7CS039-INYR-UM1: Summative Assessment 1

Statistics for Data Science, Faculty of Science and Engineering, University of Wolverhampton

Faraz Yusuf Khan 2307375 7CS039

Contents

1.0 Introduction

The dataset contains information regarding the progression of gastric ulcers for 205 patients with a time frame from 1962 to 1977. Information provided is based on the status of the disease, year of observation, tumour thickness measured in mm, and presence or absence of an ulcer. Patients underwent surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark between 1962 and 1977. Key variables in the provided dataset can be identified as survival time, tumour thickness, and ulceration. Tumour thickness and ulceration are considered important prognostic factors with thicker or ulcerated tumours associated with a higher mortality risk.

2.0 Summary Statistics

Summary statistics have been calculated with the help of R-Studio. Initially, an examination of summary statistics revealed that there are categorical variables namely, status, sex, and ulcer. For better visual and statistical interpretation, they have been encoded in accordance with the description and presented in Code Snippet 1 (Pargent et al., 2022).

```
# Get a summary of the 'melanoma' dataset
summary(melanoma) # This will provide statistical summary for each column in the dataset
# Convert 'status' variable to factor with labels 'death_melanoma', 'alive', 'death_other'
melanoma$status <- factor(melanoma$status, labels = c("death_melanoma", "alive", "death_other"))
# Convert 'sex' variable to factor with labels 'Female' and 'Male'
melanoma$sex <- factor(melanoma$sex, labels = c("Female", "Male"))
# Convert 'ulcer' variable to factor with labels 'Absent' and 'Present'
melanoma$ulcer <- factor(melanoma$ulcer, labels = c("Absent", "Present"))
# View the 'melanoma' dataset again to confirm if changes have been made
View(melanoma)
# Get a summary of the 'melanoma' dataset again to see the changes
summary(melanoma)
```

```
      X              time                status             sex           age              year     
 Min.   :  1   Min.   :  10   death_melanoma: 57   Female:126   Min.   : 4.00   Min.   :1962  
 1st Qu.: 52   1st Qu.:1525   alive         :134   Male  : 79   1st Qu.:42.00   1st Qu.:1968  
 Median :103   Median :2005   death_other   : 14                Median :54.00   Median :1970  
 Mean   :103   Mean   :2153                                     Mean   :52.46   Mean   :1970  
 3rd Qu.:154   3rd Qu.:3042                                     3rd Qu.:65.00   3rd Qu.:1972  
 Max.   :205   Max.   :5565                                     Max.   :95.00   Max.   :1977  
   thickness         ulcer     
 Min.   : 0.10   Absent :115  
 1st Qu.: 0.97   Present: 90  
 Median : 1.94                
 Mean   : 2.92                
 3rd Qu.: 3.56                
 Max.   :17.42
```

Code Snippet 1: Summary Statistics

Time Variable

Wide variability has been discovered for the time variables which indicates survival time since the operation. Half of the patients passed away within roughly 5.5 years (2005 days), while 25% died within 4 years (1525 days) and 75% within 8 years (3042 days). The average survival time was found to be equal to almost 6 years (2153 days)

Age Variable

The ages of patients range greatly from 4 years to 95 years of age however, a majority of them can be categorized into the middle age group. The mean age of 52.46 years with a median age of 54 years. The interquartile range lies from 42-65 years which confirms that a majority of patients have undergone tumour removal surgery.

Tumour Thickness

The Tumour variable is centred around thinner tumours with a minimum value of 0.1 mm and a maximum value of 17.42 mm with an average value of 2.92 mm. The middle 50% range in mm from 0.97 to 3.56. The extreme outliers are the minimum of 0.1 mm and the greatest of

17.42 mm. Although the distribution is right-skewed due to prominent outliers, the data tends to favour smaller sizes.

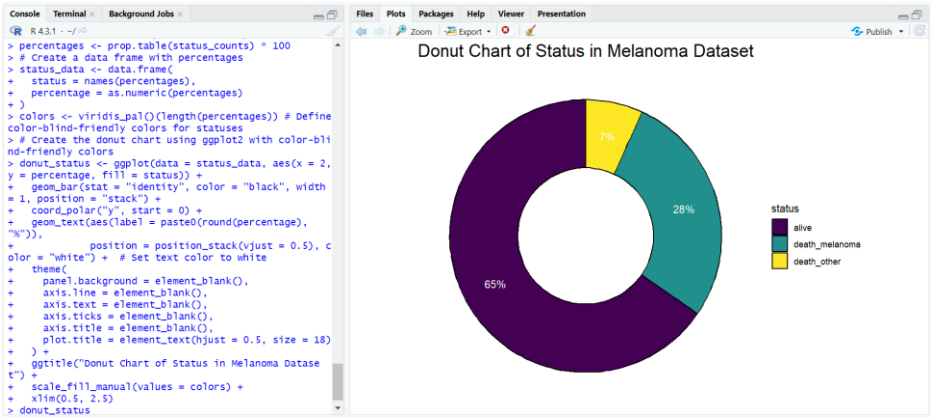
Status

In the melanoma dataset, 27.8% (57 individuals) died from melanoma, 6.8% (14 persons) died from non-melanoma causes, and 65.4% (134 patients) were alive at the trial's end in 1977. Most patients are alive, with a smaller proportion succumbing to non-melanoma causes and a larger percentage surviving melanoma. In the dataset, the gender of the patients is represented in the sex column. It shows that there are 126 female patients and 79 male patients.

Ulcer

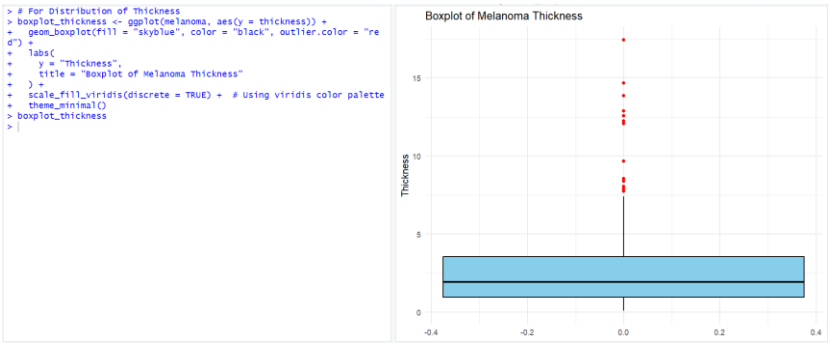
The ulcer column in the dataset signifies whether the melanoma was ulcerated. Ulceration is a factor that can influence the prognosis of melanoma. According to the data, ulceration was absent in 115 cases and present in 90 cases.

