Exploratory Analysis of the 1977 Survival from Malignant Melanoma University of Odense Denmark Data Set

By

Opeyemi Olaosebikan

University of Wolverhampton

Student No: 2338168

Abstract

“As to the remote and exciting causes of melanosis, we are quite in the dark, nor can more be said of the methodus medendi*. We are hence forced to confess the incompetency of our knowledge of the disease under consideration, and to leave to future investigators the merit of revealing the laws which govern its origin and progress….and pointing out the means by which its ravages may be prevented or repressed”* - Thomas Fawdington, The Manchester Royal Infirmary, 1826.

Introduction

This report is an exploratory analysis of data collected from patients who had surgical removal of a type of skin cancer tumor known as “Malignant Melanoma” at the Department of Plastic Surgery, University Hospital of Odense, Denmark from 1962 to 1977. The surgery included the complete removal of the cancer tumor and 2.5cm (about 0.98 in) of the surrounding skin to ensure the removal of all affected cells. Patients were observed for 15 years, and a comparative analysis of measurements from patients with malignant melanoma were considered to find the correlation with the likelihood of death.

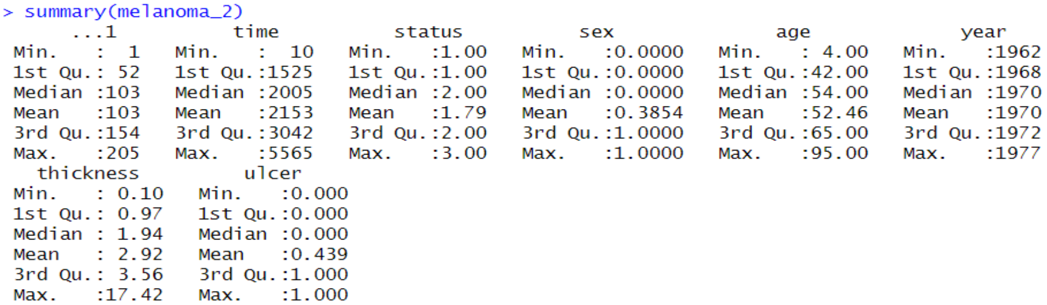
1.1 The 1977 Malignant Melanoma Dataset

Among the measurements taken were the thickness of the tumor in ‘mm’, ulcerated tumors were recorded categorical variable ‘1’ and non-ulcerated with “0”.These are thought to be important prognostic variables in that patients with a thick and/or ulcerated tumor have an increased chance of death from melanoma. Other data collected included the age and sex of the patient, the year the operation was carried out, ‘Survival time in days’ since the operation, and the status of patients at the end of the study in 1977, ‘1’ indicating that they had died from melanoma, ‘2’ indicates that they were still alive and ‘3’ indicates that they had died from causes unrelated to melanoma.

2 Exploratory Data Analysis

In this section we explore numerical statistical and graphical summaries for each variable in the dataset.

2.1 Numerical Summary



```
> summary(melanoma_2)
      ...1           time          status           sex               age             year     
 Min.   :  1   Min.   :  10   Min.   :1.00   Min.   :0.0000   Min.   : 4.00   Min.   :1962  
 1st Qu.: 52   1st Qu.:1525   1st Qu.:1.00   1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1968  
 Median :103   Median :2005   Median :2.00   Median :0.0000   Median :54.00   Median :1970  
 Mean   :103   Mean   :2153   Mean   :1.79   Mean   :0.3854   Mean   :52.46   Mean   :1970  
 3rd Qu.:154   3rd Qu.:3042   3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:65.00   3rd Qu.:1972  
 Max.   :205   Max.   :5565   Max.   :3.00   Max.   :1.0000   Max.   :95.00   Max.   :1977  
   thickness         ulcer      
 Min.   : 0.10   Min.   :0.000  
 1st Qu.: 0.97   1st Qu.:0.000  
 Median : 1.94   Median :0.000  
 Mean   : 2.92   Mean   :0.439  
 3rd Qu.: 3.56   3rd Qu.:1.000  
 Max.   :17.42   Max.   :1.000  
```

Variable	Standard Deviation	Variance
Time	1122.061	1259020
Status	0.5512041	0.3038259
Sex	0.487873	0.2380201
Age	16.67171	277.946
Year	2.575563	6.633525
Thickness	2.959433	8.758242
Ulcer	0.4974829	0.2474892

The variable "time" exhibits a mean duration of 2153 days, equivalent to approximately 6 years, with a notable range from a minimum of 10 days to a maximum of 5565 days (approximately 15 years). The large standard deviation of 1122.061 and a variance of 1259020 highlight considerable variability around the mean. It's important to note that, through visualization, it becomes evident that only one surgery was conducted in 1962, explaining the maximum duration of 5565 days, indicating the patient remained alive at the study's conclusion.

