EXPLORATORY DATA ANALYSIS OF THE SURVIVAL FROM MELANOMA DATASET
Modestus Akushie
7CS039 Statistics for AI & Data Science

Introduction

The primary aim of this report is to present an Exploratory Data Analysis conducted with R to gain insight into the “Survival from Malignant Melanoma” dataset, which comprise of measurements from 205 patients diagnosed with Malignant Melanoma. The University Hospital of Odense, Denmark, conducted surgical removal of the malignant tumours, including a margin of approximately 2.5cm of surrounding skin, in patients from 1962 to 1977. Key prognostic variables, such as tumour thickness and ulceration status, were measured due to their significance in predicting melanoma-related mortality. Patients were thereafter observed until 1977.

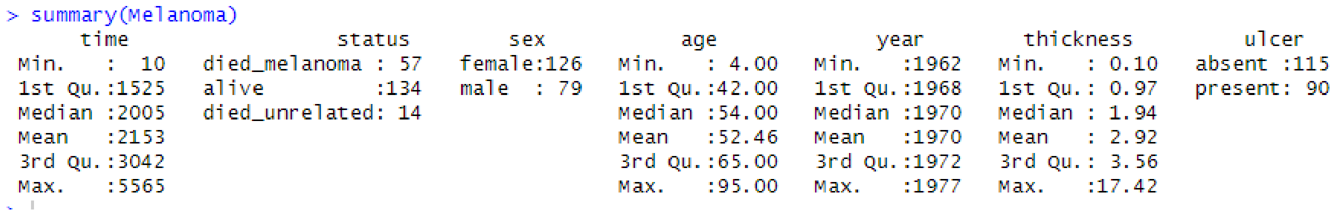
2. Data Summary

The data frame comprises 7 columns with specific variables viz:

- ➢ time - indicates days since the operation, measured in survival time
- ➢ status - This represents the patient's condition at the end of the operation and observation. 1 shows that they had died from melanoma, 2 indicates that they were still alive and 3 denotes death unrelated to melanoma.
- ➢ sex - denotes the patient's gender with 1=male and 0=female.
- ➢ age - indicates the patient’s age in years at the time of the operation.
- ➢ year - represents the year of operation.
- ➢ thickness - Denotes tumour thickness measured in millimetre (mm).
- ➢ ulcer - an indicator of ulceration where 1=present, and 0=absent.

2.1 Numerical Summary

Numerical summaries of every column in the data set are shown below

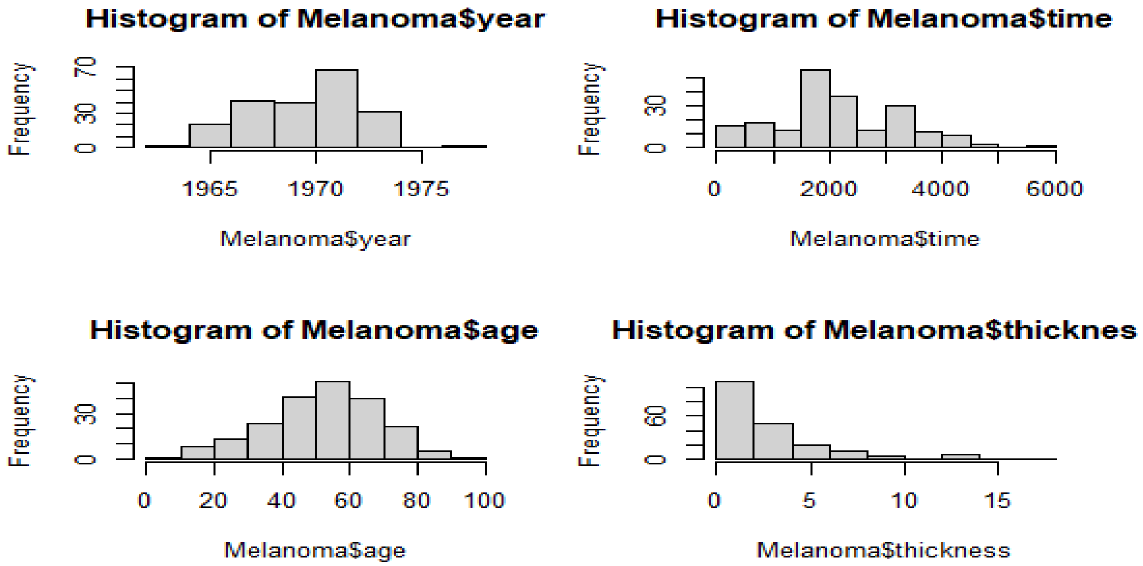


```
> summary(Melanoma)
      time                   status         sex           age             year        thickness         ulcer
 Min.   :  10   died_melanoma : 57   female:126   Min.   : 4.00   Min.   :1962   Min.   : 0.10   absent :115
 1st Qu.:1525   alive         :134   male  : 79   1st Qu.:42.00   1st Qu.:1968   1st Qu.: 0.97   present: 90
 Median :2005   died_unrelated: 14                Median :54.00   Median :1970   Median : 1.94
 Mean   :2153                                     Mean   :52.46   Mean   :1970   Mean   : 2.92
 3rd Qu.:3042                                     3rd Qu.:65.00   3rd Qu.:1972   3rd Qu.: 3.56
 Max.   :5565                                     Max.   :95.00   Max.   :1977   Max.   :17.42
```

Drawing insights from the above summary, it can be deduced from the time variable, that some of the patients died some few days after the surgery for at least a minimum of 10 days while others lived up to a maximum of 15 years. 57 of the total record population died from melanoma, capped at about 27.8% of total from the record taken. We can also see that females were more than the males by 23%. For thickness, the greatest measurement was 17.42, the minimum was 0.10, and we have a median of 1.94 and a mean of 2.92.

