough_days   pm25    no2 mouse mouse_allergic
## 1  fd171e2d 14.72   male 1.78 141          2 15.560     NA  2423            yes
## 2  a66fc33a 13.56 female 2.12  68          8 18.890 12.827   939            yes
## 3  fc038e68 14.50   male 2.73 210          2 17.808     NA   200             no
## 4  8e28b8c2 13.96 female 2.36  23          0 13.956     NA    NA             no
## 5  b2699b54 16.63 female 3.13  18         14 17.864 30.976 10371             no
## 6  7d4f3508 16.56 female 2.59 128          0 43.784 10.590  4789            yes
## 7  3d8242a6 17.18 female 2.60  19          0 26.313     NA   760            yes
## 8  f401998a 15.46   male 3.49  26          4 39.864 32.902   264             no
## 9  357fdacb 12.68 female 2.29  17          1 27.081 25.659   419             no
## 10 2a722e16 16.02   male 2.37 134         14 64.578 17.953   187            yes
```

Part 1

We will first consider the relationship between FEV1 and age. In general, it is expected that as children get older (and hence, larger in size), their FEV1 values should get higher.

Consider the statement “FEV1 values in children are higher in older children relative to younger children”.

Write a function in R that takes the `dataset` object as an argument and returns `TRUE` if the statement above is true for the dataset and `FALSE` otherwise.

NOTE: In order to write this function, you will need to translate the statement above into something that can be checked with the data. There are many ways in which you can do that translation correctly and you only need to pick one way here.

NOTE: For this part, do not use any plots.

```r
## Write your function here
```

```
fev1_age_corr <- function(dataset) {
  r <- cor(dataset$age, dataset$fev1, use = "complete.obs")
  return(r > 0)
}
```

In this approach, I calculate the correlation between age and FEV1. If it is positive, the statement is true. However, one can also compare the mean for different age groups.

Part 2

Fit a linear regression model with FEV1 as the outcome and age as a predictor.

How much does FEV1 change for a 1-year increase in the child's age?

```
## Add your code here
fit_lin_reg_age <- lm(fev1 ~ age, data = dataset)
summary(fit_lin_reg_age)
```

```
##
## Call:
## lm(formula = fev1 ~ age, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.51099 -0.27100  0.00196  0.23616  2.15477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.001199   0.134696  -0.009    0.993
## age          0.171162   0.010984  15.583   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4827 on 131 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.6496, Adjusted R-squared:  0.6469
## F-statistic: 242.8 on 1 and 131 DF,  p-value: < 2.2e-16
```

Write your data analysis statement interpreting the regression model here:

Given the results of the linear regression, FEV1 increases approximately 0.17 liters every 1-year increase of age. This supports the statement that: older children tend to have higher FEV1 values.

Part 3

Develop **three** supporting premises derived from the data that support the statement you wrote in Part 2. These can be plots, other summary statistics, or model results.

NOTE:

- At least one supporting premise should use a plot.
- Do not use the code you write in Part 1 as a supporting premise

```
## Add your code here
#SP 1
medianAge <- median(dataset$age, na.rm = TRUE)
meanYoung <- mean(dataset$fev1[dataset$age <= medianAge], na.rm = TRUE)
```