The "status" variable, with a mean of 1.79, suggests that a majority of patients survived until the end of the study, given the variable's scale where 2 indicates survival at end of study. The low standard deviation of 0.5512041 indicates a clustering of values around the mean.

The mean age of patients at the time of surgery is 52.46 years, with a standard deviation of 16.6 years. The age variable spans from a minimum of 4 years to a maximum of 95 years, reflecting a diverse age range among the patients.

For the "year" variable, denoting the year of surgery, the mean is 1970, with a standard deviation of 2.57. The surgeries span from the earliest year in 1962 to the latest in 1977.

Regarding "thickness," the mean thickness is 2.92 mm, with a standard deviation of 2.95 mm, indicating a diverse range of cases. The thickness values range from 0.1 mm to 17.42 mm.

The "sex" variable, with a mean of 0.3854, suggests that a higher proportion of patients were female. Meanwhile, the "ulcer" variable, with a mean of 0.44, implies that a majority of patients did not exhibit ulceration.

In summary, the numerical summaries provide a comprehensive understanding of the central tendencies, variabilities, and ranges within each variable in the dataset, laying the foundation for further analysis and interpretation.

2.2 Graphical Summary

In our exploration of the dataset, we utilized various graphical representations such as histograms, boxplots, and pie charts to visually inspect and extract meaningful patterns and insights.

Survival Time Histogram: The histogram depicting survival time after surgery reveals a notable concentration of patients with a survival duration ranging from approximately 1500 to 2500 days. Interestingly, a distinct gap is observed between the 5000 to 5500-day mark.

Status Distribution Pie Chart: A pie chart was employed to illustrate the distribution of patient status at the conclusion of the study in 1977. Majority of patients were found to be alive, constituting a significant portion of the dataset. Further analysis of the status distribution indicated that 134 patients were still alive, 57 had succumbed to melanoma, and 14 had passed away due to unrelated causes. This visualization effectively communicates the overall survival outcomes within the studied cohort.

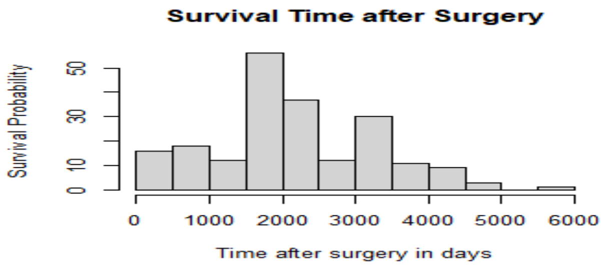
Age Distribution Boxplot: An exploration of the age distribution among patients was conducted using a boxplot. The results unveiled a relatively even distribution of ages, suggesting a diverse representation of age groups within the dataset. Additionally, the identification of an outlier significantly below the first quartile, indicates the presence of a patient with an age considerably lower than the rest of the cohort. This outlier warrants further investigation to discern its potential impact on the overall dataset.

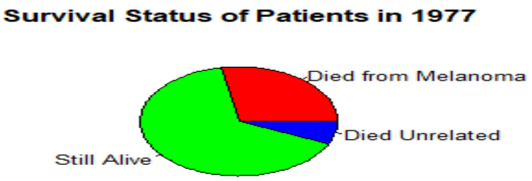
Gender Distribution Boxplot: As evident from the boxplot illustrating gender distribution, a higher proportion of female patients underwent treatment for melanoma compared to their male counterparts. Specifically, 126 female patients underwent cancer tumor removal surgery, nearly double the number of males, which stood at 79.

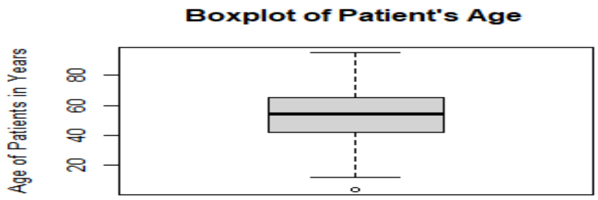
Year of Surgery LineChart: Examining the line chart depicting the distribution of surgeries over the years, it becomes apparent that a significant number of surgeries were performed between 1968 and 1972. The peak frequency occurred within the years 1970 to 1972, while no surgeries were conducted between 1974 and 1976. The year 1977 recorded the lowest surgical intervention, aligning with the conclusion of the study.