2.2 Graphical Summaries

We would be highlighting the summaries of variables in the dataset with some graphical visualisation to derive meaning from their relationship.

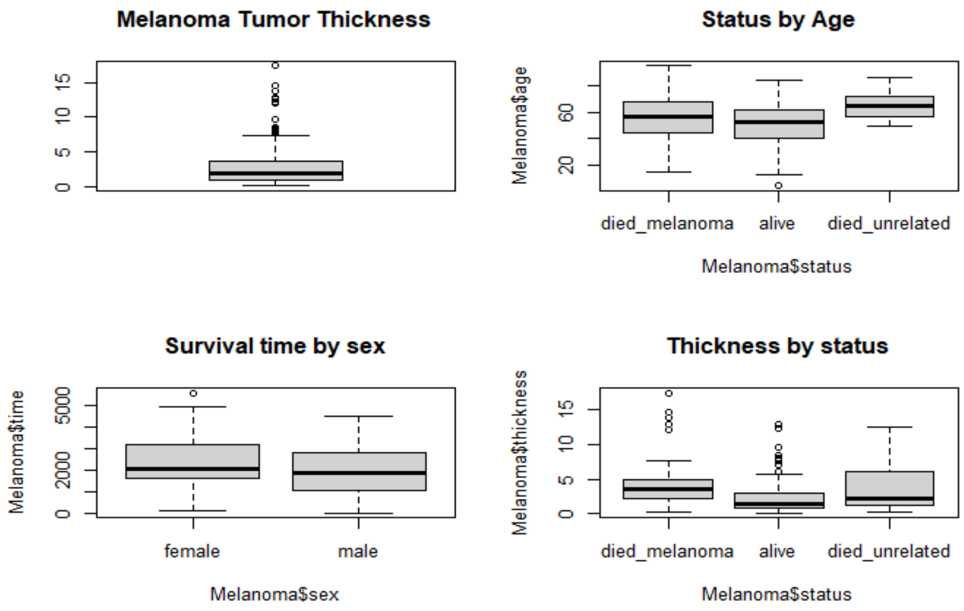


Histogram of year, time, age and thickness

In the above histogram figure, the thickness histogram clearly shows that the distribution is positively skewed while that of age is a normal distribution. The histogram of time however seems to be a normal distribution. These assumptions will be validated later with the Quantile Quantile plot.

Also, the histogram shows records of some missing data in time, year and thickness which could affect the reliability of the data and lead to uncertainty in result findings.

Finally, the distribution of time shows a trend of peak in survival time of patients since the operation between 1500 and 2500 days.



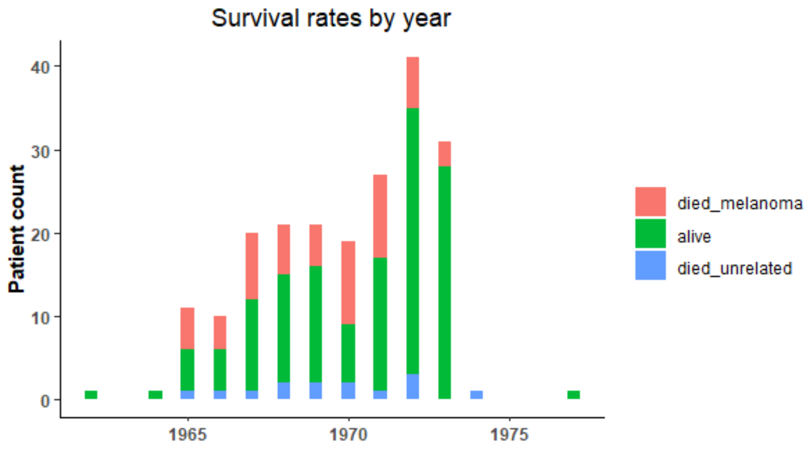
- The boxplot in ‘Melanoma tumor thickness’ shows that most of the recorded values of thickness size falls in the range of 0.10 - 7, leaving those higher as an outlier.

- It also shows in ‘Status by Age’ that there were more people who died than those who survived after the operation. Females had more survival time than males.

- We can also deduce that Thickness is closely associated with a high number of deaths from melanoma.

- With respect to age, most people who died from melanoma were in their middle age.

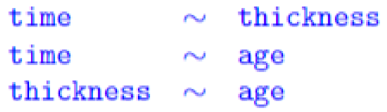
Bar chart of survival rate by year



This particular chart above goes further to illustrate the survival rates of patients by year. It is observed that there were no cases of death until 1965, showing that the first death case was in 1965, and no cases of death after 1973.

