Factors that affect tumor thickness of melanoma

Introduction

The goal of this report is to find which factors has the most effect on the survival rate of melanoma cancer. In order to do this, I have chosen the dataset “melanoma” which was composed of 205 observations with 6 variables. The variables were status(status of the patients), sex(sex of the patients), age(age of the patients), year(year of the operation), thickness(thickness of the tumor in mm), ulcer(presence of ulceration). I expect to find thicker tumors among males and in presence of ulceration. The dataset was already tidy so I did not tidy them.

EDA

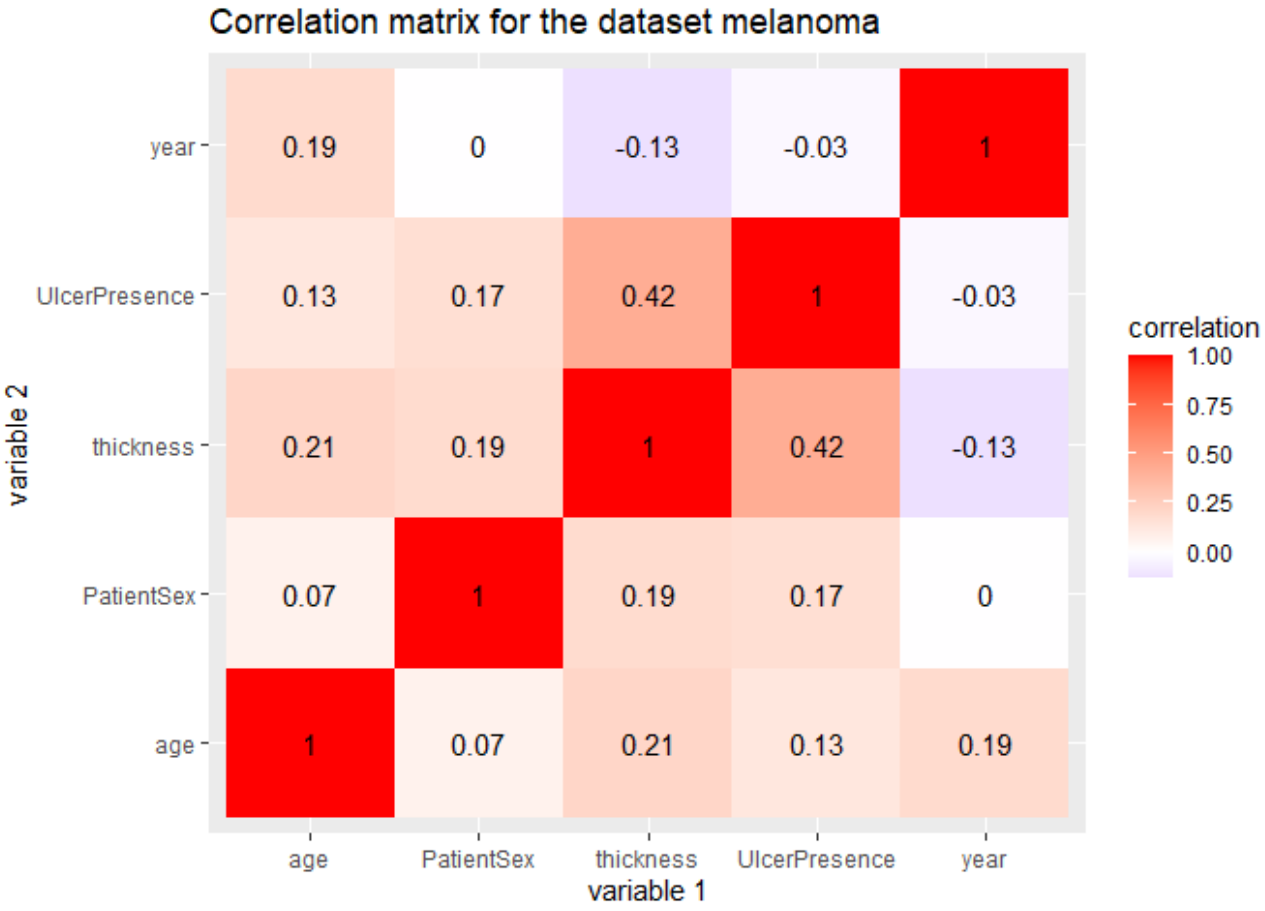
```
# Bring and open the data
melanoma <- read_excel("C:/Users/okcij/Documents/R/melanoma.xlsx")

# Creates a new data with binary outcomes
melanoma_new <-melanoma %>%
  as.data.frame %>%
  mutate(UlcerPresence = ifelse(ulcer == "Present", 1, 0)) %>%
  mutate(PatientSex = ifelse(sex == "Male", 1, 0))

# Create a new data with only numeric variables
melanoma_num <- melanoma_new %>%
  select_if(is.numeric)

# Save as data
cor(melanoma_num, use = "pairwise.complete.obs") %>%
  # Convert each row to a variable
  as.data.frame %>%
  # All correlations appear in the same column
  rownames_to_column %>%
  # Change the scale for a neutral appeal
  pivot_longer(-1, names_to = "other_var", values_to = "correlation") %>%
  ggplot(aes(rowname, other_var, fill=correlation)) +
  geom_tile() +
  # Overlay the values
  scale_fill_gradient2(low="blue",mid="white",high="red") +
```

```
  # Write title and labels the axis
  geom_text(aes(label = round(correlation,2)), color = "black", size = 4) +
  labs(title = "Correlation matrix for the dataset melanoma", x = "variable 1", y = "
```




```
melanoma %>%
  #groups by sex
  group_by(sex) %>%
  #selects the variable thickness
  select(thickness) %>%
  #summarizes the mean thickness between male and female
  summarize(mean_thickness=mean(thickness))
```

```
## Adding missing grouping variables: `sex`
```

```
## # A tibble: 2 x 2
##   sex    mean_thickness
##   <chr>           <dbl>
## 1 Female           2.49
## 2 Male             3.61
```

```
