Exploratory Data Analysis of the Melanoma Dataset in R

Student Number: 2128468

Name: Subas Thapa

Module Code: 7CSO39

Module Name: Statistics for AI and Data Science

1. Introduction

The first Melanoma case was thought to be found in the bones of nearly 2400 years old mummies [1]. In 1757, several cases of melanoma were recorded in the medical literature [1]. In 1787, John Hunter, a Scottish surgeon, removed a patient's tumour [1]. Since then, removal of the tumour by conducting surgical resection became the solution to it [1]. Later melanoma was found to be developed from moles and differs based on hereditary [2][3]. Melanoma is a skin cancer which can be deadly if not treated or removed early [3]. Further found that some people might have a chance of developing melanoma before 10 years of age [4][5]. In 1970, Alexander Breslow classified the tumour's depth into 5 stages based and the unit millimetres [6]. Stage 1(<= 0.75mm), stage 2(0.76-1.5mm), stage 3(1.51 – 2.25 mm), stage 4(2.26-3.0mm) and stage 5(>3.0 mm).

This report consists of an exploratory data analysis of the dataset which consists of detail of 205 Melanoma patients. The dataset consists of 79 male and 126 female patients recorded by the University Hospital of Odense, Denmark. Here are some of the variables measured:

Time: Survival time(days) of a patient after the tumour is removed. This is a discrete variable.
Thickness: Thickness of the tumour in mm and is a continuous variable.
Status: Nominal variable which indicates the patient status.
Sex: Nominal variable, used to indicate the gender of the patient.
Age: Patient's age at the time of the operation. It's a discrete variable.
Year: Records the year patient's tumour was removed.
Ulcer: Nominal variable which indicates whether the ulcer is present in the patient or not.
Category: New custom ordinal variable has been introduced based on Breslow's classification.

2. Summary Statistics

```
> summary(melanoma)
      time                  status             sex           age              year       
 Min.   :  10   Died - Melanoma: 57   Male  : 79   Min.   : 4.00   Min.   :1962  
 1st Qu.:1525   Alive          :134   Female:126   1st Qu.:42.00   1st Qu.:1968  
 Median :2005   Died - Other   : 14                Median :54.00   Median :1970  
 Mean   :2153                                      Mean   :52.46   Mean   :1970  
 3rd Qu.:3042                                      3rd Qu.:65.00   3rd Qu.:1972  
 Max.   :5565                                      Max.   :95.00   Max.   :1977  
   thickness         ulcer        category  
 Min.   : 0.10   Present: 90   Stage 1:34  
 1st Qu.: 0.97   Absent :115   Stage 2:49  
 Median : 1.94                 Stage 3:30  
 Mean   : 2.92                 Stage 4:20  
 3rd Qu.: 3.56                 Stage 5:72  
 Max.   :17.42
```

fig.i Summary of our dataset

There are more female patients (61.5%) diagnosed with melanoma compared to males (38.5%) during the period. Additionally, the ulcer was absent for 56% of the patients. The number of patients who are alive after removing the tumour is high. Half of the patients have tumour thicknesses of size 0.97mm to 3.56mm. The maximum value of the tumour size is 17.42mm, greater than (Q3 + 1.5 * IQR = 3.56 + 1.5*2.59= 7.445mm) which is an outlier. We have a few patients with huge tumour sizes. Range, interquartile range, variance, and standard deviation values for thickness are 17.32, 2.59, 8.758242 and 2.959433 respectively. This indicates that our data has low dispersion with few patients as outliers. Looking at the central tendency of the tumour size, the mean size (2.92mm) of a tumour is greater than the median (1.94mm), indicating that the distribution is right-skewed.

The minimum number of days a patient lived after diagnosis of Melanoma is 10, which indicates that the patient might have died from causes unrelated to Melanoma or a patient might have a huge tumour. 50% of our patients lived around 1525 to 3042 days. Additionally, the range, interquartile range, standard deviation, and variance values obtained are 5555, 1517, 1122.061 and 1259020 respectively. This indicates that our data are spread widely. The mean survival time of patients is greater than the median, this indicates that the distribution is right-

skewed.