Correlation and Regression Analysis

We would be computing the correlation between variables, using the Pearson method to find a linear relationship, and then build linear models to run a regression analysis. Let's therefore, first find the relationship between the following variables

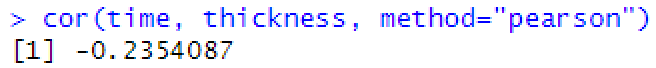


```
time       ~  thickness
time       ~  age
thickness  ~  age
```

Correlation

time ~ thickness

Below is the calculated correlation coefficient result for time and thickness

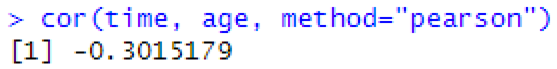


```
> cor(time, thickness, method="pearson")
[1] -0.2354087
```

Therefore, our correlation coefficient (r) between time and thickness is = -0.2354087

time ~ age

We present our calculated correlation coefficient result for time and age as

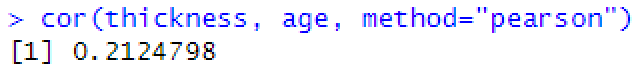


```
> cor(time, age, method="pearson")
[1] -0.3015179
```

For time and age, we therefore have as our coefficient r = -0.3015179

thickness ~ age

The calculated correlation coefficient value for thickness and age is highlighted thus



```
> cor(thickness, age, method="pearson")
[1] 0.2124798
```

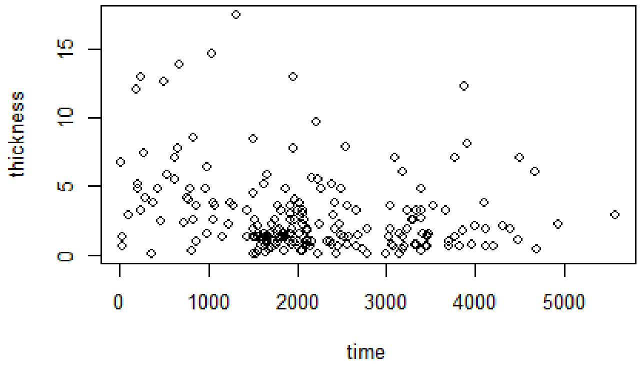
So we have our correlation coefficient r = 0.2124798

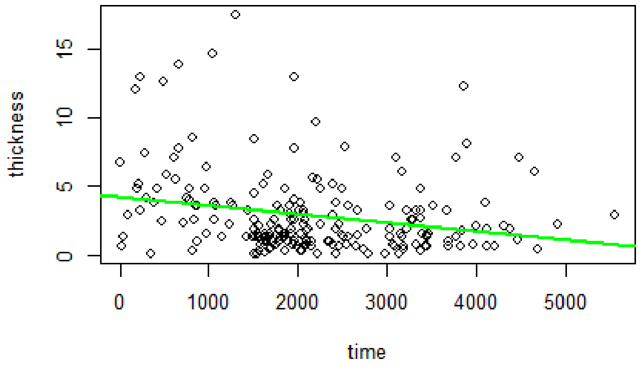
Regression analysis

In conducting our regression analysis, we first construct a scatterplot of

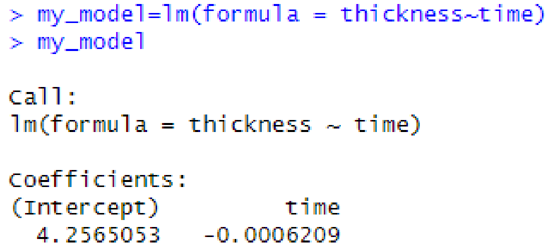
- time (x − axis) vs thickness (y − axis),
- time (x − axis) vs age (y − axis) and
- thickness (x − axis) vs age (y − axis)

time ~ thickness





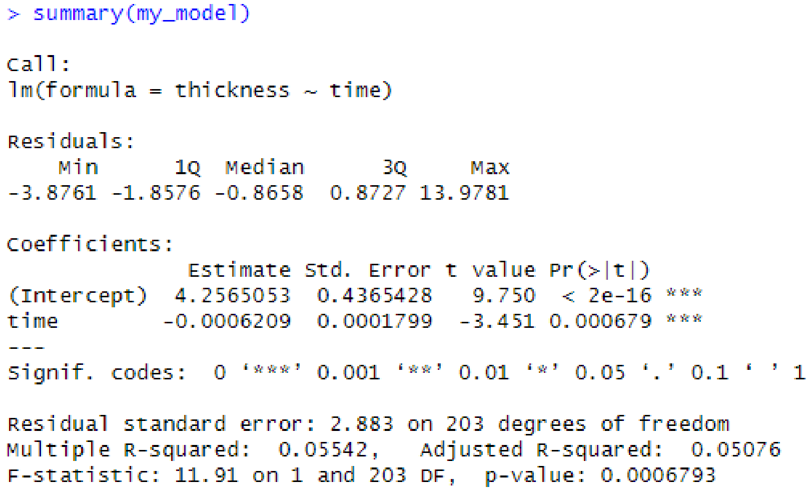
Now constructing our regression model, we employ
my_model = lm(formula = thickness~time) which gives us



```
> my_model=lm(formula = thickness~time)
> my_model

Call:
lm(formula = thickness ~ time)

Coefficients:
(Intercept)         time  
  4.2565053   -0.0006209  
```

