AdvStDaAn, Worksheet, Week 4

Micheal Lappert

12.04.2022

Contents

Exercise 1

```
path <- file.path('Datasets', 'turbines.dat')
df <- read.table(path, header=TRUE)

summary(df)
```

Dataset loading and sanity check:

```
##      Hours         Turbines        Fissures
##  Min.   : 400   Min.   :13.00   Min.   : 0.000
##  1st Qu.:1600   1st Qu.:33.50   1st Qu.: 4.500
##  Median :2600   Median :39.00   Median : 7.000
##  Mean   :2582   Mean   :39.27   Mean   : 9.636
##  3rd Qu.:3600   3rd Qu.:41.00   3rd Qu.:15.000
##  Max.   :4600   Max.   :73.00   Max.   :22.000
```

```
str(df)
```

```
## 'data.frame':    11 obs. of  3 variables:
##  $ Hours   : int  400 1000 1400 1800 2200 2600 3000 3400 3800 4200 ...
##  $ Turbines: int  39 53 33 73 30 39 42 13 34 40 ...
##  $ Fissures: int  0 4 2 7 5 9 9 6 22 21 ...
```

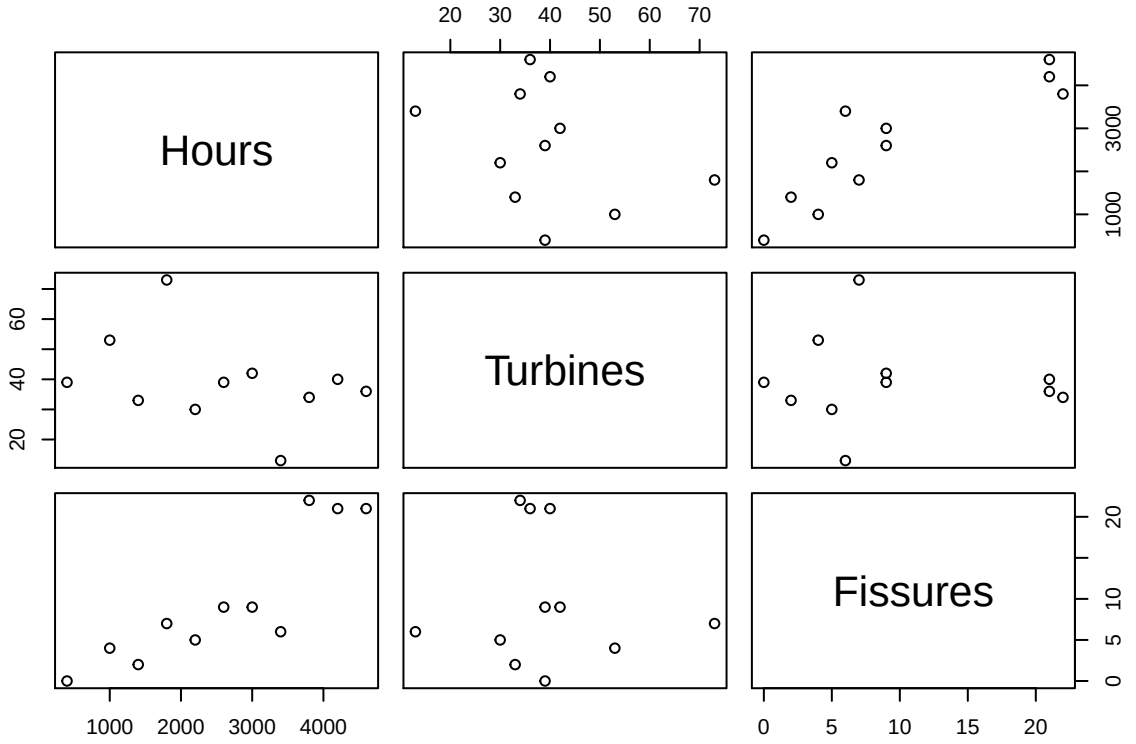
```
head(df)
```

```
##   Hours Turbines Fissures
## 1   400       39        0
## 2  1000       53        4
## 3  1400       33        2
## 4  1800       73        7
## 5  2200       30        5
## 6  2600       39        9
```

```
tail(df)
```

```
##    Hours Turbines Fissures
## 6   2600       39        9
## 7   3000       42        9
## 8   3400       13        6
## 9   3800       34       22
## 10  4200       40       21
## 11  4600       36       21
```