Half of the patients diagnosed with melanoma are aged 42 to 65 years. The minimum age of the patient is 4 years, which indicates that the child might have been born with a mole which is a type of melanoma.

We also introduced a new ordinal variable based on tumour thickness using the Breslow classification. We found that there are a huge number of patients with tumour thicknesses greater than 3.0mm.

3. Graphical Summaries

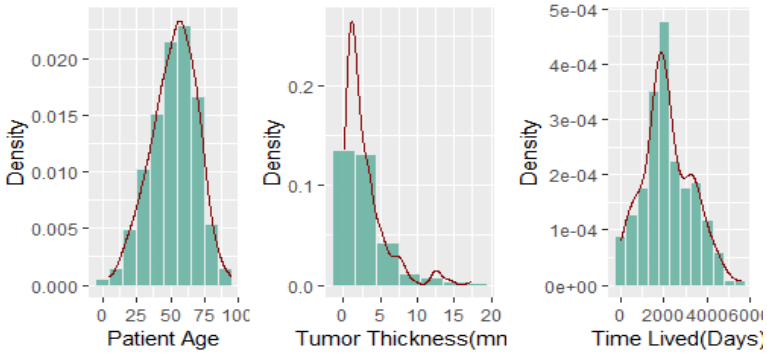

fig. ii Histogram and density plot of variables Age, Thickness and Time

The histogram of the patient's age looks bell-shaped, and the distribution is symmetric. There are a huge number of patients with tumour thicknesses 0-5 mm while few patients have tumour thicknesses greater than 10mm. Looking at the histograms of tumour thickness and survival time, we can say that the distribution of both data is right-skewed.

To examine outliers, distribution of data and skewness among male and female patients, boxplots are created below.

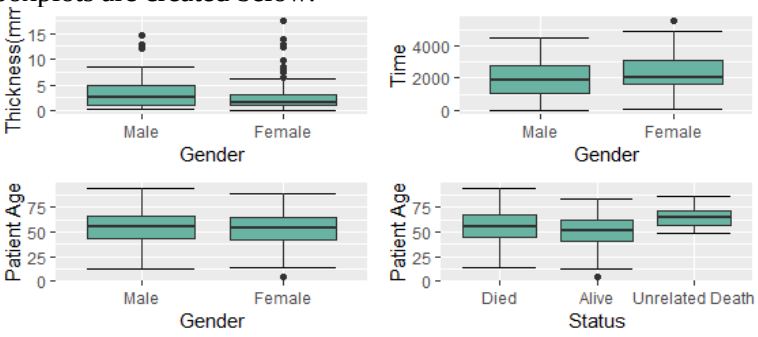

fig. iii Boxplots of thickness, time and age group on sex and status

We can see from the boxplot that the mean thickness of tumours for female patients is slightly lower than that of male patients. Female patients tend to have a smaller tumour size of thickness (range 0-6mm) and have low dispersion compared to male patients(0-8mm). There are few male and female patients with huge tumour thickness which are outliers. The mean survival time of female patients seems to be slightly higher compared to male patients while looking at the boxplots. There is one female patient of small age who survived more days compared to others. The mean age at which the tumour was removed among male and female patients looks identical if we look at the boxplots above. Also, the dispersion of male and female age at which the tumour was removed seems to be identical. The median age of the patients who died from other diseases is about 63 years. This indicates that patients who died not by Melanoma are of older age. We can further conclude that old people with poor health conditions might have died from other diseases.

