Data Analysis of cancer:veteran Dataset

Budurean Marius-Mihai
Faculty of Mathematics and Computer Science
Big Data - Data Science, Analytics and Technologies

January 2024

1 Introduction

The dataset **cancer:veteran** belongs to the **survival** package from R. The data from this dataset is made from the findings of a randomized study of two different treatments for lung cancer.

Nr Crt	trt	celltype	time	status	karno	diagtime	age	prior
1	1	squamous	72	1	60	7	69	0
2	1	squamous	411	1	70	5	64	10
3	1	squamous	228	1	60	3	38	0
4	1	squamous	126	1	60	9	63	10
5	1	squamous	118	1	70	11	65	10
6	1	squamous	10	1	20	5	49	0
7	1	squamous	82	1	40	10	69	10
8	1	squamous	110	1	80	29	68	0
9	1	squamous	314	1	50	18	43	0
10	1	squamous	100	0	70	6	70	0
11	1	squamous	42	1	60	4	81	0
12	1	squamous	8	1	40	58	63	10
13	1	squamous	144	1	30	4	63	0
14	1	squamous	25	0	80	9	52	10
15	1	squamous	11	1	70	11	48	10
16	1	smallcell	30	1	60	3	61	0
17	1	smallcell	384	1	60	9	42	0
18	1	smallcell	4	1	40	2	35	0
19	1	smallcell	54	1	80	4	63	10
20	1	smallcell	13	1	60	4	56	0

Table 1: Sequence from Cancer Veteran Dataset

The dataset 1 contains a total of 137 records covering the following characteristics were collected from the subjects on whom this study was conducted:

Variable	**Description**
trt	Type of treatment
celltype	Type of cancerous cell
time	Individual survival time
status	Censorship status of the information
karno	Karnofsky's performance score
diagtime	Months from diagnosis to randomisation
age	Age in years
prior	Prior therapy

Table 2: Dataset's Variables

In this paper, I will describe in more detail the component variables of the dataset, and then, based on them, verify a set of hypotheses that result from its analysis and create a predictive model that can estimate the values of a variable based on the values of a set of predictors.

2 Descriptive Analysis of the dataset

The dataset contains 8 variables. Their meaning was explained in the introduction. In this chapter, all of them will be analyzed individually. The analysis will come together with the modifications I made to the dataset and an explanation for each one.

2.1 TRT

The **TRT** is a numerical variable, that describes the type of treatment taken by the subject (1 = standard, 2 = test). I found it more suitable to transform it into a nominal qualitative variable because 1 and 2 do not describe properly the type of treatment.

From Fig. 1 it can be seen that the two existing samples are relatively balanced with a small difference between them.

2.2 CELLTYPE

The **celltype** variable is a nominal qualitative variable because the values its values are representative of the type or size of the individual's cancer cells (smallcell, squamous, large, adeno).

It can be observed that the number of subjects with smallcell is preponderant.

2.3 TIME

The **time** variable is a quantitative variable because it represents the life span of an individual. The unit of measurement of this variable is not specified.