3.0 Graphical Summary



Code Snippet 2: Donut Chart displaying status of melanoma in the dataset

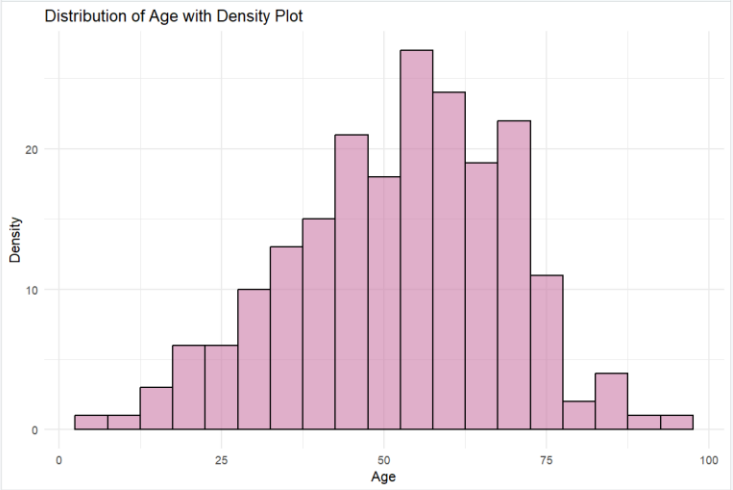
For the status variable, a donut chart is shown in Code Snippet 2. Since a donut chart successfully depicts the proportions of categorical data, it has been used to illustrate the status variable. In this instance, Figure 1 shows that most patients were still alive at the end of the research, with 6.8% dying from other causes and 27.8% dying from melanoma.



Code Snippet 3: Box Plot for distribution of thickness

Using Code Snippet 3, a Box plot was produced to indicate thickness. The median thickness is found to be right-skewed, closer to the box's bottom bound. Additionally, a sizable number of outliers have been discovered outside the box plot's upper bound.

```
> #For Distribution of Age
> # Set colorblind-friendly palette
> cb_palette <- c("#0072B2", "#E69F00", "#009E73", "#F0E44
2", "#D55E00", "#CC79A7")
>
> # Histogram with density plot for age
> ggplot(melanoma, aes(x = age)) +
+   geom_histogram(binwidth = 5, fill = cb_palette[6], color
= "black", alpha = 0.6) +
+   geom_density(fill = cb_palette[6], alpha = 0.6) +
+   labs(title = "Distribution of Age with Density Plot",
+        x = "Age",
+        y = "Density") +
+   theme_minimal()
>
```



Code Snippet 4: Distribution of Age with Density plot

With the use of Code Snippet 4, an age density plot has been produced. The majority of the people in the sample are middle-aged. In melanoma cases, two outliers—a 4-year-old and a 95-year-old—stand out and should be closely examined for accuracy and possible scientific importance.

4.0 Regression and Correlation

4.1 Regression (time ~ thickness)

Parameter	Estimate	Comments
Intercept	2413.41	Predicted time when thickness is 0
thickness	-89.25	Time decreases 89.25 units per 1 unit increase in thickness
R-squared	0.05542	Thickness explains only 5.542% of variance in time
Residual Standard Error	1093	High standard deviation of residuals indicating poor model fit
F-statistic	11.91	Significant linear relationship between thickness and time
p-value	0.0006793	Thickness coefficient is statistically significant

Table 1: Regression for Time and Thickness

The findings of a linear regression model analysing tumour thickness and melanoma survival time are shown in Table 1. The R-squared value of 0.055 indicates that thickness accounts for just 5.5% of the variation in survival time. With a negative coefficient of -89.25, the model predicts that for every 1 mm increase in thickness, there will be an 89.25-day decrease in projected survival. This effect, however, is far less than the significant residual standard error of 1093 days, suggesting a significant amount of unexplained variability.

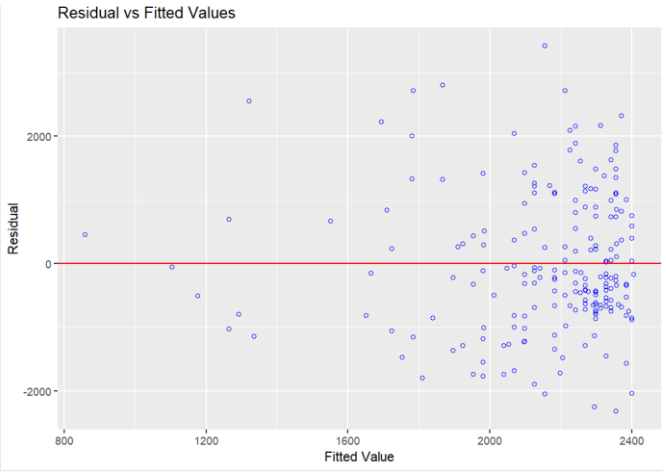
Model constraints cast doubt on the coefficient's practical value despite statistical significance. Given the limitations of the model, the p-value indicates a relationship that is unlikely to occur

by chance, but it provides little information about clinical significance. In summary, the model's ability to forecast melanoma survival time is limited because it only takes into account tumour thickness. The low R-squared, considerable residual error, and moderate effect size all highlight how crucial it is to take into account extra variables in order to fully comprehend the prognosis of melanoma.