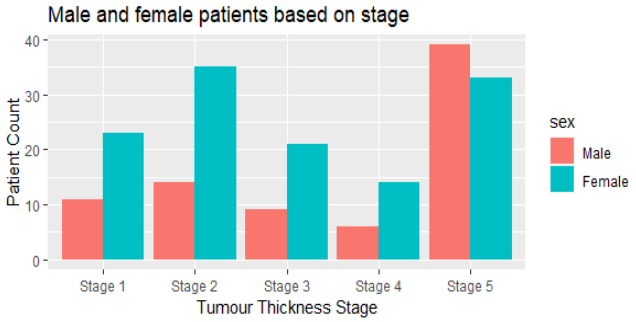

fig. iv

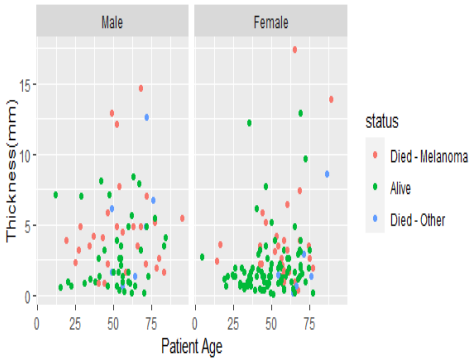

fig. v

Patients with ulcers have bigger tumour sizes compared to patients who don't have them. Survival time seems to be lower for patients with ulcers. Most patients who died of Melanoma are of older age. Female patients of the age group 15-50 years have a low mortality rate compared to male patients. Despite having few male patients compared to female patients, a greater number of male patients have tumour thickness (>3.0mm) and the proportion of male patients who died at this stage is greater than female patients. We can also see that more female patients are alive compared to male patients except for stage 5.

4. Regression Analysis and Correlation Computations

In the scatterplots below, the bold red line represents the regression line.

4.1 Thickness versus time

The variable being predicted here survival time is a response variable while thickness is a predictor variable. The correlation coefficient between time and thickness is -0.2354087.

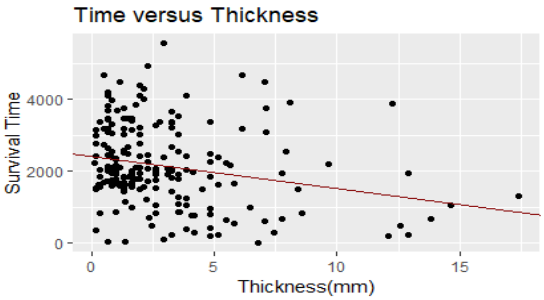

fig. vi

Our regression equation for our model is: y = 2413.41 - 89.25 x
where y = time, x = thickness, gradient = - 89.25, and intercept = 2413.41
The estimated mean survival time can be calculated by replacing the value of tumour thickness(x) in the above equation. For e.g., a patient having a tumour thickness of 5 mm is estimated to live about 1967 days, which can be observed in figure(vi).

4.2 Age versus Time

The variable being predicted here survival time is a response variable while patient age is a predictor variable. The correlation coefficient between time and age is -0.3015179. Both variables are negatively correlated with each other.

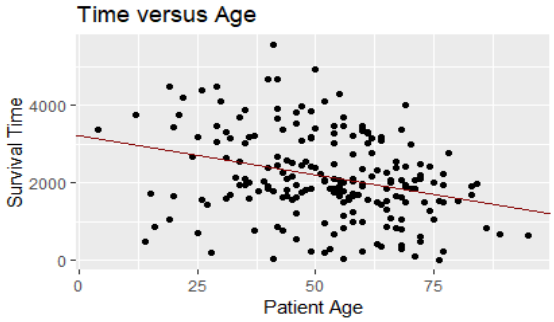

fig. vii

Our regression equation for our model is: y = 3217.448 - 20.293 x
where y = time and x = age, gradient = - 20.293, intercept = 3217.448
The estimated mean survival time can be calculated by replacing the value of age in the above equation. For e.g., a patient of age 60 is estimated to live about 2000 days, which can be observed in figure(vii).

4.3 Age versus thickness

The variable being predicted here is tumour thickness is a response variable while patient age is a predictor variable. The correlation coefficient between the thickness of the tumour and patient age is found to be 0.2124798 and is much less than 1. This indicates that there is a weak correlation between these two variables.

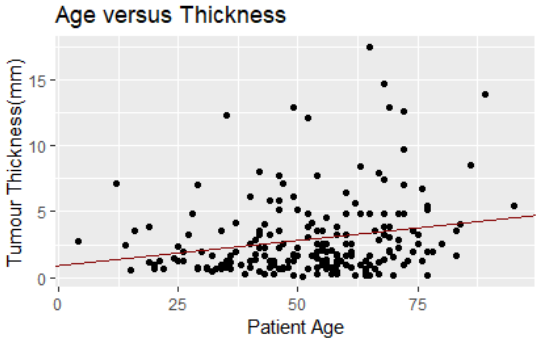

fig. viii

Our regression equation for our model is y = 0.94105 + 0.03772 x
where y = thickness, x = age, gradient = 0.03772 and intercept = 0.94105
The estimated mean thickness can be calculated by replacing the value of age in the above equation. For e.g., a patient of age 60 is estimated to have a tumour of thickness 3.20 mm, which can be observed in figure(viii).