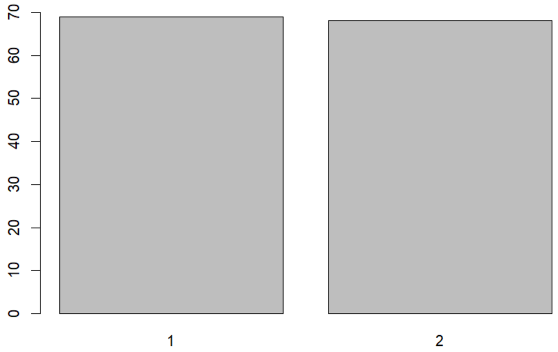

Figure 1: TRT Histogram

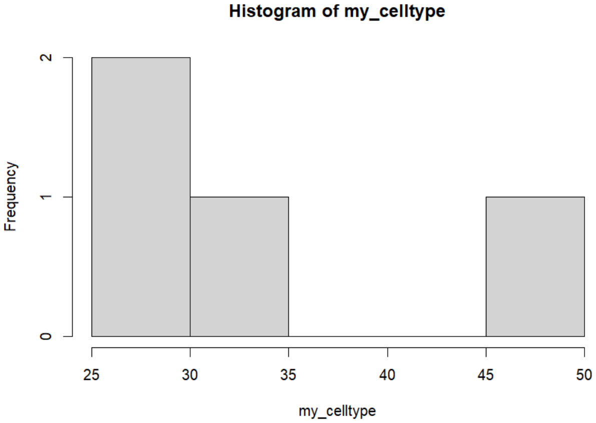

Figure 2: Celltype Histogram

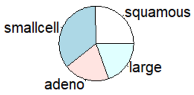

Figure 3: Celltype Pie Chart

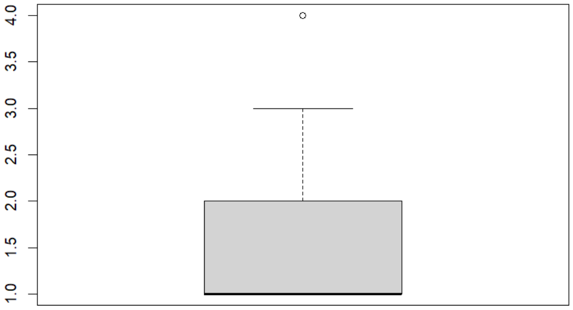

Figure 4: Time Boxplot

As can be seen in Fig. 4, most of the values of this variable is centered on the sample mean, but there are some exceptions.

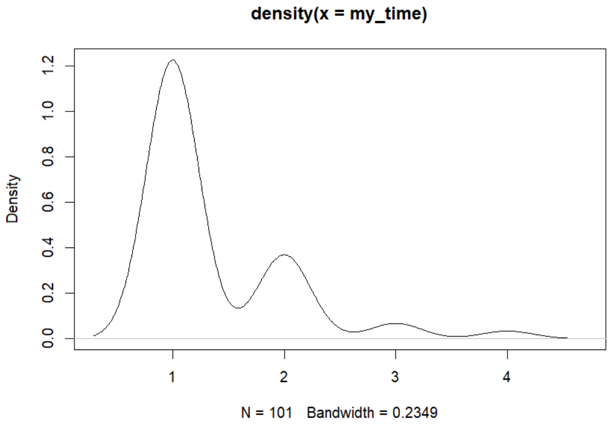

Figure 5: Time Density Plot

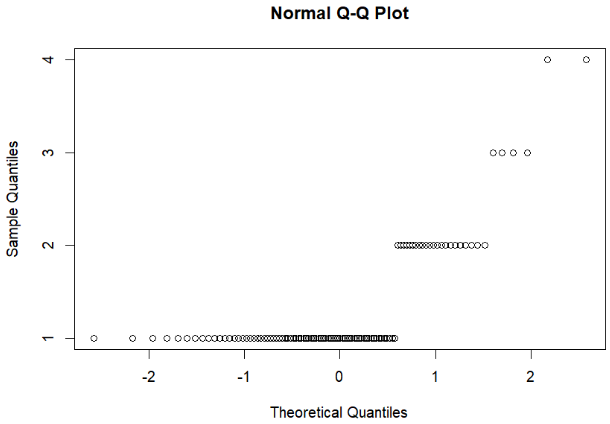

Figure 6: Time Q-Q Norm

In Fig. 5 it can be seen that the distribution comes from a normal distribution, which is distorted by the existence of outliers.

2.4 STATUS

The **status** is a qualitative numerical value because it signifies the censorship status of the individual's information. Since all values are 0 or 1, I considered that a better approach for this variable is to be logical.