melanoma %>%
  #groups by sex
  group_by(sex) %>%
  #selects the variable thickness
  select(thickness) %>%
  #summarizes the mean thickness between male and female
  summarize(sd_thickness=sd(thickness))
```

```
## Adding missing grouping variables: `sex`
```

```
## # A tibble: 2 x 2
##   sex    sd_thickness
##   <chr>         <dbl>
## 1 Female         2.75
## 2 Male           3.16
```

```
melanoma %>%
  #groups by sex
  group_by(sex) %>%
  #selects the variable thickness
  select(thickness) %>%
  #summarizes the mean thickness between male and female
  summarize(median_thickness=median(thickness))
```

```
## Adding missing grouping variables: `sex`
```

```
## # A tibble: 2 x 2
##   sex    median_thickness
##   <chr>             <dbl>
## 1 Female             1.62
## 2 Male               2.58
```

```
melanoma %>%
  #groups by sex
  group_by(sex) %>%
  #selects the variable thickness
  select(thickness) %>%
  #summarizes the mean thickness between male and female
  summarize(quantile_thickness=quantile(thickness))
```

```
## Adding missing grouping variables: `sex`
```

```
## `summarise()` has grouped output by 'sex'. You can override using the `.groups` ar
```

```
## # A tibble: 10 x 2
## # Groups:   sex [2]
##    sex    quantile_thickness
##    <chr>               <dbl>

##  1 Female               0.1
##  2 Female               0.97
##  3 Female               1.62
##  4 Female               3.06
##  5 Female              17.4
##  6 Male                 0.16
##  7 Male                 1.05
##  8 Male                 2.58
##  9 Male                 4.84
## 10 Male                14.7
```

This correlation heatmap shows that thickness of the tumor and presence of ulceration has a moderate relationship, which means that the thickness of tumor will increase or decrease with a certain level with presence of ulceration and vice versa. Age/sex of the patients and thickness of tumor does not have a high correlation coefficient, which means that their relationship is very weak and we may assume that the thickness of tumor does not depend on age or sex of the patients, but details will be tested in this report using multiple tests.