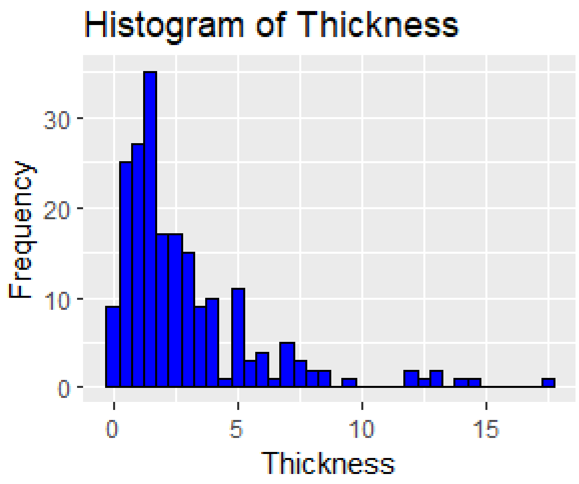
Ulceration Histogram : An exploration of the presence of ulcers in the removed tumors indicates that a higher count of tumors lacked ulcers compared to those with ulcers. This visual representation sheds light on the prevalence of ulcers in the tumors subjected to removal procedures.

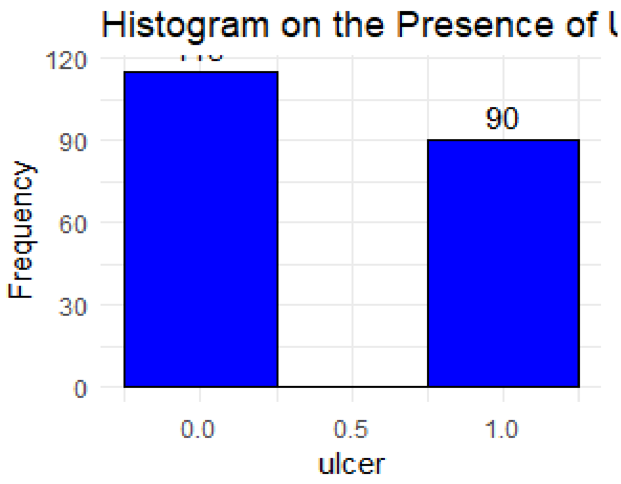
In conclusion, our graphical summaries have provided valuable insights into the survival time, status distribution, gender distribution, surgery frequency over the years, and the presence of ulcers in tumors and age composition of the melanoma patient dataset. These visualizations serve as indispensable tools for researchers and clinicians alike, offering a clearer understanding of the dataset's characteristics and paving the way for more in-depth analyses and informed decision-making in the field of melanoma research and patient care.

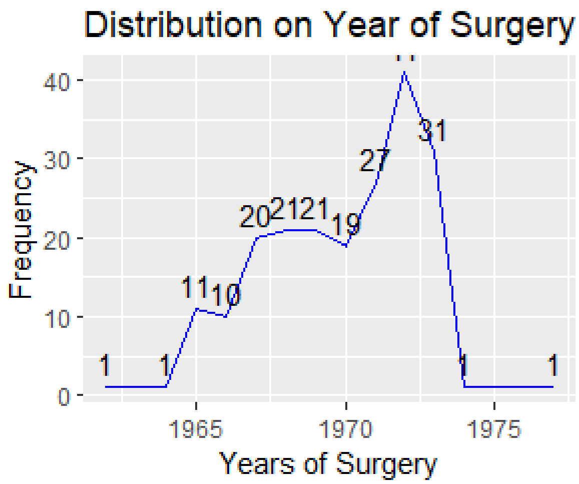


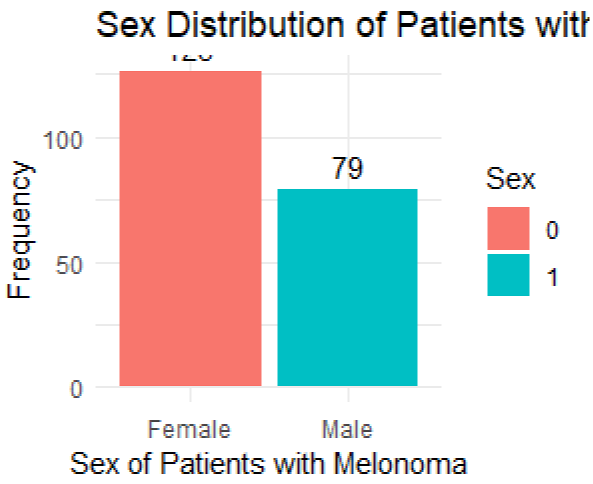












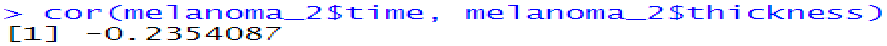
3 Regression and Correlation Analysis

In this section we want to find out if a linear relationship exists between Age, thickness, and time variables, how strong are those relationships, and what the relationships are if they exist

3.1 Time and Thickness

In this section, we will be looking to see if there is any linear relationship between time and thickness of the tumor and then calculating how strong it is if so.