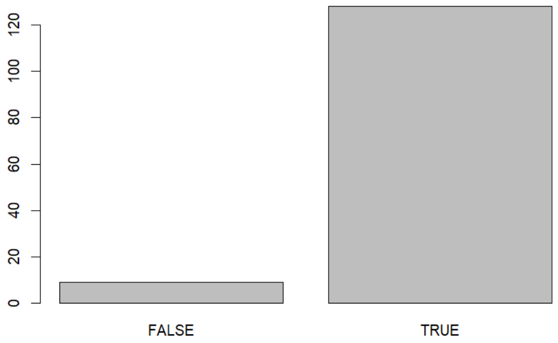

Figure 7: Status Barplot

2.5 KARNO

The **karno** variable is a quantitative variable, which represents the performance score of Karnosky. Its values can be between 0 (bad) and 100 (good). This variable is a degree of degradation for each individual.

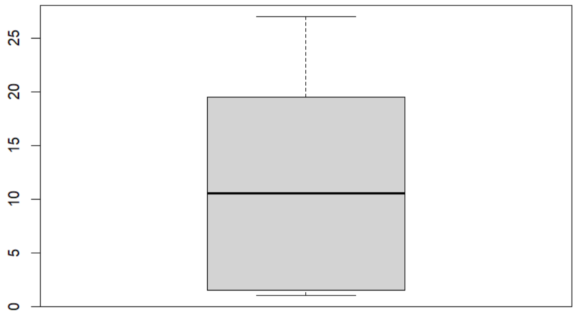

Figure 8: Karno Boxplot

In Fig. 8 it can be seen that the tests recorded in this dataset were performed, with a few exceptions on individuals in a relatively poor condition, according to the Karnofsky index.

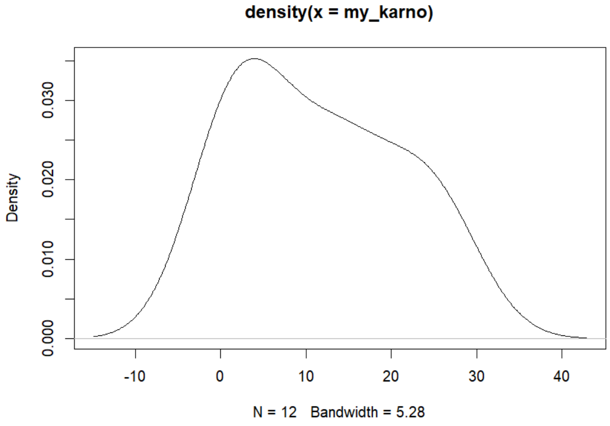

Figure 9: Karno Density Plot

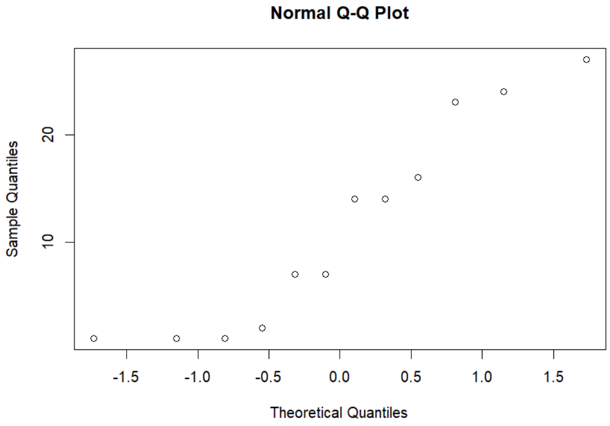

Figure 10: Karno Q-Q Norm

2.6 DIAGTIME

The **digtime** variable is a quantitative variable, which signifies the period from diagnosis to a point of interest in the evolution of the disease. This variable is expressed in months.