#Additional summary statistics were performed on the variable thickness. I chose thickness because thickness and ulcer presence had the highest correlation out of all the variables and summary statistics could only be performed on numerical data. The summary statistics included mean, sd, median and IQR. The mean of thickness for male and female was 3.611 and 2.486, the standard deviation of thickness for male and female was 3.156 and 2.755, and the median of thickness for male and female was 2.58 and 1.62. The summary statistics showed that males have thicker tumors than females.

MANOVA

```
# Find the means of thickness and age between each sex
melanoma %>%
  group_by(sex) %>%
  summarize(mean(thickness), mean(age))
```

```
## # A tibble: 2 x 3
##   sex    `mean(thickness)` `mean(age)`
```

```
##   <chr>              <dbl>       <dbl>
## 1 Female              2.49        51.6
## 2 Male                3.61        53.9
```

```
# Perform MANOVA with 2 response variables listed in cbind()
manova_melanoma <- manova(cbind(thickness,age) ~ sex, data = melanoma)

# Output of MANOVA
summary(manova_melanoma)
```

```
##            Df   Pillai approx F num Df den Df  Pr(>F)
## sex         1 0.035255   3.6909      2    202 0.02665 *
## Residuals 203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Since MANOVA is significant then we perform one-way ANOVA for each variable
summary(aov(thickness~sex, data=melanoma))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## sex           1   61.4   61.42   7.227 0.00778 **
## Residuals   203 1725.3    8.50
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(age~sex, data=melanoma))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## sex           1    265   264.8   0.952   0.33
## Residuals   203  56436   278.0
```

```
# Since ANOVA is significant then we can perform post-hoc analysis for thickness
pairwise.t.test(melanoma$thickness,melanoma$sex, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  melanoma$thickness and melanoma$sex
##
```

```
##      Female
## Male 0.0078
##
## P value adjustment method: none
```

The MANOVA test tells us that there is a significant difference in both tumor thickness and age among males and females. Since MANOVA test results are significant, I performed ANOVA test for both of my variables. Only tumor thickness was significant with ANOVA test since it had a p-value of 0.007777, which is less than 0.05. Thus I performed post-hoc t test and and confirmed that males and females have difference in their tumor thickness. I performed total of 2 tests, and according to this the probability of at least one type I error would be 0.975. Significance level did not need to be adjusted because the probability of type I error is small.

#The probablity of type I error would be 0.0975. I seem to have missed a zero in my original calculation. All assumptions were assumed to be met but details were discussed in the later part of the project.