3.1.1 Correlation Analysis: Time ~ Thickness

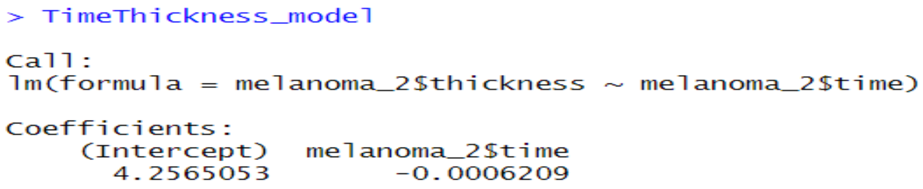


```
> cor(melanoma_2$time, melanoma_2$thickness)
[1] -0.2354087
```

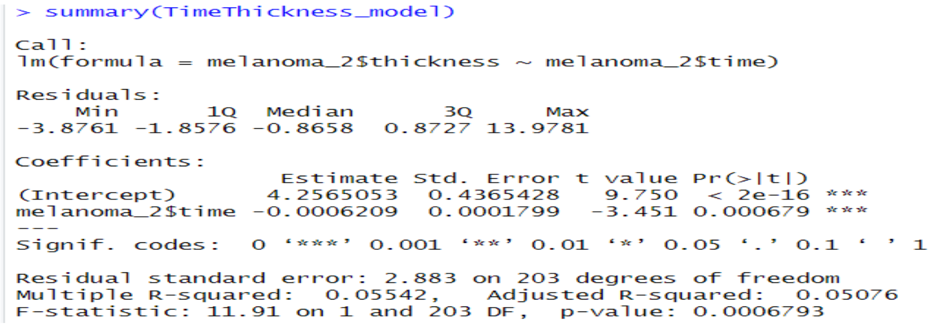
This indicates a small negative correlation between the survival time after surgery and the tumor removed thickness. This means that as one of the variables increases the other decreases, we can suggest that sometimes the larger the thickness of the tumor removed the shorter the survival time.

3.1.2 Regression

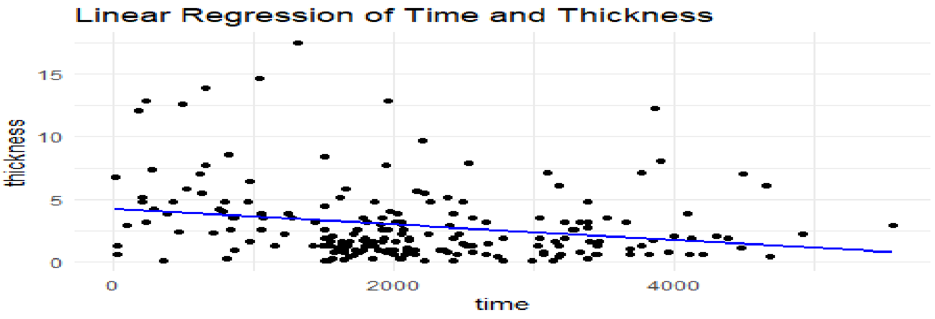
In our regression analysis, we examined the relationship between the "time" and "thickness" variables through a scatterplot. The regression equation obtained is expressed as Y = (-0.0006209)X + 4.2565053, where Y represents the dependent variable (time), X represents the independent variable (thickness), -0.0006209 is the coefficient for thickness, and 4.2565053 is the intercept term. This equation serves as the basis for predicting the time variable based on the thickness variable



```
> TimeThickness_model

Call:
lm(formula = melanoma_2$thickness ~ melanoma_2$time)

Coefficients:
    (Intercept)  melanoma_2$time  
      4.2565053       -0.0006209  
```



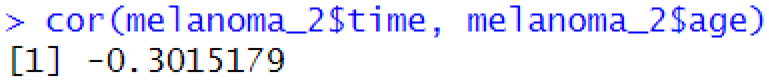
```
> summary(TimeThickness_model)

Call:
lm(formula = melanoma_2$thickness ~ melanoma_2$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 ***
melanoma_2$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
```



The depicted figure illustrates a negative correlation between the survival time after surgery and the thickness of the tumor. This implies that as the thickness of the tumor increases, the survival time tends to decrease.

3.2 Time and Age

3.2.1 Correlation Analysis: Time ~ Age



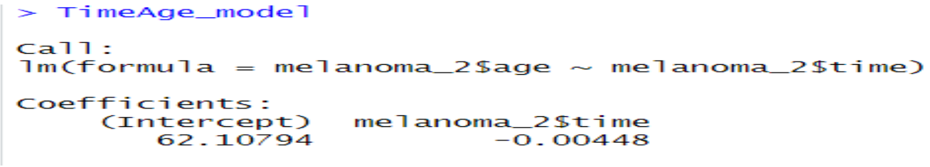
```
> cor(melanoma_2$time, melanoma_2$age)
[1] -0.3015179
```

The magnitude of the correlation coefficient (-0.2354) is relatively small, indicating a weak negative linear relationship between the two variables. In practical terms, this suggests that there is a tendency for patients with thicker melanomas to have shorter survival times, but the relationship is not strong