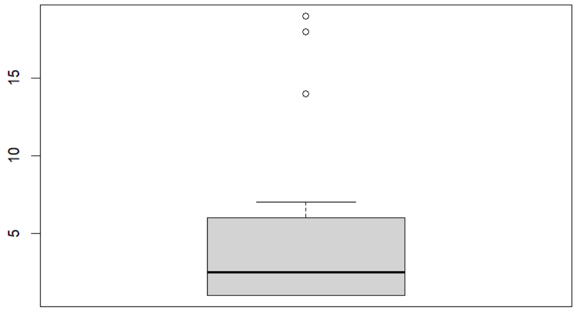

Figure 11: Diagtime Boxplot

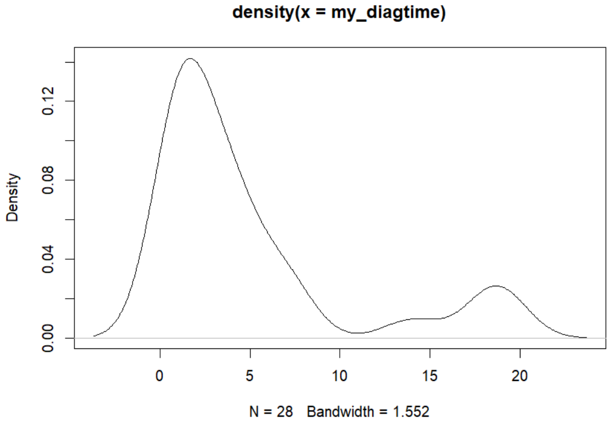

Figure 12: Diagtime Density Plot

2.7 AGE

The **age** variable is numeric, and it represents the age of the subject, measured in years.