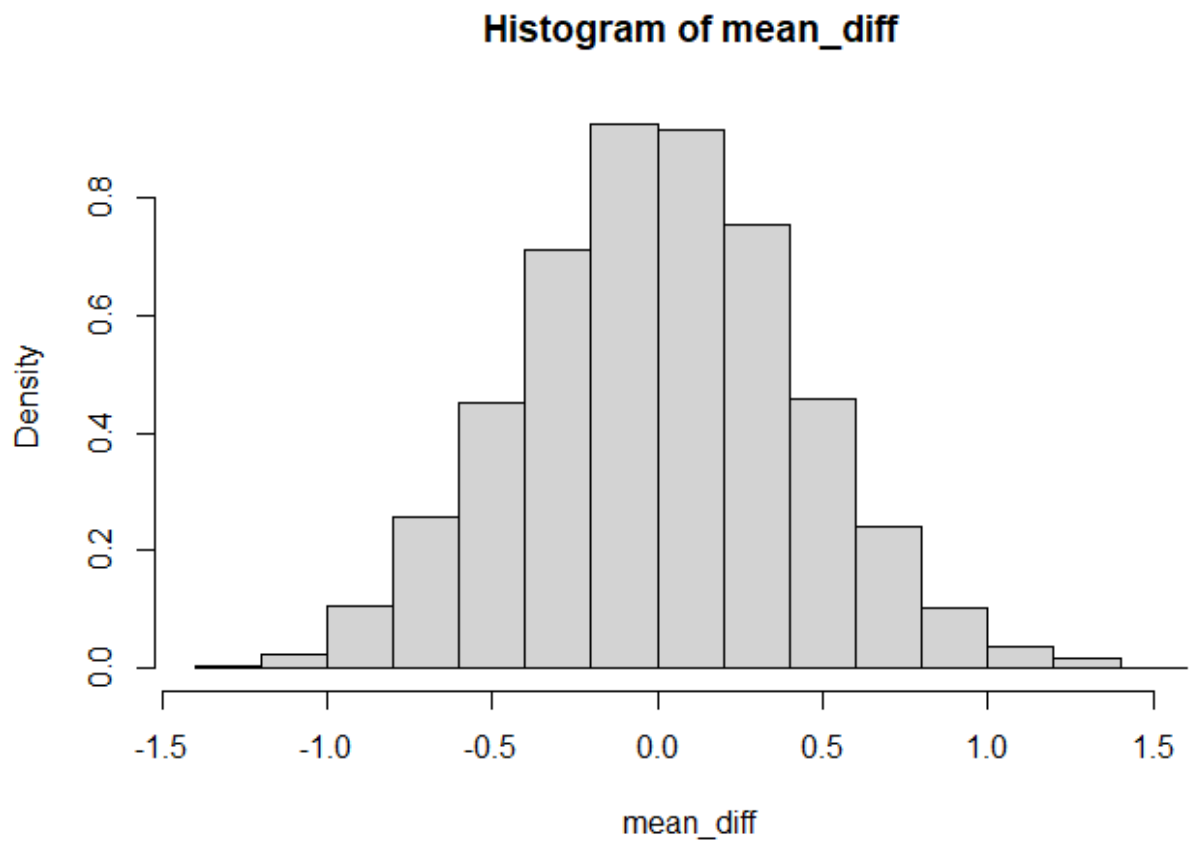
Randomization Test

```
set.seed(348)
mean_diff <- vector()

# Randomization test (using replicate)
for(i in 1:5000){
  temp <- data.frame(sex=melanoma$sex,
                     thickness=sample(melanoma$thickness))
  mean_diff[i] <- temp %>%
    group_by(sex) %>%
    summarize(means = mean(thickness)) %>%
    summarize(mean_diff = diff(means)) %>%
    pull
}

# Represent the distribution of the F-statistics for each randomized sample
hist(mean_diff, prob=T); abline(v = 7.227, col="red", add=T)
```

```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): "add" is
## not a graphical parameter
```



Using randomization test, I tested the further relationship between tumor thickness and sex. My null hypothesis was that there would be no significant difference in the mean tumor thickness between males and females. My alternative hypothesis was that there would be a significant difference in the mean tumor thickness between males and females. The null hypothesis would be rejected because the distribution of histogram is normal and the F value is far from the majority.

#The observed test statistic, which was 7.227, was not visible on the distribution because the F-statistics value was too big. This would indicate that the test results are significant.