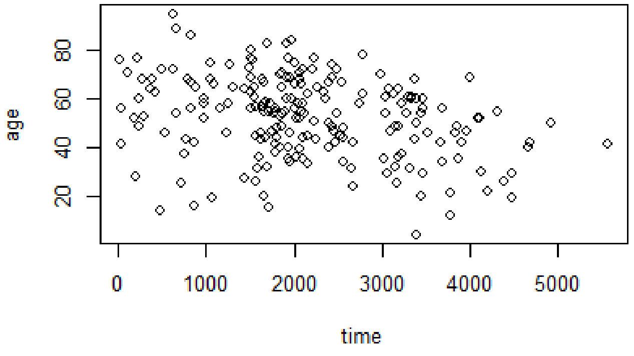
Adapting our regression equation y = mx + b, for our model (where y = thickness, x= time, m = the slope or gradient, and b= intercept)
we have; ***y = -0.00062x + 4.26***

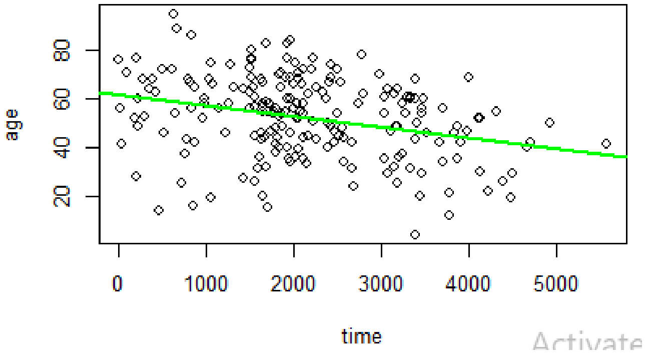
The summary of our overall model performance shows thus;



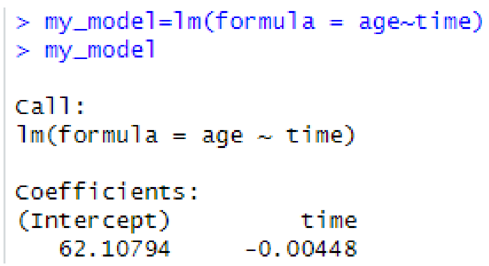
```
> summary(my_model)

Call:
lm(formula = thickness ~ time)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8761 -1.8576 -0.8658  0.8727 13.9781 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.2565053  0.4365428   9.750  < 2e-16 ***
time        -0.0006209  0.0001799  -3.451 0.000679 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.883 on 203 degrees of freedom
Multiple R-squared:  0.05542,	Adjusted R-squared:  0.05076 
F-statistic: 11.91 on 1 and 203 DF,  p-value: 0.0006793
```

time ~ age





Our regression model analysis is depicted below.

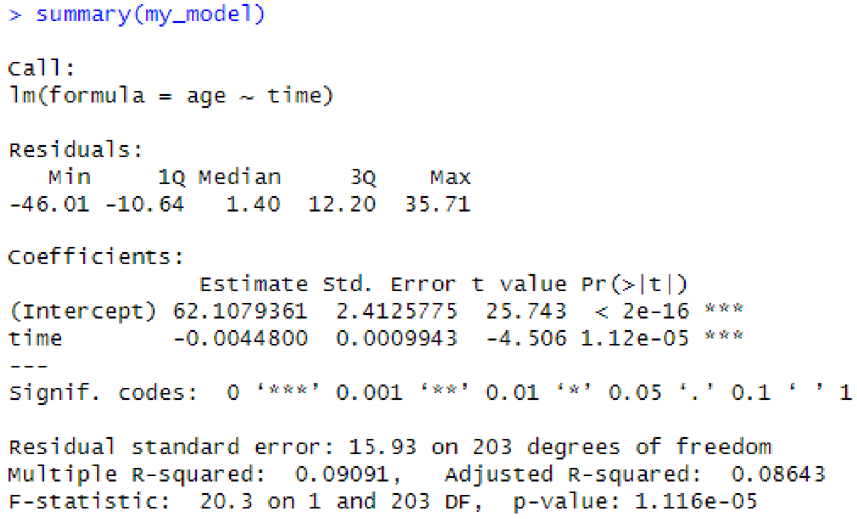


```
> my_model=lm(formula = age~time)
> my_model

Call:
lm(formula = age ~ time)

Coefficients:
(Intercept)         time  
   62.10794     -0.00448  
```

Fitting the coefficients values into our regression equation, we have

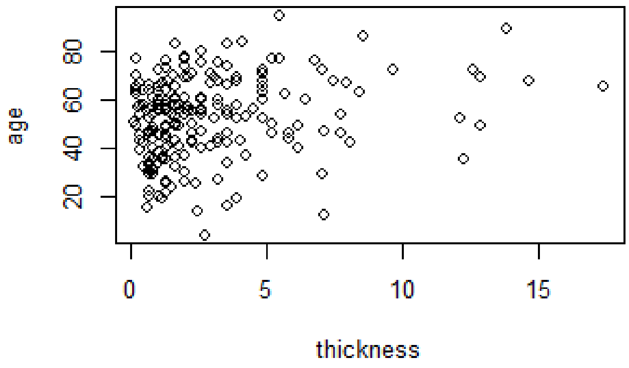
y = -0.00448x + 62.11

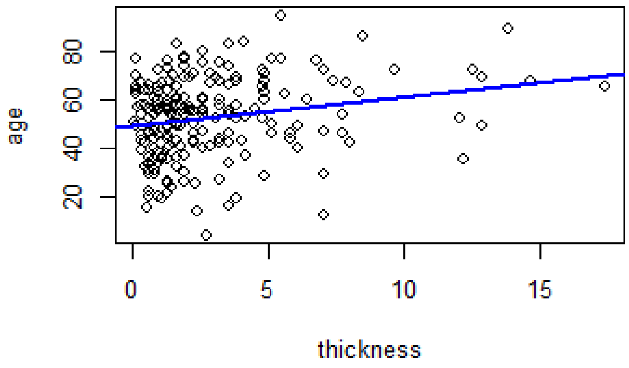
The computed summary of our model overall performance for age ~ time shows thus;