```
meanOld <- mean(dataset$fev1[dataset$age > medianAge], na.rm = TRUE)
meanYoung; meanOld
```

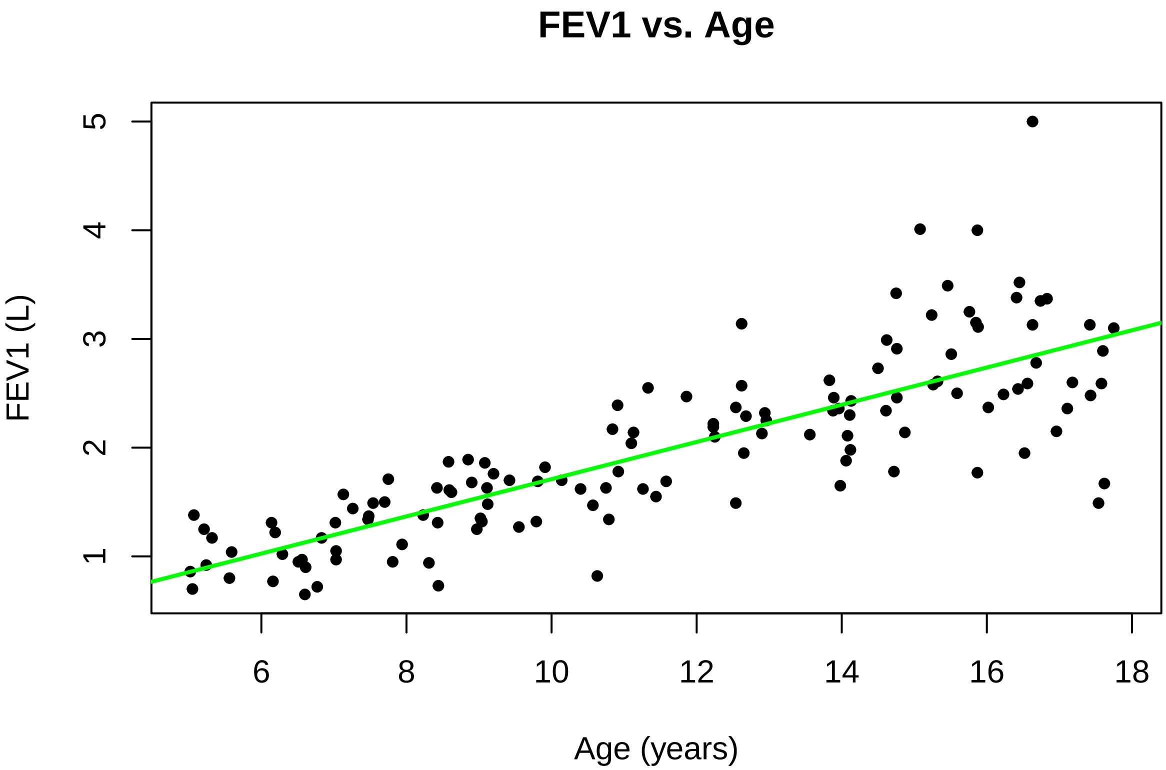
```
## [1] 1.3305
```

```
## [1] 2.539178
```

```
#SP 2
fit_age <- lm(fev1 ~ age, data = dataset)
coef(fit_age)["age"]
```

```
##       age
## 0.1711623
```

```
#SP 3
plot(dataset$age, dataset$fev1,
     xlab = "Age (years)",
     ylab = "FEV1 (L)",
     main = "FEV1 vs. Age",
     pch = 20, col = "black")
abline(lm(fev1 ~ age, data = dataset), col = "green", lwd = 2)
```



Write the three supporting premise statements here:

1. The mean FEV1 of children who are older than the average age is higher than those of the younger than or equal to average age children.
2. The estimated regression slope for age is positive, indicating FEV1 does increase as age increases.
3. The scatterplot of FEV1 versus age indicated a upward-slopeing regression, meaning there is a positive relationship.

Part 4

For each of the supporting premises above, write a function that takes the `dataset` object as an argument and returns `TRUE` if the supporting premise statement above is true for the dataset and `FALSE` otherwise.

For statements involving plots, instead of returning `TRUE` or `FALSE`, you function should do two things:

1. Produce the plot that is used in the statement
2. Produce a hypothetical version of the plot in the event that the statement is true. This can be done using simulated data or by simply hand drawing a plot.