Linear Regression Model

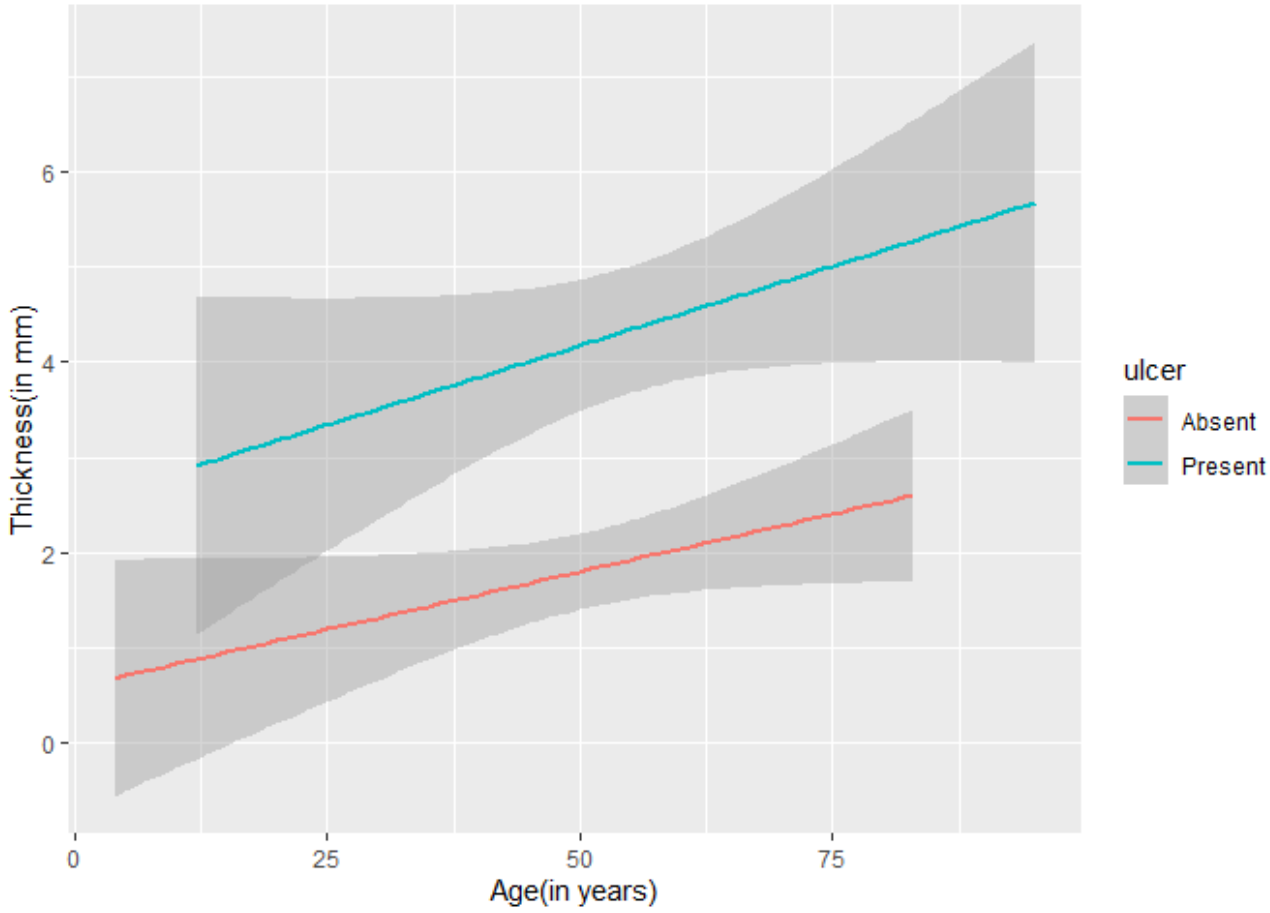
```
# Mean-center numeric variable involved in the interaction
melanoma$age_c <- melanoma$age - mean(melanoma$age, na.rm=TRUE)

# Build the linear regression model
fit <- lm(thickness ~ age_c * ulcer, data = melanoma)
summary(fit)
```