4.1.1 Assumptions: Regression (time ~ thickness)

The Breusch Pagan test was conducted and it was found to reject homoskedasticity, suggesting heteroskedasticity, and the residual plot displays a fanning pattern as suggested by Code Snippet 5. This goes against the constant variance assumption of linear regression, which could skew standard errors and conclusions.

```
# Run regression for time ~ thickness
reg_model1 <- lm(time ~ thickness, data = melanoma)
# Summary of the regression results
summary(reg_model1)
#Test for Homoskedasticity
ols_plot_resid_fit(reg_model1) #a funnel shape is found
ols_test_breusch_pagan(reg_model1)
#Test for Normality
ols_plot_resid_hist(reg_model1)
shapiro.test(reg_model1$residuals)
#Test for Linearity
# Scatter plot for 'Thickness' vs. 'Time'
plot(melanoma$thickness, melanoma$time,
     xlab = "Thickness (mm)", ylab = "Time (days)",
     main = "Scatter Plot: Thickness vs. Time",
     pch = 20,          # Use solid circle point shape
     cex = 1.5,         # Set point size
     col = "black")     # Set point color to black
fit <- lm(melanoma$time ~ melanoma$thickness)
abline(fit, col = "red", lwd = 2)
# Trendline in red with increased line width
legend("topleft", legend = "Trendline", col = "red",
       lty = 1, lwd = 2)
grid(col = "gray")  # Add gridlines
```

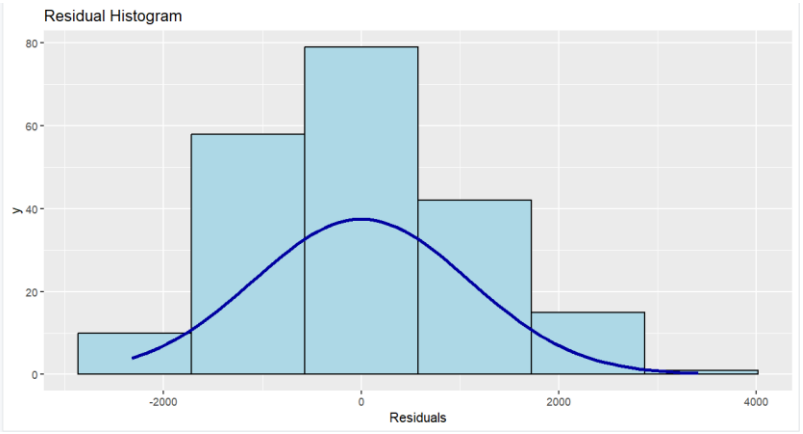


Code Snippet 5: Residual vs Fitted Values

```
> #Test for Normality
> ols_plot_resid_hist(reg_model1)
> shapiro.test(reg_model1$residuals)

        Shapiro-Wilk normality test

data:  reg_model1$residuals
W = 0.9777, p-value = 0.002413
```



Code Snippet 6: Test for Normality (time and thickness)

Regression model residuals subjected to the Shapiro-Wilk test yielded a W-statistic of 0.9777 and a p-value of 0.002413, rejecting the null hypothesis that the residuals were normally distributed. The test found strong evidence of non-normality, which went against the presumptions of the model. To increase model adequacy, further research is necessary. This research may involve changing variables, utilizing different models, or utilizing techniques resistant to non-normality. Code Snippet 6 has been used to construct the Shapiro-Wilk test and histogram.

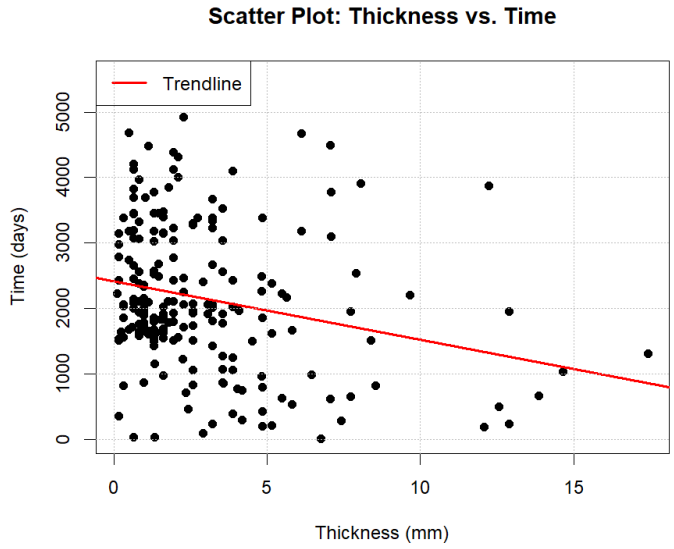


Figure 3: Scatterplot for Linearity (Time and Thickness)

The scatterplot in Figure 3 indicates that there are no significant linear correlations between the variables. Therefore, linear regression's requirement of linearity between variables is not satisfied.

4.2 Regression (time ~ age)

Parameter	Estimate	Comments
Intercept	3217.448	Predicted time when age is 0
Age	-20.293	Time decreases 20.293 units per 1 unit increase in age
R-squared	0.09091	Age explains only 9.091% of variance in time
Residual SE	1072	Typical deviation between predicted and actual time
F-statistic	20.3	Overall model is statistically significant
p-value	1.12E-05	Age coefficient is statistically significant

Table 2: Regression for Time and Age

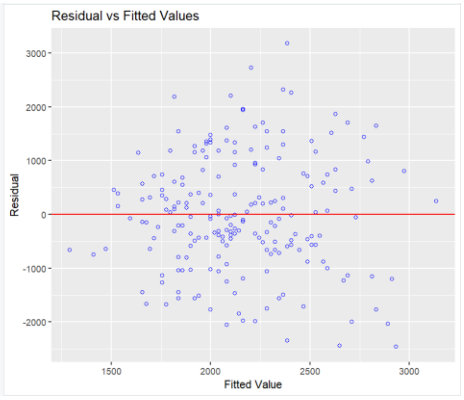
The association between patient age and survival time for the melanoma dataset is shown by the linear regression model in Table 2, which produces an incredibly low R-squared value of 0.091. This suggests that age has very little prognostic power as a single predictor, explaining only 9.1% of the variability seen in survival time. For every year that an individual's age increases, the model yields a statistically significant coefficient that estimates a 20.3-day drop in survival. However, this effect is clinically inconsequential due to its tiny magnitude in comparison to the enormous residual error of 1072 days.

Ultimately, any conclusions on the significance of age derived from this analysis are seriously undermined by the poor model fit, limited predictive power, and violation of assumptions such as heteroskedasticity. Attempting to characterize age as a driver of survival time based on these results alone is unwarranted. In summary, this regression analysis is inadequate to support definitive conclusions about age as a lone predictor and highlights the need for expanded clinical investigation using appropriate statistical methods.