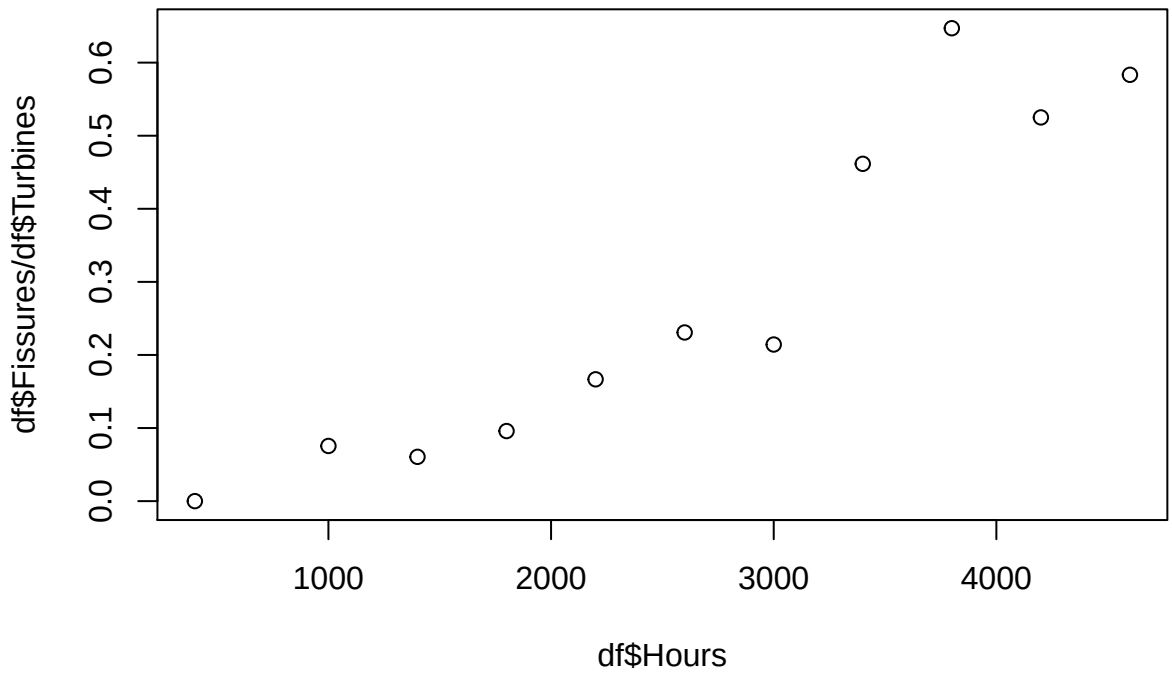
```
plot(df)
```



The data is ascending sorted in hours and looks fine.

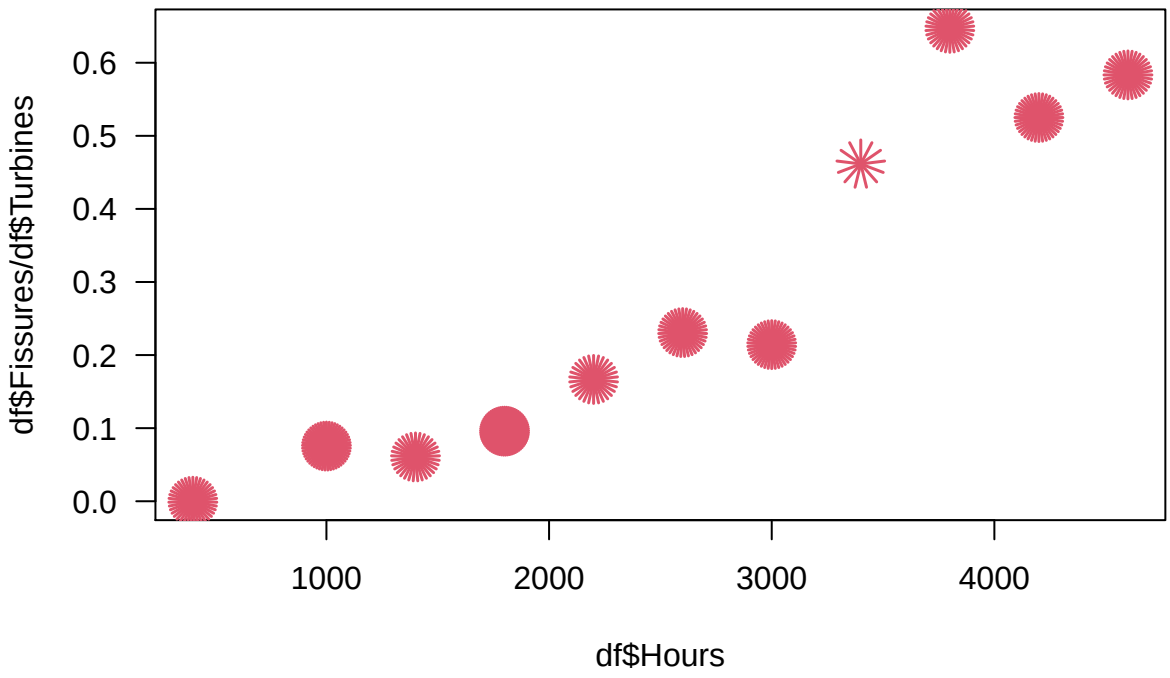
Exercise 1.a)

```
par(mfrow=c(1,1))
plot(df$Hours, df$Fissures/df$Turbines)
```



This plot does not show the density per observation. So one might consider an alternative plot where the densitiy is visualized as well.

```
sunflowerplot(df$Hours, df$Fissures/df$Turbines,
              number = df$Turbines, las = 1)
```



Exercise 2.b)

Let Y_i be the number of wheels with fissures. Then
$Y_i \sim$ independent Binomial(π_i, #Turbines)
with
$log(\frac{pi_i}{1-pi_i}) = \beta_0 + \beta_1 * Hours$

```
glm1.1 <- glm(cbind(Fissures, Turbines-Fissures) ~ Hours, family = binomial, data = df)
summary(glm1.1)
```

```
##
## Call:
## glm(formula = cbind(Fissures, Turbines - Fissures) ~ Hours, family = binomial,
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5055  -0.7647  -0.3036   0.4901   2.0943
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.9235966  0.3779589 -10.381   <2e-16 ***
## Hours        0.0009992  0.0001142   8.754   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 112.670  on 10  degrees of freedom
## Residual deviance:  10.331  on  9  degrees of freedom
## AIC: 49.808
##
## Number of Fisher Scoring iterations: 4
```

Exercise 2.b)

```
coef(glm1.1)
```

```
##   (Intercept)         Hours
## -3.9235965551  0.0009992372
```

$log(\frac{pi_i}{1-pi_i}) = -3.9235965551 + 0.0009992372 * Hours$

Henc the probability of fissures increases by a facotr of exp(0.0009992372) = ´r exp(0.0009992372)´

Question 2.b)

How do we know that the increase of the probability of fissures is related to 100 hours? Why not per 1 hour?