```r
## Function for supporting premise statement 1
sp1 <- function(dataset) {
  ok <- complete.cases(dataset$age, dataset$fev1)
  data <- dataset[ok, c("age", "fev1")]
  if (nrow(data) < 4) return(FALSE)

  avg_age <- median(data$age)
  YoungMean <- mean(data$fev1[data$age <= avg_age])
  OldMean <- mean(data$fev1[data$age > avg_age])

  OldMean > YoungMean

}
```

```r
## Function for supporting premise statement 2
sp2 <- function(dataset) {
  ok <- complete.cases(dataset$age, dataset$fev1)
  data <- dataset[ok, c("age", "fev1")]
  if (nrow(data) < 3) return(FALSE)

  fit <- lm(fev1 ~ age, data = data)
  coef(fit)["age"] > 0
}
```

```r
## Function for supporting premise statement 3
sp3 <- function(dataset) {
  ok <- complete.cases(dataset$age, dataset$fev1)
  data <- dataset[ok, c("age", "fev1")]
  if (nrow(data) < 3) stop("Not enough complete cases to plot this.")

  plot(data$age, data$fev1,
       xlab = "Age (in Years)",
       ylab = "FEV1 (L)",
       pch = 20)
  abline(lm(fev1 ~ age, data = data), color = "green", lwd = 2)

  set.seed(42)
  age_center <- data$age - mean(data$age)
  fev1_simu <- 2 + 0.15 * age_center + rnorm(length(age_center), sd = 0.2)

  plot(data$age, fev1_simu,
       xlab = "Age (in Years)",
       ylab = "FEV1 (L, simulated)",
       main = "Hypothetical: FEV1 vs Age",
       pch = 20)
  abline(lm(fev1_simu ~ data$age), col = "green", lwd = 2)