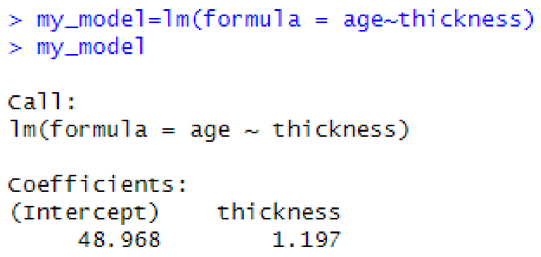
```
> summary(my_model)

Call:
lm(formula = age ~ time)

Residuals:
   Min     1Q Median     3Q    Max 
-46.01 -10.64   1.40  12.20  35.71 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 62.1079361  2.4125775  25.743  < 2e-16 ***
time        -0.0044800  0.0009943  -4.506 1.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.93 on 203 degrees of freedom
Multiple R-squared:  0.09091,	Adjusted R-squared:  0.08643 
F-statistic:  20.3 on 1 and 203 DF,  p-value: 1.116e-05
```

thickness ~ age





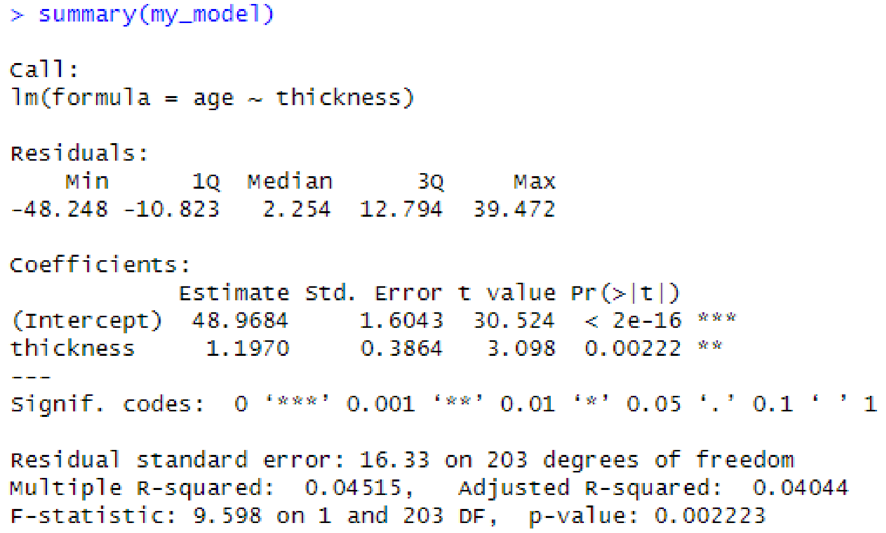
Our regression model analysis is as seen below.



```
> my_model=lm(formula = age~thickness)
> my_model

Call:
lm(formula = age ~ thickness)

Coefficients:
(Intercept)    thickness  
     48.968        1.197  
```

Fitting into our regression equation, we have; ***y = 1.197x + 48.968***

Our summary of model overall performance, therefore, shows



```
> summary(my_model)

Call:
lm(formula = age ~ thickness)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.248 -10.823   2.254  12.794  39.472 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  48.9684     1.6043  30.524  < 2e-16 ***
thickness     1.1970     0.3864   3.098  0.00222 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.33 on 203 degrees of freedom
Multiple R-squared:  0.04515,	Adjusted R-squared:  0.04044 
F-statistic: 9.598 on 1 and 203 DF,  p-value: 0.002223
```

Commentary

Based on the correlation and regression analysis carried out in this exploratory data analysis, we can infer the following relationships:

- For **time ~ thickness**, the obtained coefficient r = -0.2354, indicates a weak negative correlation between the two variables. Furthermore, our R-squared value which is 0.05542, indicates 5.542% of the variability in thickness is explained by time. The model has limited explanatory power. The p-value for time is 0.0006793, indicating that time is a statistically significant predictor of thickness.