4.2.1 Assumptions: Regression (time ~ age)

```r
# Run regression for time ~ age
reg_model2 <- lm(time ~ age, data = melanoma)
# Summary of the regression results
summary(reg_model2)
#Test for Homoskedasticity
ols_plot_resid_fit(reg_model2)
ols_test_breusch_pagan(reg_model2)
#Test for Normality
ols_plot_resid_hist(reg_model2)
shapiro.test(reg_model1$residuals)
# Scatter plot for 'Time' vs. 'Age' with a trendline
plot(melanoma$age, melanoma$time,
     xlab = "Age", ylab = "Time (days)",
     main = "Scatter Plot: Time vs. Age",
     pch = 20,          # Use solid circle point shape
     cex = 1.5,         # Set point size
     col = "black")     # Set point color to black

fit <- lm(melanoma$time ~ melanoma$age)
abline(fit, col = "red", lwd = 2)  # Trendline in red
#with increased line width
legend("topright", legend = "Trendline", col = "red", lty = 1, lwd = 2)
grid(col = "gray")  # Add gridlines
```



Code Snippet 7: Test for Heteroskedasticity (Time and Age)

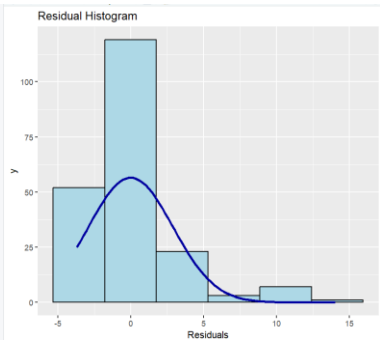
Code Snippet 7 reveals a fanning pattern for the Breusch Pagan test thereby rejecting homoskedasticity and indicating heteroskedasticity. Therefore, the constant variance assumption of linear regression is not met.

```
> ols_plot_resid_hist(reg_model3)
> shapiro.test(reg_model3$residuals)

        Shapiro-Wilk normality test

data:  reg_model3$residuals
W = 0.79949, p-value = 1.659e-15

>
```



Code Snippet 8: Shapiro Wilk test and Histogram for Normality (Time and Age)

The graphical and statistical evidence shown in Code Snippet 8 suggests that residuals are adequately normally distributed. Therefore, meeting the assumption for linear regression residuals should be normally distributed.

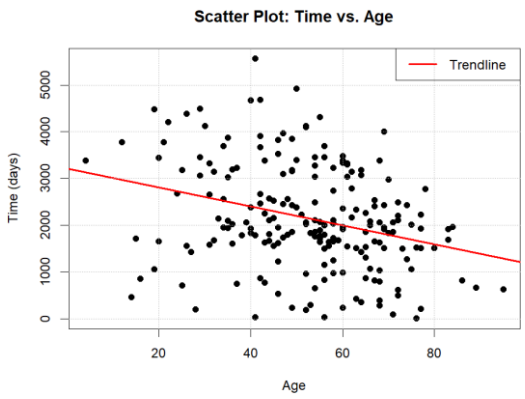


Figure 2: Scatterplot for Linearity (Time vs Age)

There are no discernible linear correlations between the various variables, as the scatterplot in Figure 2 suggests. As a result, linear egression's linearity assumption between variables is not satisfied.

4.3 Regression (thickness ~ age)

Parameter	Estimate	Comments
Intercept	0.94105	Predicted time when age is 0
age	0.03772	Positive value indicates thickness increases with age.
R-squared	0.04515	The model explains 4.5% of variation in thickness.
F-statistic	9.598	Overall model is statistically significant
Residual SE	2.899	Error in predicting thickness. Quite high relative to coefficient estimates.
p-value	0.00222	Age coefficient is statistically significant

Table 3: Regression for Thickness and Age

Table 3 summarizes the outcomes of a regression model investigating the association between thickness and age. The intercept suggests a predicted thickness of 0.94105 when age is 0. The positive age coefficient of 0.03772 implies that thickness tends to increase with age, with an estimated increase of approximately 0.038 mm for each 1-year rise in age.

Although the F-statistic of 9.598 deems the overall model statistically significant, the R-squared value of 0.04515 indicates that only 4.5% of the variation in thickness is explained by age. The residual standard error of 2.899 is relatively high compared to the coefficient estimates, suggesting a notable amount of error in predicting thickness.

The p-value of 0.00222 for the age coefficient confirms its statistical significance. In conclusion, while a statistically significant positive relationship between thickness and age exists, the limited explanatory power (as indicated by the low R-squared) and the relatively high residual standard error raise caution. The model captures a weak linear association, emphasizing the need for consideration of other factors influencing thickness.

4.3.1 Testing for Assumptions

The Breusch-Pagan test revealed strong evidence of heteroskedasticity (p=0.0002) in the linear model predicting melanoma thickness, suggesting that the residual variance violates the constant variance assumption (Appendix 1). Additionally, as shown in Appendix 1, heteroskedasticity was discovered by graphical examination of the residuals and fitted values.

The Shapiro-Wilk test for normality found strong evidence (p=1.659e-15) that the residuals of the melanoma thickness regression model deviate from a normal distribution, violating model assumptions. Furthermore, Histogram of residuals confirms the presence of right skewness as shown in Appendix 2. Appendix 3 displays a scatterplot showing strong evidence of non-linearity between age and melanoma tumour thickness, violating the linearity assumption of the regression model. This suggests a linear model is inappropriate and that non-linear transformations of age or non-parametric modelling should be explored to better fit the relationship.