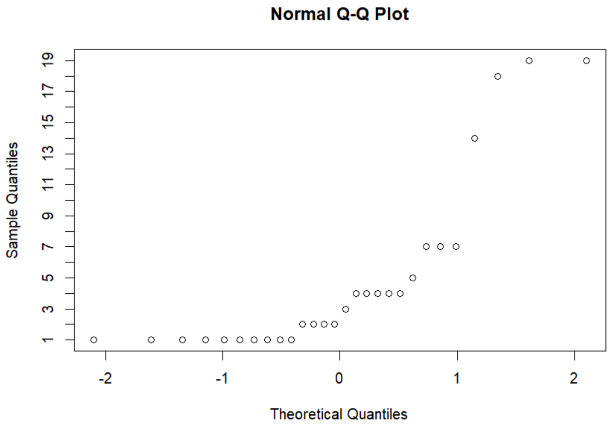

Figure 13: Diagtime Q-Q Norm

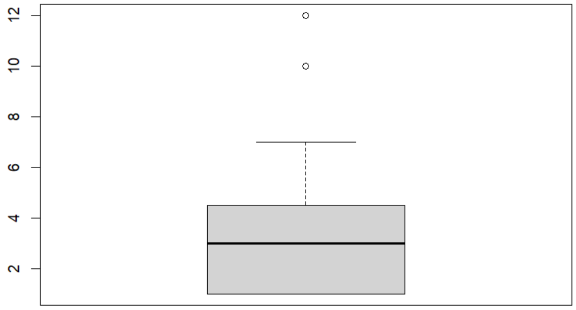

Figure 14: Age Boxplot

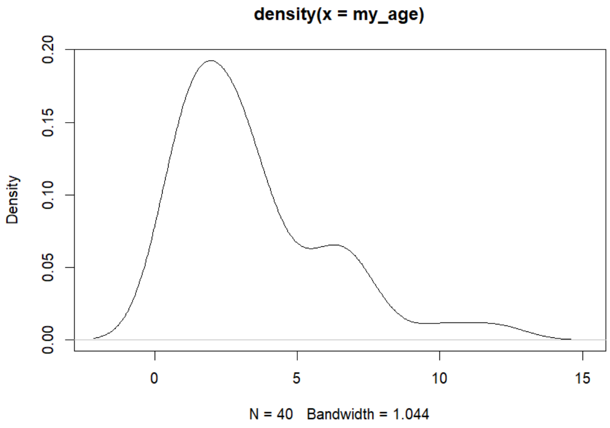

Figure 15: Age Density Plot

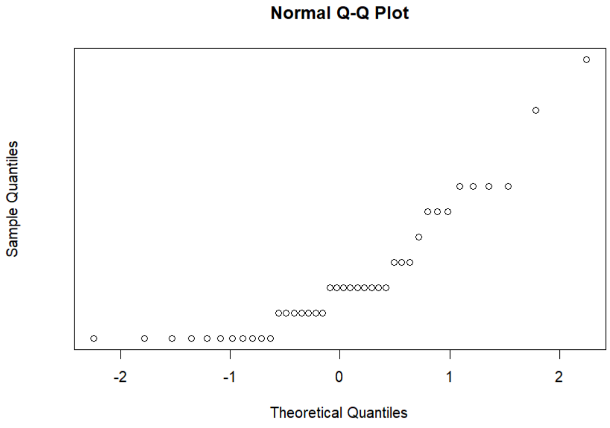

Figure 16: Age Q-Q Norm

2.8 PRIOR

The **prior** variable is numeric, which describes if the subject has received a treatment before the experiment. A better approach would be to use logical value because the problem presented is of type if or if not.

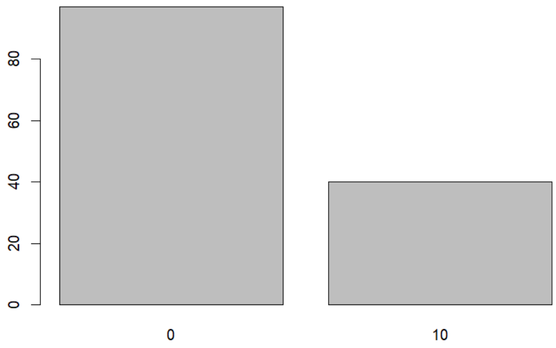

Figure 17: Prior Histogram

3 Statistic Hypotheses

3.1 Greater lifespan for affirmative prior value

We assume that those who have had prior disease treatment have a longer lifespan. We test this claim at a significance level of 0.05.

To conclude this hypothesis, we divided the sample into two subsets on which we applied a t-test, obtaining the following result:

```
treatment <- my_veteran$time[my_veteran$prior == TRUE]
no_treatment <- my_veteran$time[my_veteran$prior == FALSE]

t.test(treatment, no_treatment, mu = 0, alternative = "
    greater")
```

```
    Welch Two Sample t-test
data:  treatment and no_treatment
t = 0.87013, df = 48.959, p-value = 0.1942
alternative hypothesis: true difference in means is
    greater than 0
95 percent confidence interval:
 -30.07086       Inf
```

```
sample estimates:
mean of x mean of y 
 144.6000  112.1546 
```

3.2 Relations between karno and celltype

Grouping the values of the variable **karno** by the type of cells passed in tables, compare the Karnofsky coefficients according to the grouping performed.

To achieve this we performed an ANOVA test, the result being the following:

```
summary(aov(my_veteran$karno ~ my_veteran$celltype))
```

```
                     Df Sum Sq Mean Sq F value Pr(>F)  
my_veteran$celltype   3   2519   839.6   2.143 0.0978 .
Residuals           133  52097   391.7                 
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ 
     ’ 1
```

To see the differences between the groups we also applied the pairwise.t.test function:

```
pairwise.t.test(my_veteran$karno, my_veteran$celltype)
```

```

  Pairwise comparisons using t tests with pooled SD 

data:  my_veteran$karno and my_veteran$celltype 

   sq   sc   ad  
sc 0.49 -    -   
ad 1.00 1.00 -   
lg 1.00 0.10 0.81

P value adjustment method: holm 
```