  invisible(NULL)
}
```

Execute each of your function and show that the produce the expected output.

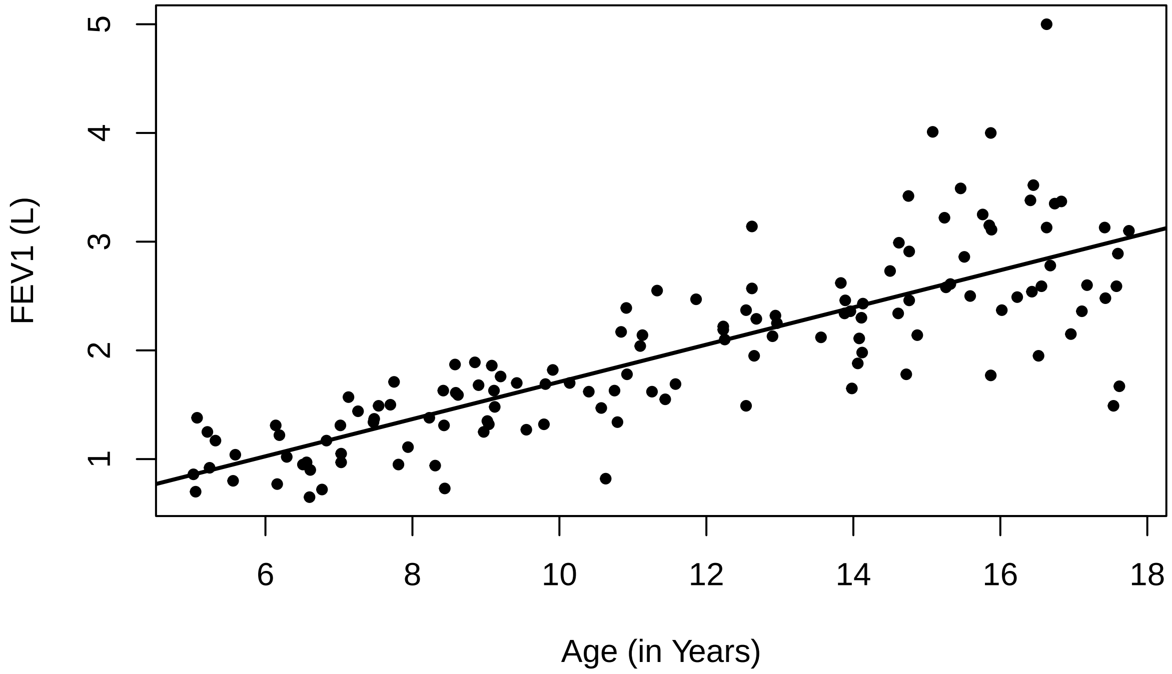
```
sp1(dataset)
```

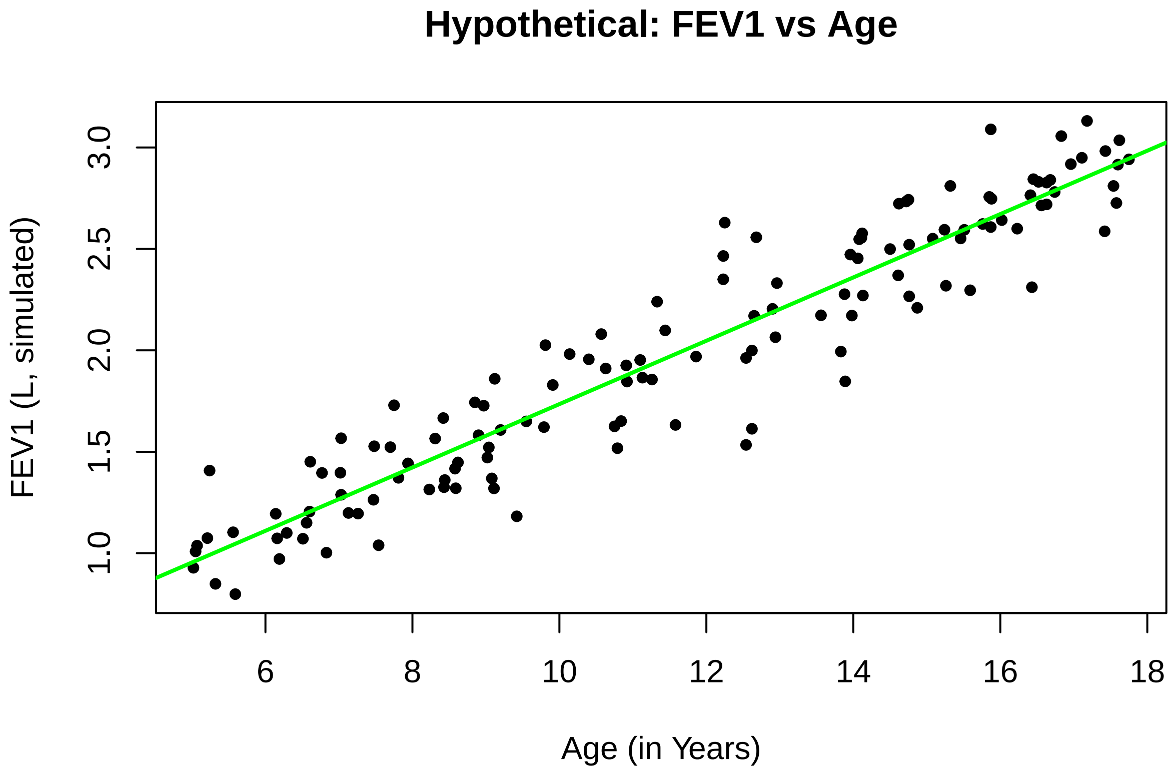
```
## [1] TRUE
```

```
sp2(dataset)
```

```
##  age
## TRUE
```

```
sp3(dataset)
```





Part 5

Describe one alternative to the primary statement "FEV1 values in children are higher in older children relative to younger children".

Another Primary Statement could be:

"FEV1 values aren't associateed to age; the FEV1 values on similar."

Create a fault tree for the alternative outcome describing how the alternative outcome could be realized in the data even if the primary statement were true.

Your fault tree should be created as a separate image and does not need to be created in R. Upload the image of the fault tree to Canvas.