4.4 Correlation

Variable 1	Variable 2	Correlation	p-value	Significance	Orientation	Nature
time	thickness	-0.2354087	0.00067928	Statistically significant	Weak	Negative
time	age	-0.3015179	1.12E-05	Statistically significant	Weak	Negative
thickness	age	0.2124798	0.00222338	Statistically significant	Weak	Positive

Table 4: Correlation

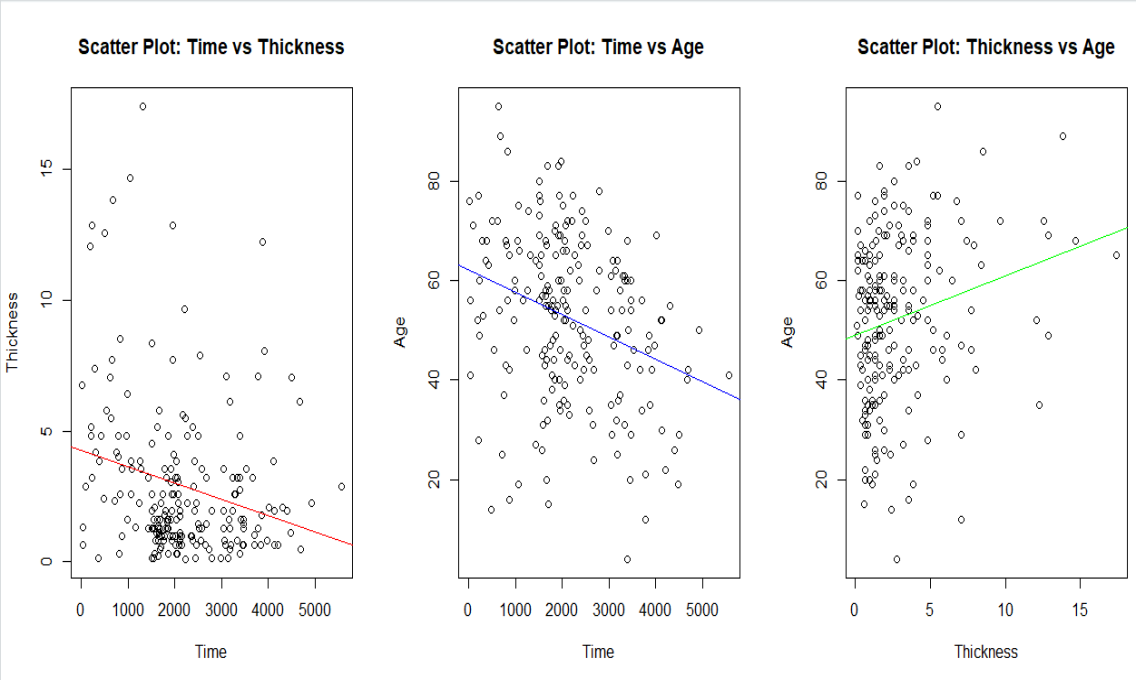


Figure 4: Correlation Scatterplots

The correlations between time, thickness, and age are clarified by Table 4 and visualized in Figure 4, which presents the results of correlation tests performed on the melanoma dataset. All three variable pairings have statistically significant associations, according to the analysis (Bekheit et al., 2020).

Time and Thickness: A weak negative association (r=-0.235, p=0.000679) indicates that thickness has a slight tendency to decline with time.

Time and Age: Likewise, a weak negative correlation (r=-0.301, p=1.12E-05) suggests that age shows a propensity to decline with time, albeit at a very low rate.

Age and Thickness: There is a weak but statistically significant positive association (r=0.212, p=0.002223) showing that tumour thickness generally tends to grow with age.

In conclusion, the melanoma dataset's correlation analysis reveals important correlations between thickness, age, and time. Age and thickness show a slight positive correlation, whereas time and thickness show negative associations. Code to generate correlation and plots for Figure 4 have been generated through Appendix 4.

4.5 Two Sample Significance Tests

4.5.1 Thickness

A. Hypotheses:

H_0 (null hypothesis): There is no difference in the mean thickness between males and females.

H_1 (alternative hypothesis): There is a difference in the mean thickness between males and females.

B. Significance level:

5 % Significance level with Alpha = 0.05.

C. Test statistic:

t-statistic from two sample t-test using the t.test() function in R.

D. Critical values:

Determine critical t-value based on alpha and degrees of freedom.

E. Decision rule:

If t > critical t, reject H_0.

If t < critical t, fail to reject H_0.

F. Conclusion:

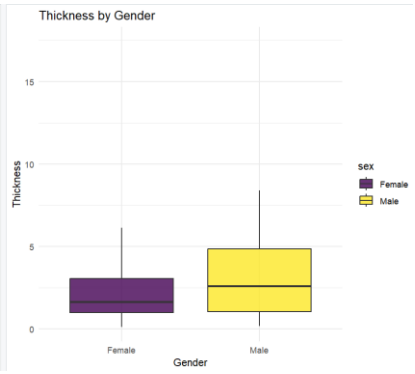
For thickness: t = 2.6059, p = 0.01009 < 0.05. Reject H_0.

```
> ##THICKNESS
> # Subset data by gender
> thickness_female <- melanoma$thickness[melanoma$sex == "Female"]
> thickness_male <- melanoma$thickness[melanoma$sex == "Male"]
> # Perform t-test
> t_test_thickness_gender <- t.test(thickness_female, thickness_male)
> # Print test results
> print(t_test_thickness_gender)

        Welch Two Sample t-test

data:  thickness_female and thickness_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

> # Create a boxplot with color-friendly palette
> ggplot(melanoma, aes(x = sex, y = thickness, fill = sex)) +
+   geom_boxplot(alpha = 0.8, outlier.shape = NA) +
+   scale_fill_viridis(discrete = TRUE) +
+   labs(title = "Thickness by Gender", x = "Gender", y = "Thickness") +
+   theme_minimal()
>
```



Code Snippet 14: T-Test for Thickness by Gender with Boxplot

There is sufficient statistical evidence at the 5% significance level to conclude there is a difference in the mean-time and thickness between males and females in the sample. Further confirmed by box-plot generated in Code Snippet 14.

4.5.2 Age

A. Hypotheses:

H_0: There is no difference in the mean or age between males and females.

H_1: There is a difference in the mean age between males and females.

B. Significance level:

Alpha = 0.05

C. Test statistic:

t-statistic from two sample t-test

D. Critical values:

Determine critical t-value based on alpha and degrees of freedom.