5. Observed Relationship

5.1 Thickness versus time

The correlation coefficient between time and thickness is -0.2354087. The correlation coefficient between the two variables survival time and thickness is found to be negative. Both variables are negatively correlated with each other. Patients seem to live longer when they have small tumour thickness and vice versa. A patient with a tumour thickness of 2 mm seems to live about 2235 days while a patient with a tumour thickness of 5 seems to live 1967 days. This shows that the survival time is less for patients with bigger tumour sizes and vice versa.

5.2 Age versus time

The correlation coefficient between the two variables age and time is -0.3015179. This indicates that there is a negative relationship between these two variables. Younger patients seem to live longer compared to older people. A negative gradient is obtained for the two variables. Young patients of age 20 years seem to live about 2811 days while an older patient of age 65 years seem to live 1898 days. Based on this we can draw a conclusion that younger patients seem to

live longer days compared to older.

5.3 Age versus thickness

For the two variables thickness and age, the correlation coefficient is found to be 0.2124. This indicates that the relationship between these two variables is not that strong. An upward slope is created when a relationship is positive. In our dataset, we found that younger patients have small tumour sizes compared to older patients. The thickness of a tumour also depends on whether the patient has an ulcer or not. Patients of age 30 years are predicted to have an average thickness of 2.07mm while for 60-year-old patients is 3.20mm.

6. Two Sample Significance Test

Two sample significance test is performed to check whether the survival time, thickness and age among male and female patients are the same. H0 is our null hypothesis and H1 is our alternative hypothesis. We use a T-test to approve or reject our null hypothesis. We reject or accept the null hypothesis based on the p-value we received against the level of significance. Our default level of significance is 0.05, which is 5%.

6.1 Tumour thickness of male and female patients

Let's check whether the tumour thickness in male and female patients are same or not.
H0: Mean tumour thickness among male and female patients are same $(\mu_1 = \mu_2)$
H1: Mean tumour thickness among male and female patients are different $(\mu_1 \neq \mu_2)$
The default level of significance $\alpha = 0.05$ and p-the value we received from the t-test is 0.01009. The p-value we received is less than the level of significance, so we reject the null hypothesis (H0) and conclude that our alternative hypothesis(H1) is correct.

6.2 Survival time of male and female patients

After looking at the boxplots above, we observed that the mean survival time among male and female patients is different. To test this, we set the null hypothesis(H0) and alternative hypothesis(H1) as below:
H0: Mean survival time among male and female patients are same $(\mu_1 = \mu_2)$
H1: Mean survival time among male and female patients are different $(\mu_1 \neq \mu_2)$
The default level of significance $\alpha = 0.05$ and p-the value we got from the t-test is 0.0386. The p-value we received from the test is less than the level of significance, so we reject the null hypothesis(H0) and conclude that our alternative hypothesis(H1) is correct.

6.3 Age of male and female patients diagnosed with melanoma

After observing the boxplots in fig. iii, the mean age of patients for both genders look similar. To verify that two sample significance test is conducted below:
H0: Mean age among male and female patients at which tumour was removed are same $(\mu_1 = \mu_2)$
H1: Mean age among male and female patients at which tumour was removed are different $(\mu_1 \neq \mu_2)$
The default level of significance $\alpha = 0.05$ and p-the value we got from the t-test is 0.3408. The p-value we received is significantly higher than the level of significance, so we accept the null hypothesis(H0).

7. QQ Plots

Below plots show the distribution of variables age, time and thickness among male and female patients in an order. We plot those plots to see if sample data are drawn from normally distributed populations or not.