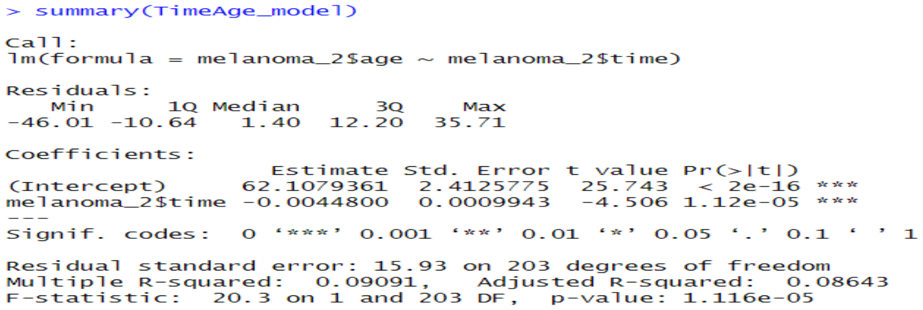
3.2.2 Regression

The regression analysis, as observed through the scatterplot depicting the relationship between the time and thickness variables, yields a regression model represented by the equation Y = (-0.00448) + 62.10794. This equation encapsulates the linear relationship between the variables, where Y represents the dependent variable (survival time after surgery), and X represents the independent variable (thickness of the tumor)

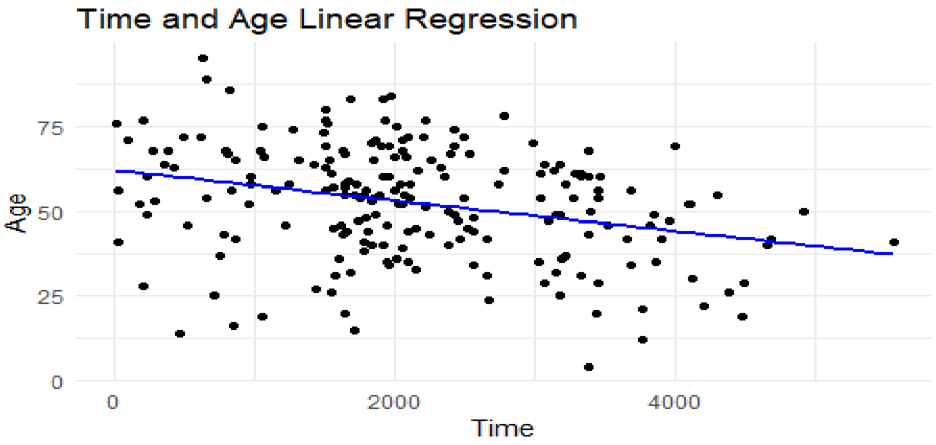
The presented scatterplot below illustrates a negative correlation between the survival time after surgery and the age of the patient. The downward slope indicates as the age increases, the survival time tends to decrease. Additionally, the scatterplot reveals individual patient data points, highlighting an outlier in age that corresponds to a distinct survival time. This finding challenges the assumption of prolonged survival beyond the observed period, suggesting potential outliers influencing the overall



```
> TimeAge_model

Call:
lm(formula = melanoma_2$age ~ melanoma_2$time)

Coefficients:
    (Intercept)  melanoma_2$time  
       62.10794         -0.00448  
```

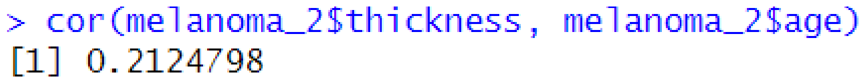


```
> summary(TimeAge_model)

Call:
lm(formula = melanoma_2$age ~ melanoma_2$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 ***
melanoma_2$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
```



3.3 Thickness and Age

3.3.1 Correlation Analysis Thickness ~ Age

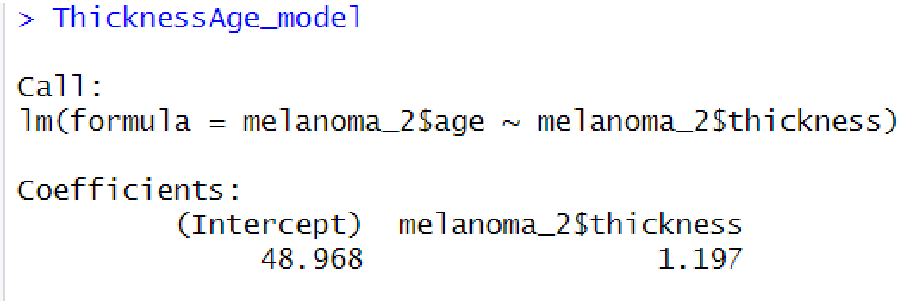


```
> cor(melanoma_2$thickness, melanoma_2$age)
[1] 0.2124798
```