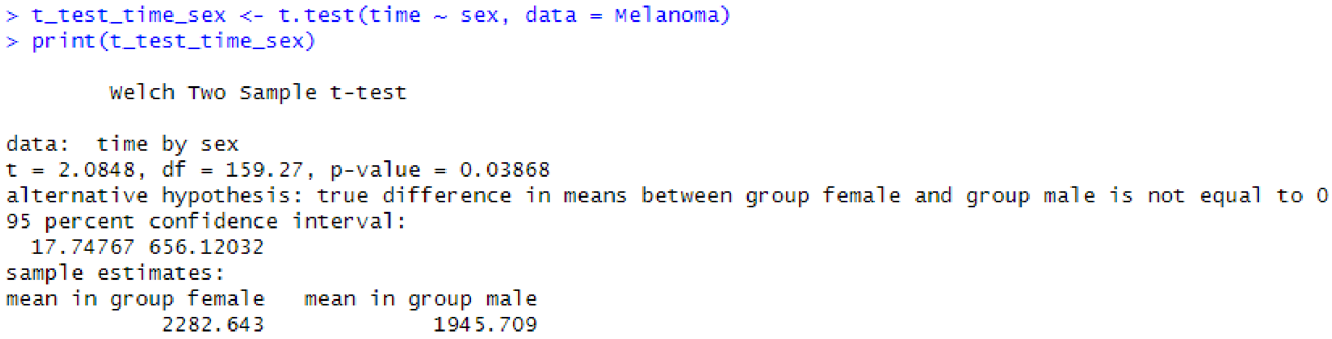
- For **time ~ age**, r = -0.3015, indicates a weak negative correlation between the two variables. Furthermore, our R-squared value which is 0.09091, indicates 9.091% of the variability in age is explained by time. The model has limited explanatory power. The p-value for time is 0.00001116, indicating that time is a statistically significant predictor of Age.

- For **thickness ~ age**, r = 0.2125, indicates a weak positive correlation between the two variables. Furthermore, our R-squared value which is 0.04515, indicates 4.515% of the variability in age is explained by thickness. The model has limited explanatory power. The p-value for Thickness is 0.002223, indicating that Thickness is a statistically significant predictor of Age.

Test of Significance

The results of the Two sample t-test for the 3 variables (time, thickness, age) grouped by sex are shown below:

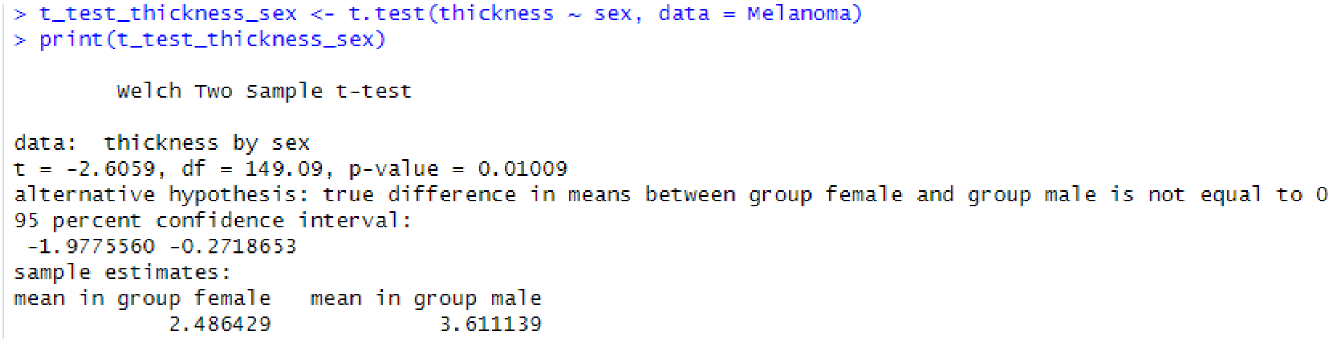
Time by sex



```
> t_test_time_sex <- t.test(time ~ sex, data = Melanoma)
> print(t_test_time_sex)

        Welch Two Sample t-test

data:  time by sex
t = 2.0848, df = 159.27, p-value = 0.03868
alternative hypothesis: true difference in means between group female and group male is not equal to 0
95 percent confidence interval:
  17.74767 656.12032
sample estimates:
mean in group female   mean in group male 
            2282.643             1945.709 
```

The p-value (0.03868) is below the significance level (α = 0.05), leading to the rejection of the null hypothesis (H0). Gender-based differences in real mean time are supported by the data.

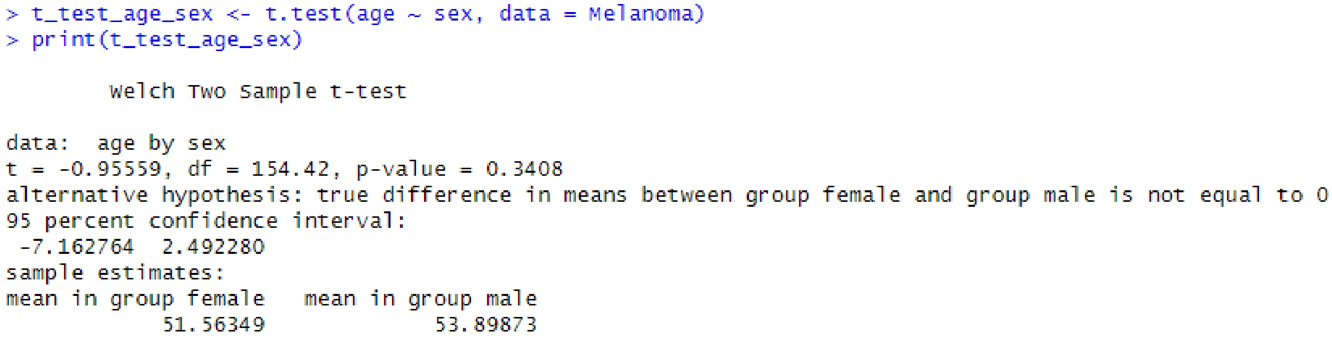
Thickness by sex



```
> t_test_thickness_sex <- t.test(thickness ~ sex, data = Melanoma)
> print(t_test_thickness_sex)

        Welch Two Sample t-test

data:  thickness by sex
t = -2.6059, df = 149.09, p-value = 0.01009
alternative hypothesis: true difference in means between group female and group male is not equal to 0
95 percent confidence interval:
 -1.9775560 -0.2718653
sample estimates:
mean in group female   mean in group male 
            2.486429             3.611139 
```

We gather here that our p-value (0.01009) is much lower than the default significance level of α = 0.05. We can therefore rule out hypothesis H0 and conclude that gender-based differences in real mean thickness are backed by the data.