E. Decision rule:

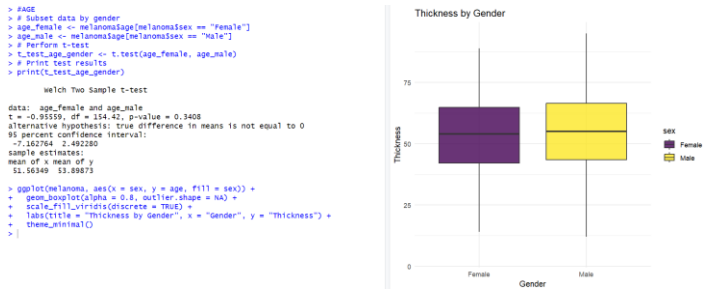
If t > critical t, reject H_0.

If t < critical t, fail to reject H_0.

F. Conclusions:

t = -0.95559, p = 0.3408 > 0.05

Fail to reject H0. There is insufficient evidence to conclude there is a difference in mean age.



Code Snippet 15: T-Test for Age by Gender with Boxplot

It can be said that with 95% Statistical confidence there exists no difference between the mean age for males and females as suggested by the T-test at 5% significance level and illustrated in box plot generated in Code Snippet 15.

4.5.3 Time

A. Hypotheses:

H_0: There is no difference in the mean time between males and females.

H_1: There is a difference in the mean time between males and females.

B. Significance level:

Alpha = 0.05

C. Test statistic:

t = 2.0848

D. Critical value:

With df = 159.27, the critical t = 1.974

E. Decision rule:

If t > critical t, reject H_0.

If t < critical t, fail to reject H_0.

F. Conclusion:

Since t = 2.0848 > critical t = 1.974, reject H_0.

Code Snippet 16: T-Test for Time by Gender with Boxplot

The hypothesis test for time between males and females yielded a p-value of 0.03868, which is less than the 5% significance level (Oti el al., 2021). Therefore, there is sufficient evidence to reject the null hypothesis and conclude there is a statistically significant difference in the mean time. Specifically, the sample data suggests females have a higher average time compared to males as evidenced by boxplot generated in Code Snippet 16.

4.6 Q-Q Plots

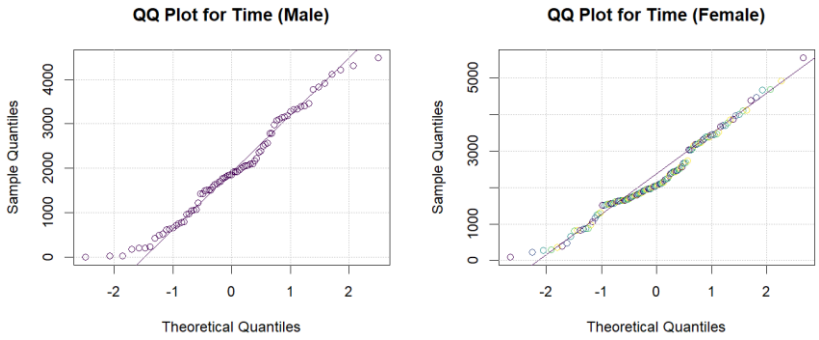


Figure 5: QQ Plots for Time

The QQ plot for the male time data indicates the distribution aligns relatively closely with the theoretical normal distribution, as seen by the data points following the diagonal line with only minor deviations as displayed in Figure 5. Specifically, the points exhibit no major systematic curvature and generally adhere to the linear trend, with slight departures observed in the moderate tails. Given the overall conformity to the normal shape, notwithstanding some expected random fluctuation, the male time data QQ plot suggests this group satisfies the assumption of normality reasonably well based on visual inspection.

In contrast, the QQ plot for the female time data reveals more substantial departures from the normal shape, evidenced by the observable curvature in the centre and heavier tails (Dansana et al., 2020). The data points diverge from the theoretical diagonal, especially around the median and in the upper tail, demonstrating skewness and kurtosis inconsistent with a normal distribution. The obvious systematic deviation from the linear trend provides a graphical indication that the female time data violates the normality assumption, complementing formal statistical testing. In summary, examination of the QQ plots reveals the male time distribution aligns more closely with normality than the female time distribution.

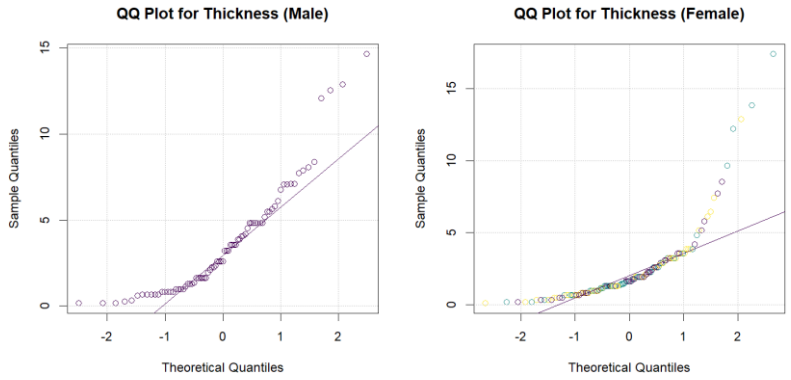


Figure 6: QQ Plots for Thickness

The QQ plot for male thickness data clearly departs from the normal distribution line, exhibiting substantial systematic curvature and heavy tails. The data points diverge noticeably from the theoretical diagonal, with an S-shaped pattern demonstrating both skewness and

excessive kurtosis inconsistent with normality. The upper tail shows particularly severe divergence, with extreme values deviating severely from the expected trend. This substantial graphical departure from the normal shape provides strong visual evidence that the assumption of normality is violated for male thickness data.

Similarly, the QQ plot for female thickness data also displays marked deviation from the normal shape, with visible S-shaped curvature indicating skewness in the distribution. The plotted points exhibit heavy tails, especially in the upper end, reflecting kurtosis beyond that expected in a normal distribution. While less extreme than the males, the obvious systematic graphical deviation from the normal theoretical line provides a clear indication that the female thickness data also violates the assumption of normality. In summary, examination of both gender QQ plots conveys the thickness variable is non-normal in distribution for either group.