4 Predictive Model

4.1 Construction and Analysis

We construct a linear regression model describing the relationship between individual age and Karnofsky's performance score. Based on this we want to determine a 95% confidence interval for the Karnofsky index of an individual aged 68.

```
karno <- my_veteran$karno
age <- my_veteran$age
```

```
model  <- lm(karno ~ age)
residuals <- resid(model)
fitted <- fitted(model)
summary(model)
```

```
Call:
lm(formula = karno ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-44.472 -16.819   2.278  17.042  40.195

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.0975     9.6493   7.161 4.67e-11 ***
age          -0.1806     0.1629  -1.109     0.27
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’
     ’ 1

Residual standard error: 20.02 on 135 degrees of freedom
Multiple R-squared:  0.009022,  Adjusted R-squared:
    0.001682
F-statistic: 1.229 on 1 and 135 DF,  p-value: 0.2696
```

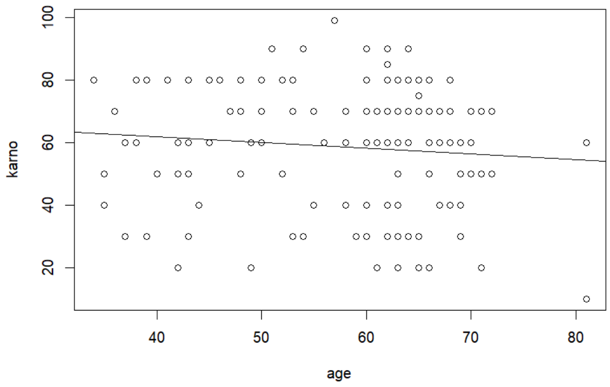

Figure 18: Age and Karno Plot

```
shapiro.test(residuals)
```

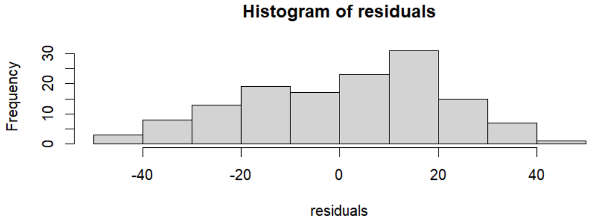

Figure 19: Histogram of Residuals

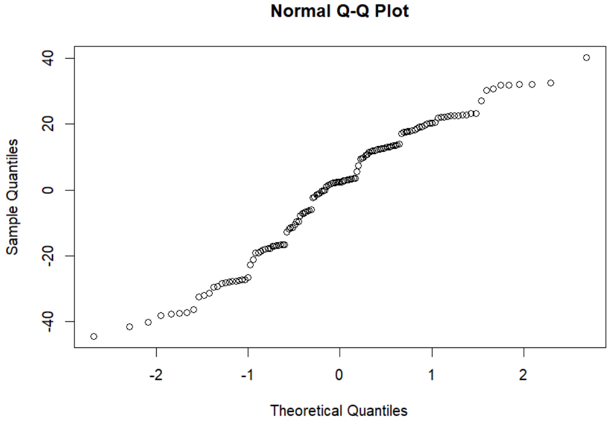

Figure 20: Residuals Q-Q Norm

```
	Shapiro-Wilk normality test

data:  residuals
W = 0.96375, p-value = 0.00106
```

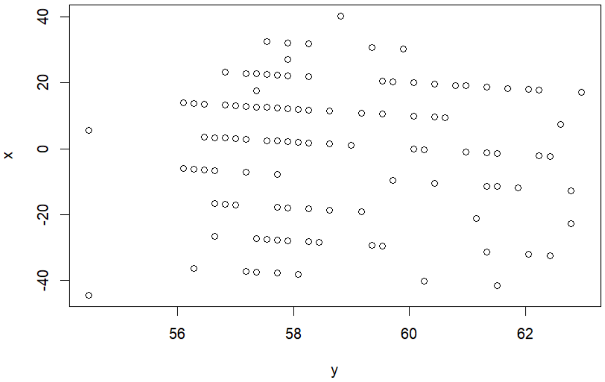

Figure 21: Residuals vs Adjusted Values Plot

4.2 Prediction

```
n_date <- data.frame(age = 68)
predict(model, n_date, interval = "predict", level = 0.95)
```

```
       fit      lwr      upr
1 56.81904 16.95351 96.68457
```