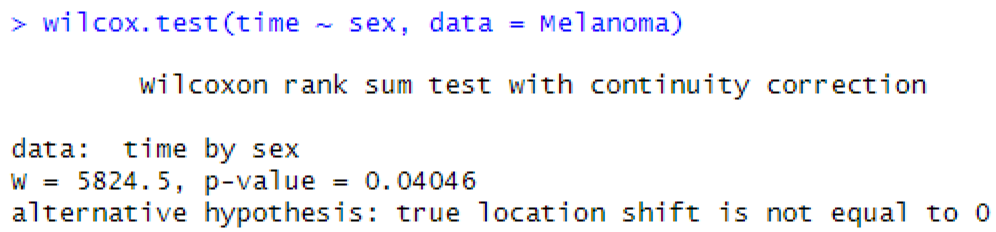
Age by sex



```
> t_test_age_sex <- t.test(age ~ sex, data = Melanoma)
> print(t_test_age_sex)

        Welch Two Sample t-test

data:  age by sex
t = -0.95559, df = 154.42, p-value = 0.3408
alternative hypothesis: true difference in means between group female and group male is not equal to 0
95 percent confidence interval:
 -7.162764  2.492280
sample estimates:
mean in group female   mean in group male 
            51.56349             53.89873 
```

Here, our p-value (0.3408) is significantly larger than the default level of significance of α = 0.05. Therefore, We may rule out hypothesis H0 and conclude that gender-based differences in real mean age are backed by the data.

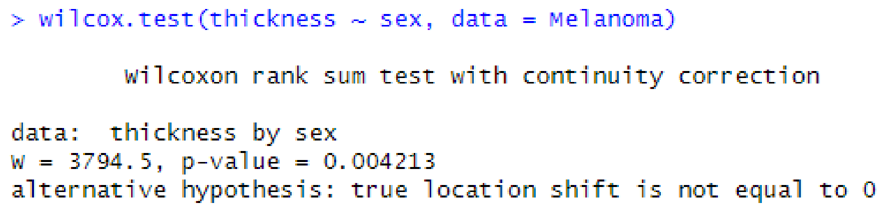
Wilcox test : **Time by sex**



```
> wilcox.test(time ~ sex, data = Melanoma)

        Wilcoxon rank sum test with continuity correction

data:  time by sex
W = 5824.5, p-value = 0.04046
alternative hypothesis: true location shift is not equal to 0
```

The p-value (0.04046) is below the significance level (α = 0.05), leading to the rejection of the null hypothesis (H0). Gender-based differences in real median time are supported by the data.

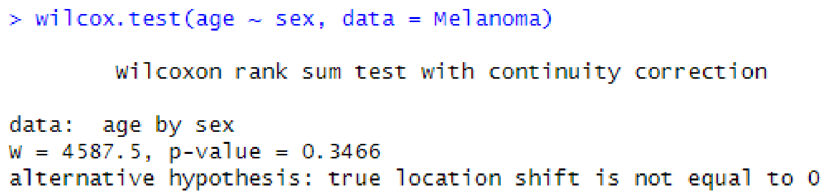
Wilcox test : **Thickness by sex**



```
> wilcox.test(thickness ~ sex, data = Melanoma)

        Wilcoxon rank sum test with continuity correction

data:  thickness by sex
W = 3794.5, p-value = 0.004213
alternative hypothesis: true location shift is not equal to 0
```

The p-value (0.004213) is below the significance level (α = 0.05), leading to the rejection of the null hypothesis (H0). Gender-based differences in real median thickness are supported by the data.

Wilcox test : **Age by sex**



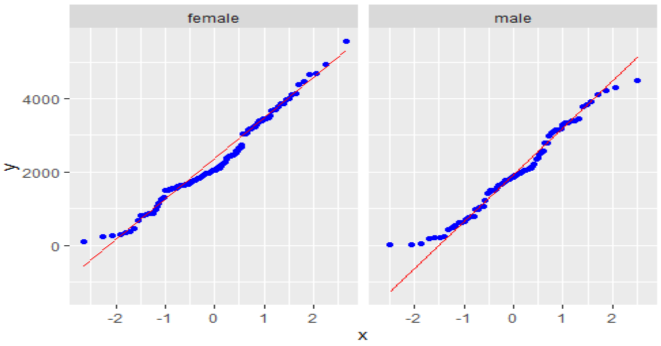
```
> wilcox.test(age ~ sex, data = Melanoma)

        Wilcoxon rank sum test with continuity correction

data:  age by sex
W = 4587.5, p-value = 0.3466
alternative hypothesis: true location shift is not equal to 0
```

The p-value (0.3466) is below the significance level (α = 0.05), leading to the rejection of the null hypothesis (H0). Gender-based differences in real median age are supported by the data.