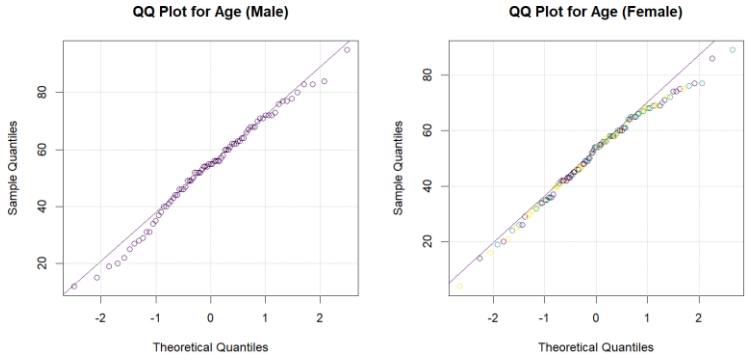


Figure 7: QQ Plots for Age

The QQ plots for both male and female age data align relatively closely to the theoretical normal distribution line, with only minor deviations observable. The male plot shows slight curvature in the moderate tails while the female plot exhibits minimal divergence across the distribution. Overall, neither gender plot displays substantial or systematic departures from the normal shape. The general conformity to a linear trend suggests the age variable is sufficiently normal in distribution for both groups based on visual inspection of the QQ plots as illustrated in Figure 7.

4.7 Insights

It is unexpected that there is such a poor link between tumour thickness and survival duration given that thickness has been shown in other research to be a significant predictive factor (Tapoi et al., 2023). This implies that additional factors, such as age, gender, and ulceration status, might potentially be significant in determining survival.

Model misspecification is indicated by assumptions such as heteroskedasticity, non-normality, and non-linearity being broken. This could be addressed by taking non-linear covariate effects into account in generalized additive models with spline terms (Mundo et al., 2022).

Model-based recursive partitioning is one segmentation technique that can be used to find heterogeneity in subgroups that are defined by interactions between factors such as thickness, age, and gender (Jones et al., 2020). Kaplan-Meier curves are the most effective way to depict the unique survival characteristics that these subgroups may have.

Due to them taking right-censored observations into account, survival analysis techniques like Cox proportional hazards regression are preferred over linear regression (Muse et al., 2022). Prior to using a Cox model, it is crucial to verify the proportionate hazards assumption and assess interactions and time-varying effects.

Traditional Cox modelling is enhanced by machine learning techniques such as random survival forests, which capture intricate non-linear correlations and higher-order interactions (Kantidakis et al., 2020). Generalizability would be assessed through external validation on a separate dataset.

The weak relationships revealed between thickness, age, and survival time suggest the presence of unmeasured confounding. It takes a multivariate approach to identify the distinct contributions of many prognostic variables by analysing their intricate interactions.

4.8 Recommendations

A. Data Collection

- Expand data collection to include additional prognostic factors beyond just tumor thickness, such as presence of ulceration, mitotic rate, patient age, gender, metastatic disease stage, and anatomical site. This allows for more complex multivariate analyses to uncover interactions and independent effects of different factors (Joshi et al., 2021).
- Increase sample size to improve statistical power for subgroup analyses, larger samples will reduce variability in estimated effects.
- Perform log or other transformations on non-normal variables like thickness to normalize distributions prior to analysis.

B. Statistical Analysis

- Utilize more advanced machine learning techniques such as multivariate adaptive regression splines, neural networks, random survival forests, and deep learning models (Zhang et al., 2022). These can capture non-linear effects and high-order interactions more effectively than basic linear regression.
- To combine several prognostic factors for survival analysis while accounting for censoring, use multivariate Cox proportional hazards regression.
- Verify all of the model's assumptions, such as proportional hazards, non-linearity, residual normality, etc. Upon confirmation and applicability, use alternate methods such as generalized additive models or data transformations (Lin et al., 2021).

C. Model Validation

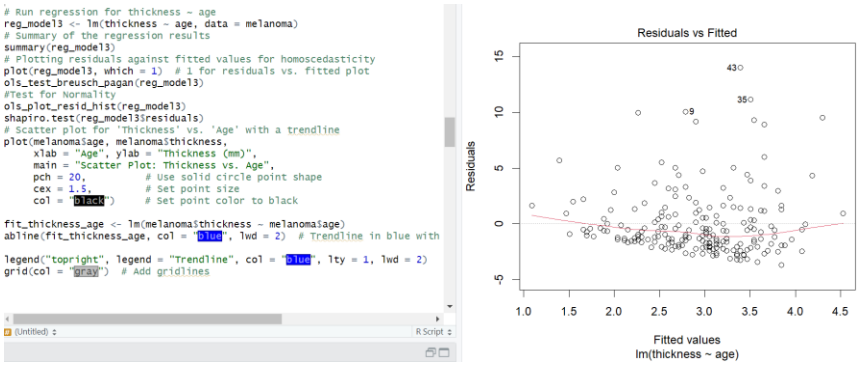
- Utilizing resampling techniques like k-fold cross validation and measures like C-index, AUC, and RMSE, evaluate prediction performance and check for excessive fitting (Hassan et al., 2023).
- Divide the data into test sets for holdouts and training to ensure independent validation then analyse the results.
- Use different datasets for external validation in order to verify generalizability.

D. Interpretation and Clinic Guidance

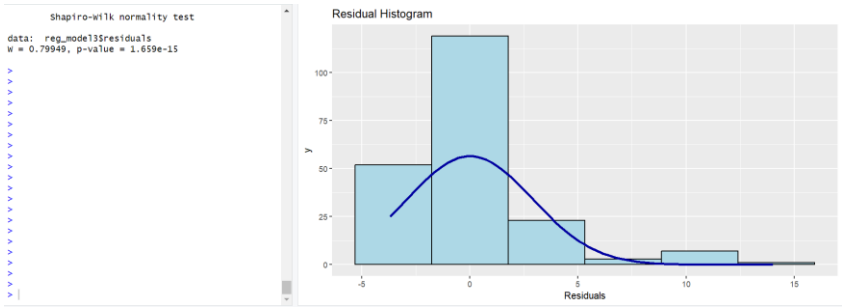
- To find non-linear effects and variable interactions, use subgroup analyses and partial dependence graphs.
- Create Kaplan-Meier curves for several subgroups in order to compare and visualize survival rates with the usage of tests for log-ranks (Chen et al., 2022).
- Utilize clinical knowledge to evaluate data constraints, spot possible unmeasured confounders, and carefully evaluate model outputs in a therapeutic setting.