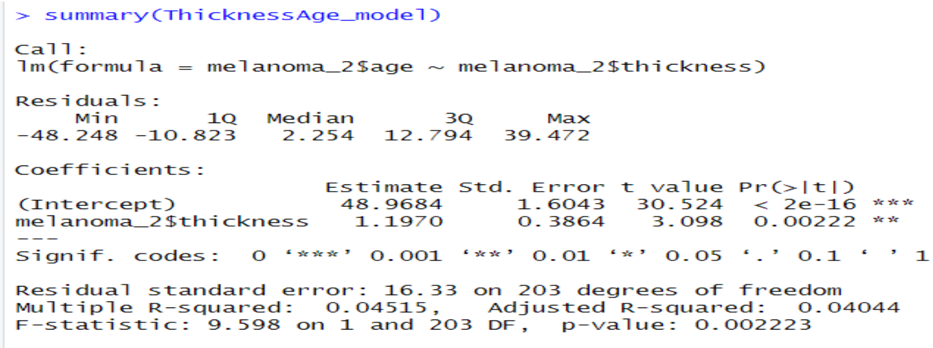
The magnitude of the correlation coefficient (0.2125) is relatively small, indicating a weak positive linear relationship between the two variables. In practical terms, this suggests older patients tend to have slightly thicker melanomas, but the relationship is not strong.

3.3.2 Regression

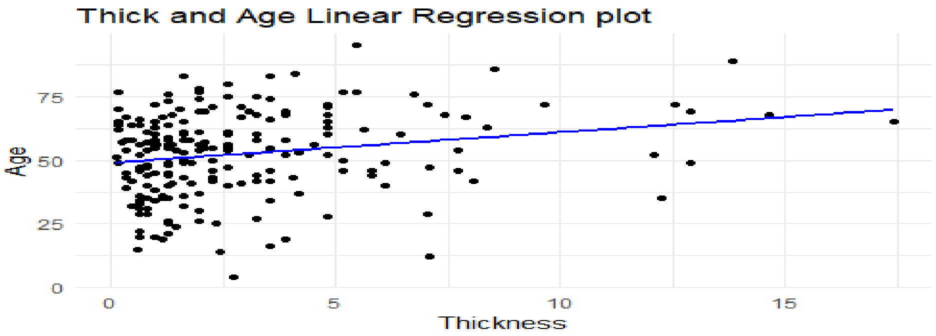
The scatterplot depicting the relationship between thickness and age variables reveals a positive correlation. As thickness increases, the age of the patients tends to rise as well. Our regression model, formulated as y = 1.197 + 48.968, quantifies this relationship. Specifically, for every unit increase in thickness, the predicted age of the patient is expected to increase by 48.968 units.



```
> ThicknessAge_model

Call:
lm(formula = melanoma_2$age ~ melanoma_2$thickness)

Coefficients:
         (Intercept)  melanoma_2$thickness
              48.968                 1.197
```



```
> summary(ThicknessAge_model)

Call:
lm(formula = melanoma_2$age ~ melanoma_2$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 ***
melanoma_2$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
```



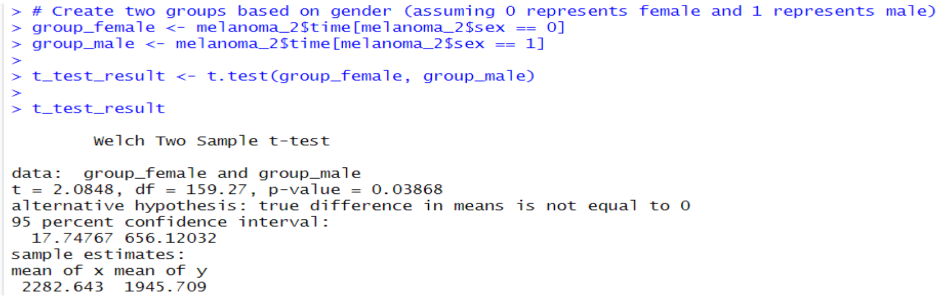
4 Two Sample Significance Test

In this section, we conducted two-sample significance tests to compare the mean of the variables time, thickness, and age between male and female groups using a t-test.

The null hypothesis (H0) states that there is no difference in the mean of variables between males and females, while the alternative hypothesis (H1) suggests that there is a difference

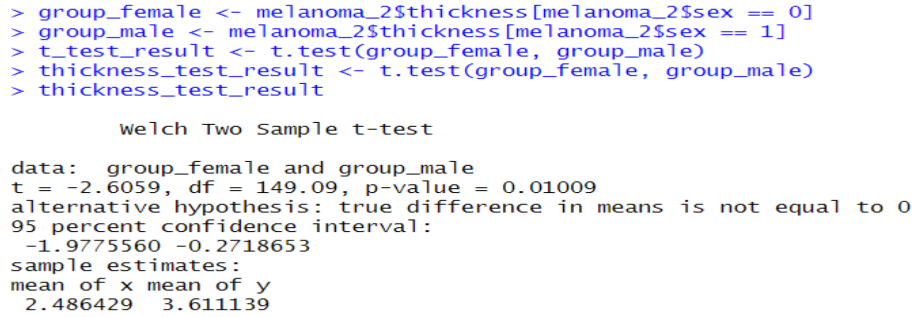
4. 1 T-Test on Time Variable grouped by Gender