QQ-plots for the 3 variables grouped by gender

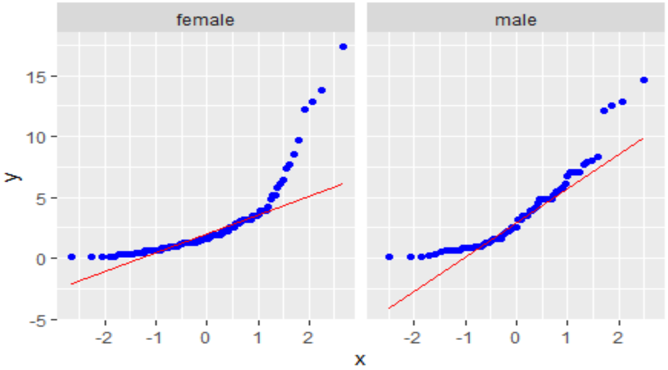
We will apply a conventional QQ-plot (quantile-quantile plot) for each of the 3 variables classified by gender, to confirm whether our data originates from a normally distributed population

Time by sex



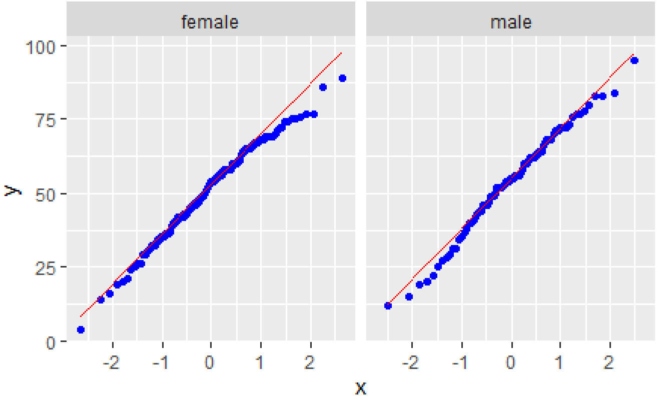
Seeing that the data points fall close to the reference line indicates that the data approximately resembles the expected distribution (normal distribution)

Thickness by sex



Seeing that the data points curve away from the reference line, we can deduce that our data is not normally distributed

Age by sex



Seeing that the data points fall closely to the reference line, we can infer that the data resembles the expected normal distribution.

Summary

Data Overview: The dataset comprises measurements on patients with malignant melanoma who underwent surgery between 1962 and 1977. Variables include survival time (time), patient status (status), sex, age, year of operation, tumor thickness (thickness), and ulcer presence (ulcer).

Numerical Summary: The summary statistics highlight the range of measurements, indicating potential variability and skewness in the data.

Graphical Summary: Histograms reveal the right skewness of the thickness variable, suggesting a majority of patients with smaller thickness. Boxplot projects that tumor thickness is closely associated with a high number of deaths from the melanoma operation, while bar chart shows the survival and death rates of patients after the operation by year

Regression Analysis: Relationships between time and thickness, time and age, and thickness and age are explored. Weak correlations are observed in all cases, with low R-squared values indicating limited explanatory power of the models.

Significance Tests: Two-sample t-tests reveal significant differences in mean time and thickness between genders. However, no significant difference is observed in mean age.

QQ plots suggest that the data for time, and age except thickness, grouped by gender, are approximately normally distributed.

Recommendation

Based on the findings, from the analysis of the "Survival from Malignant Melanoma" dataset, it is recommended that:

I. Given the presence of probable outliers in tumor thickness measures, additional examination into these cases which could reveal insights into their importance should be carried. Gaining insight into the causes of extreme numbers could help improve the predictive models and refine the analysis to reinforce that tumor thickness is the most important and predictive factor in malignant melanoma (Nield *et al.*, 1988)

II. The gaps in the dataset found, especially in the time, years, and thickness indicators, should be closed. Ensuring data completeness will increase the credibility of studies and allow a more thorough knowledge of the factors determining survival from malignant melanoma.

III. An incorporation of extra variables pertaining to patients' genetic makeup like blood group and genotype, medical histories, or treatment schedules may provide more understanding and solution for melanoma survival. This could result in more precise forecasts and a better comprehension of the illness.

IV. Relevant authorities and medical institutions like the World Health Organization should advance and fund genetic studies to explore the genetic factors influencing melanoma outcomes, especially in correlation with gender and age. This can provide deeper insights into the biological makeup and processes driving the observed disparities.

References

Nield, D.V. *et al.* (1988) 'Tumour thickness in malignant melanoma: the limitations of frozen section,' *British Journal of Plastic Surgery*, 41(4), pp. 403–407. https://doi.org/10.1016/0007-1226(88)90082-3.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

RStudio Team, (N/A) RStudio: Integrated Development Environment for R. Available at: https://www.rstudio.com/

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer. Retrieved from https://ggplot2.tidyverse.org

Wilke, C. O. (2019). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. Retrieved from https://CRAN.R-project.org/package=cowplot

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., ... & Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. Retrieved from https://doi.org/10.21105/joss.01686

Wickham, H. and Grolemund, G., (2017) R for Data Science. O'Reilly Media. Available at: https://r4ds.hadley.nz/ (Accessed: 2024-01-12).