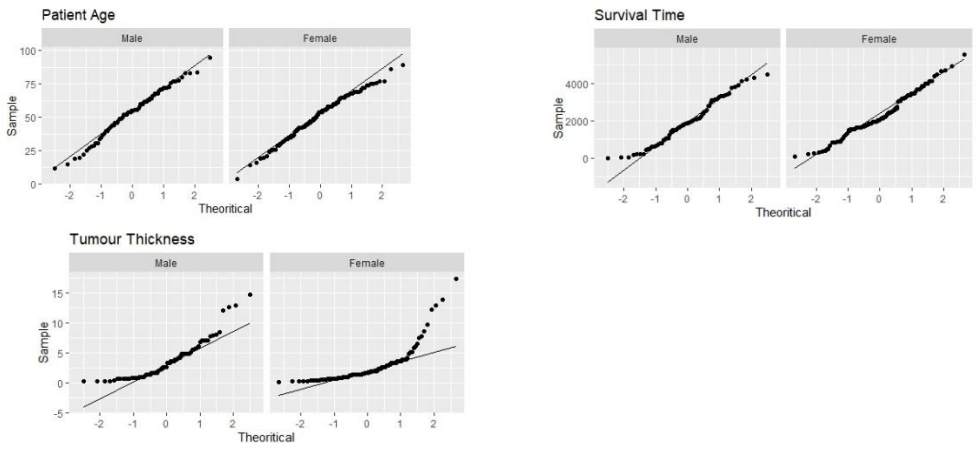

fig. ix

As we can see from the above qq-plots of the patient's age, most data lie close to the straight line. This indicates that the data are normally distributed. After that most of the survival time data are close to the straight line and can be stated that the data are distributed normally. While most of the tumour thickness data are away from the straight line and we can conclude that those data might not have been drawn from a normally distributed population. If we look at the histogram from the above figure (fig. ii) we could also see that the patient's age and survival time look bell-shaped while tumour thickness is right-skewed.

8. Data Insights and Recommendations

The distribution of the mean survival time of patients (male and female) is right-skewed. Most patients have tumour thicknesses of 0-5mm while few patients have tumour thicknesses of size greater than 7mm. Female patients tend to have small tumour sizes and the distribution is less dispersed compared to male patients. After carefully examining using a t-test we found out that the mean thickness among male and female patients is different. Out of 126 female patients, the ulcer was not present in 79 (62.6%) patients. The mean survival time and quartile values for alive patients are almost similar among both sexes despite having significant differences in mean thickness. The distribution of thickness data is not normally distributed and has different mean tumour thicknesses among male and female patients.

By looking at the boxplots, histograms and qqplots of patient age we can say that the data are normally distributed. The mortality rate is low for younger patients compared to older patients. We further found that mortality in female patients of the age range 15-50 years is low compared to male patients of a similar age range. The mortality rate and mean tumour thickness of patients having ulcers are high for both genders. We further examined that the mean age at which the tumour was removed is the same among male and female patients using two sample significance tests. Presence of ulcer also seems to decrease the mean survival time. Additionally, female patients tend to live longer compared to male patients.

We grouped and classified patients based on tumour thickness. We found that there are a huge number of patients with tumour thicknesses greater than 3.0mm and the mortality rate is high for that group.

After conducting two sample significance tests, we found that mean tumour thickness and survival time among male and female patients vary. If we want to investigate further how the dependence of tumour thickness and survival time related to gender, we need to perform the Chi-Squared independence test. To investigate how and why mortality rate in younger patients we need more data. More data need to be collected for various tumour thicknesses to make a distribution more normal.

References:

1. Vito W. Rebecca Et. Al., A Brief History of Melanoma: From Mummies to Mutations.
2. Carrie Lee Et. Al, Historical review of melanoma treatment and outcomes
3. Lauren E. Davis, Alan J. Tackett and Sara C. Shalin, Current state of melanoma diagnosis and treatment, 2019
4. A Hunter Shain, Boris C Bastian, From melanocytes to melanomas, 2016
5. Chunying Li Et. Al., Polymorphisms in the DNA repair genes XPC, XPD, and XPG and risk of cutaneous melanoma: a case-control analysis, 2006
6. Richard A. Scolyer Et. Al., Evolving concepts in melanoma classifications and their relevance to multidisciplinary melanoma patient care