The result from the T-Test of gender based on the survival time of patients has a p-value of 0.03868, the mean survival time for female patients is 2282.643, while the mean survival time for male patients is 1945.709. Given a P- Value lower than 0.05 indicates we will reject the null hypothesis that survival time in both males and females are the same.



```
> # Create two groups based on gender (assuming 0 represents female and 1 represents male)
> group_female <- melanoma_2$time[melanoma_2$sex == 0]
> group_male <- melanoma_2$time[melanoma_2$sex == 1]
>
> t_test_result <- t.test(group_female, group_male)
>
> t_test_result

        Welch Two Sample t-test

data:  group_female and group_male
t = 2.0848, df = 159.27, p-value = 0.03868
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  17.74767 656.12032
sample estimates:
mean of x mean of y
 2282.643  1945.709
```

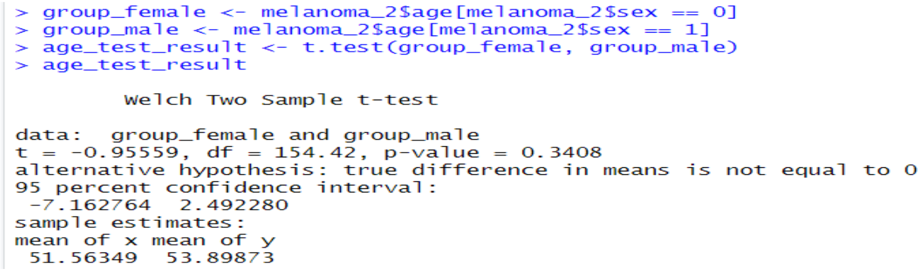
4.2 T- Test on Thickness of Tumor grouped by gender



```
> group_female <- melanoma_2$thickness[melanoma_2$sex == 0]
> group_male <- melanoma_2$thickness[melanoma_2$sex == 1]
> t_test_result <- t.test(group_female, group_male)
> thickness_test_result <- t.test(group_female, group_male)
> thickness_test_result

        Welch Two Sample t-test

data:  group_female and group_male
t = -2.6059, df = 149.09, p-value = 0.01009
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.9775560 -0.2718653
sample estimates:
mean of x mean of y
 2.486429  3.611139
```

In the T-test conducted on the thickness of tumors grouped by gender, the obtained p-value is 0.01009, which is less than the significance level of 0.05. The mean thickness for female patients is 2.486429, while for male patients, it is 3.611139. This outcome indicates a substantial difference in tumor thickness between female and male patients in the dataset. Consequently, we reject the hypothesis that the thickness of tumors is equal regardless of gender.

4.3 T- Test on Age grouped by Gender



```
> group_female <- melanoma_2$age[melanoma_2$sex == 0]
> group_male <- melanoma_2$age[melanoma_2$sex == 1]
> age_test_result <- t.test(group_female, group_male)
> age_test_result

        Welch Two Sample t-test

data:  group_female and group_male
t = -0.95559, df = 154.42, p-value = 0.3408
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.162764  2.492280
sample estimates:
mean of x mean of y
 51.56349  53.89873
```

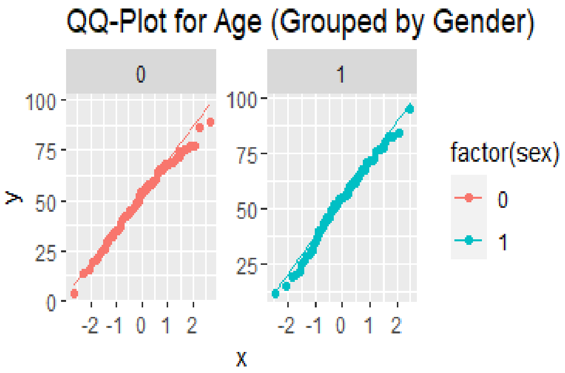
The T-test conducted on the age variable grouped by gender yields a p-value of 0.3408, which exceeds the significance level of 0.05. Consequently, we do not reject the null hypothesis, indicating that the mean age of male and female patients is not significantly different. The mean age for male patients is 53.89873, while for female patients, it is 51.56349. Therefore, according to the Welch Two Sample t-test, there is no substantial difference in age between female and male patients in the dataset.