Appendices

Appendix 1

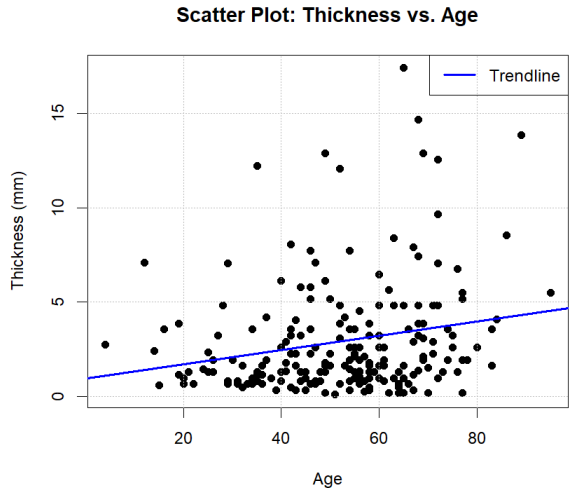


Appendix 2



Shapiro Wilk test and Histogram for Normality (Thickness and Age)

Appendix 3



Scatterplot for Linearity (Thickness vs Age)

Appendix 4

```
# Correlation between 'time' and 'age' with significance level
cor_test_time_age <- cor.test(melanoma$time, melanoma$age)
cat("Correlation between time and age:", cor_test_time_age$estimate, "\n")
cat("p-value:", cor_test_time_age$p.value, "\n\n")
```

Appendix 5

```r
##QQ Plot for Age
# Set up the plotting environment with two plots side by side
par(mfrow = c(1, 2))
# Create QQ plot for 'age' by gender (Male)
qqnorm(male_age_data$age,
       main = "QQ Plot for Age (Male)",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles",
       col = viridis(1))  # Using the first color from the viridis palette
qqline(male_age_data$age, col = viridis(2))  # Add QQ line
grid(col = "gray")  # Add gridlines
# Create QQ plot for 'age' by gender (Female)
qqnorm(female_age_data$age,
       main = "QQ Plot for Age (Female)",
       xlab = "Theoretical Quantiles",
       ylab = "Sample Quantiles",
       col = viridis(3))  # Using the third color from the viridis palette
qqline(female_age_data$age, col = viridis(4))  # Add QQ line
grid(col = "gray")  # Add gridlines
# Reset the plotting environment
par(mfrow = c(1, 1))
```

General syntax for QQ Plots

References

Bekheit, M., Ibrahim, M.Y., Tobar, W., Galal, I. and Elward, A.S., 2020. Correlation between the total small bowel length and anthropometric measures in living humans: cross-sectional study. Obesity Surgery, 30, pp.681-686.

Chen, H., Poon, I., Atenafu, E.G., Badellino, S., Biswas, T., Dagan, R., Erler, D., Foote, M., Redmond, K.J., Ricardi, U. and Sahgal, A., 2022. Development of a prognostic model for overall survival in patients with extracranial oligometastatic disease treated with stereotactic body radiation therapy. International Journal of Radiation Oncology* Biology* Physics, 114(5), pp.892-901.

Dansana, D., Kumar, R., Das Adhikari, J., Mohapatra, M., Sharma, R., Priyadarshini, I. and Le, D.N., 2020. Global forecasting confirmed and fatal cases of COVID-19 outbreak using autoregressive integrated moving average model. Frontiers in public health, 8, p.580327.

Hassan, M.M., Hassan, M.M., Yasmin, F., Khan, M.A.R., Zaman, S., Islam, K.K. and Bairagi, A.K., 2023. A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction. Decision Analytics Journal, 7, p.100245.

Jones, P.J., Mair, P., Simon, T. and Zeileis, A., 2020. Network trees: A method for recursively partitioning covariance structures. Psychometrika, 85(4), pp.926-945.

Joshi, A., Rienks, M., Theofilatos, K. and Mayr, M., 2021. Systems biology in cardiovascular disease: a multiomics approach. Nature Reviews Cardiology, 18(5), pp.313-330.

Kantidakis, G., Putter, H., Lancia, C., Boer, J.D., Braat, A.E. and Fiocco, M., 2020. Survival prediction models since liver transplantation-comparisons between Cox models and machine learning techniques. BMC Medical Research Methodology, 20, pp.1-14.

Lin, L. and Xu, C., 2020. Arcsine-based transformations for meta-analysis of proportions: Pros, cons, and alternatives. Health Science Reports, 3(3), p.e178.

Mundo, A.I., Tipton, J.R. and Muldoon, T.J., 2022. Generalized additive models to analyze nonlinear trends in biomedical longitudinal data using R: Beyond repeated measures ANOVA and linear mixed models. Statistics in Medicine, 41(21), pp.4266-4283.

Muse, A.H., Ngesa, O., Mwalili, S., Alshanbari, H.M. and El-Bagoury, A.A.H., 2022. A flexible Bayesian parametric proportional hazard model: Simulation and applications to right-censored healthcare data. Journal of Healthcare Engineering, 2022.

Oti, E.U., Olusola, M.O. and Esemokumo, P.A., 2021. Statistical Analysis of the Median Test and the Mann-Whitney U Test. International Journal of Advanced Academic Research, 7(9), pp.44-51.

Pargent, F., Pfisterer, F., Thomas, J. and Bischl, B., 2022. Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. Computational Statistics, 37(5), pp.2671-2692.

Țăpoi, D.A., Derewicz, D., Gheorghișan-Gălățeanu, A.A., Dumitru, A.V., Ciongariu, A.M. and Costache, M., 2023. The Impact of Clinical and Histopathological Factors on Disease Progression and Survival in Thick Cutaneous Melanomas. Biomedicines, 11(10), p.2616.

Zhang, G., Bateni, S.M., Jun, C., Khoshkam, H., Band, S.S. and Mosavi, A., 2022. Feasibility of random forest and multivariate adaptive regression splines for predicting long-term mean monthly dew point temperature. Frontiers in Environmental Science, 10, p.826165.