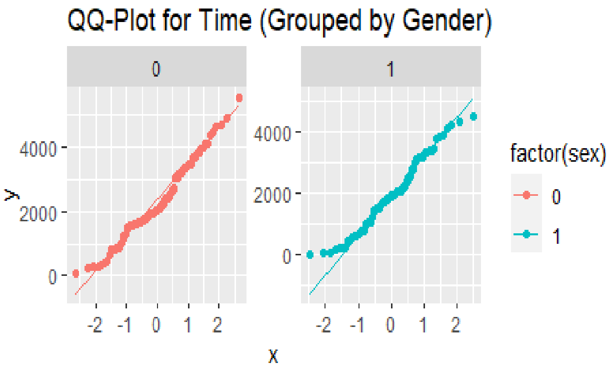
5. Test for Normality Using QQ-Plot

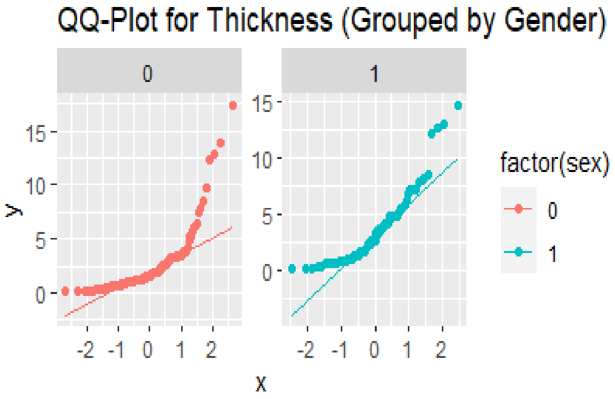
Through the application of Quantile-Quantile (QQ) plots, we examined the normality of the age, time, and thickness variables, specifically focusing on the data from male and female patients who underwent surgery to remove melanoma tumors in Denmark in 1977.

The QQ-plots provide insights into the normality of the distribution. Both the age and time variables showcase adherence to a normal distribution, as evidenced by the data points closely aligning with the diagonal line. However, a slight deviation at the end of the plots indicates the presence of outliers, suggesting potential variations from a perfectly normal distribution.

Contrastingly, the thickness distribution in both genders exhibits a right skewness. The deviation from the diagonal line suggests a departure from a perfectly normal distribution, indicating that the thickness variable is not symmetrically distributed around the mean. This means melanoma tumors removed in the dataset may be either thicker or thinner than the average thickness, rather than being evenly spread around the typical thickness level.







6. Summary:

Melanoma, characterized by its aggressive nature and potential metastasis, becomes challenging to treat once it advances beyond its primary site (Erdei E, 2010). In the latter part of the 20th century, rigorous quantitative methods played a crucial role in defining melanoma prognosis and treatment. President Nixon's declaration of the war against cancer in 1971 and the FDA approval of dacarbazine in 1975 for disseminated melanoma marked a significant milestone in treatment standards (Keiren S, 2012). The observed fluctuations in the number of surgeries over the study years, particularly the gap between 1974 and 1976, indicate a temporal gap in surgeries at the University. This temporal gap could contribute to the decline in surgeries observed afterward.

The pie chart, clustering of values around the survival mean of 1.79, and the negative relationship between survival time and melanoma tumor thickness suggest that most patients survived melanoma. The negative correlation implies that patients with thicker tumors may experience shorter survival times. The leftward skew in the histogram of tumor thickness indicates a concentration of tumors within the 0.1 to 2.17 range. Additionally, the bar chart on ulceration underscores that tumors without ulcers were more prevalent than those with ulcers, supporting observations by Allen and Spitz on size and ulceration (Victor H, 1953).

While an even distribution is observed when grouping values by gender, significant differences between male and female patients emerge in terms of survival time and tumor thickness. Female patients tend to exhibit longer survival times and thinner tumors compared to male patients. Notably, melanoma incidence variations by sex are associated with differences in melanoma anatomical sites (Natalie H, 2017). This underscores the importance of further investigation, including subgroup analyses based on factors such as ethnicity, sunlight exposure, and geography.

In conclusion, the study's findings provide valuable insights into melanoma characteristics, treatment trends, and gender-based differences. Further exploration of temporal patterns, correlations with external factors, and subgroup analyses can contribute to a more comprehensive understanding of melanoma and inform targeted interventions.

References

Natalie H. Mattews 2017 “Epidemiology of Melanoma”. Doi:http://dx.doi.org/10.15586/codon.cutaneousmelanoma.2017.chl

Keiran S.M. Smalley 2012 “A Brief History of Melanoma: From Mummies to Mutations” *DOI:* 10.1097/CMR.0b013e328351fa4d

VICTOR N. TOMPKINS CtJTANEOUS MELANOMA: ULCERATION AS A PROGNOSTIC SIGN

Erdei E, Torres SM. A new understanding in the epidemiology of melanoma. Exp Rev Anticancer Ther. 2010;10(11):1811–23. http://dx.doi.org/10.1586/era.10.170