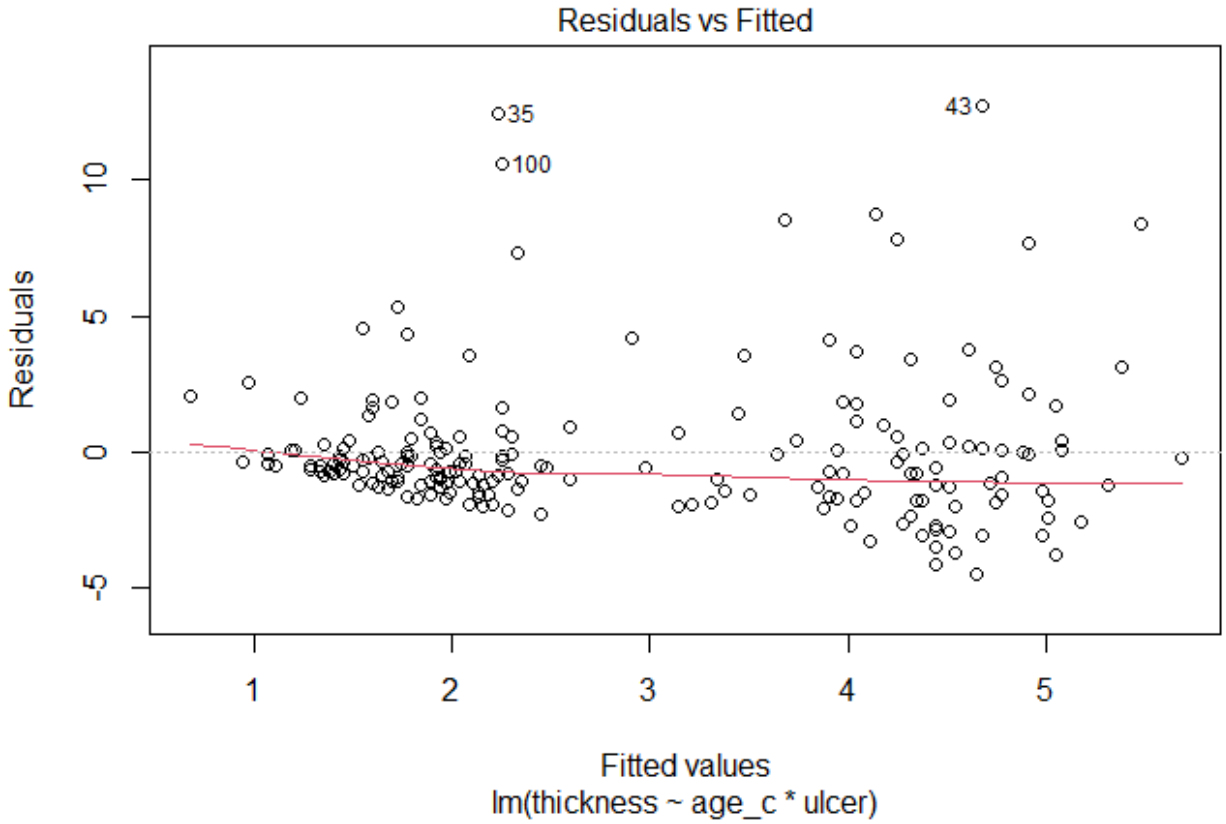
```
##
## Call:
## lm(formula = thickness ~ age_c * ulcer, data = melanoma)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4823 -1.3607 -0.5364  0.3421 12.7444
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.856339   0.249351   7.445 2.79e-12 ***
## age_c              0.024281   0.015640   1.553    0.122
## ulcerPresent       2.400900   0.376852   6.371 1.26e-09 ***
## age_c:ulcerPresent 0.009093   0.022499   0.404    0.687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.656 on 201 degrees of freedom
## Multiple R-squared:  0.2065, Adjusted R-squared:  0.1946
## F-statistic: 17.44 on 3 and 201 DF,  p-value: 4.223e-10
```

```
# Create a graph to visualize the interaction between ulcer and age
ggplot(melanoma, aes(x=age, y=thickness, col=ulcer)) +
  xlab("Age(in years)") +ylab("Thickness(in mm)") +
  geom_smooth(method=lm)
```

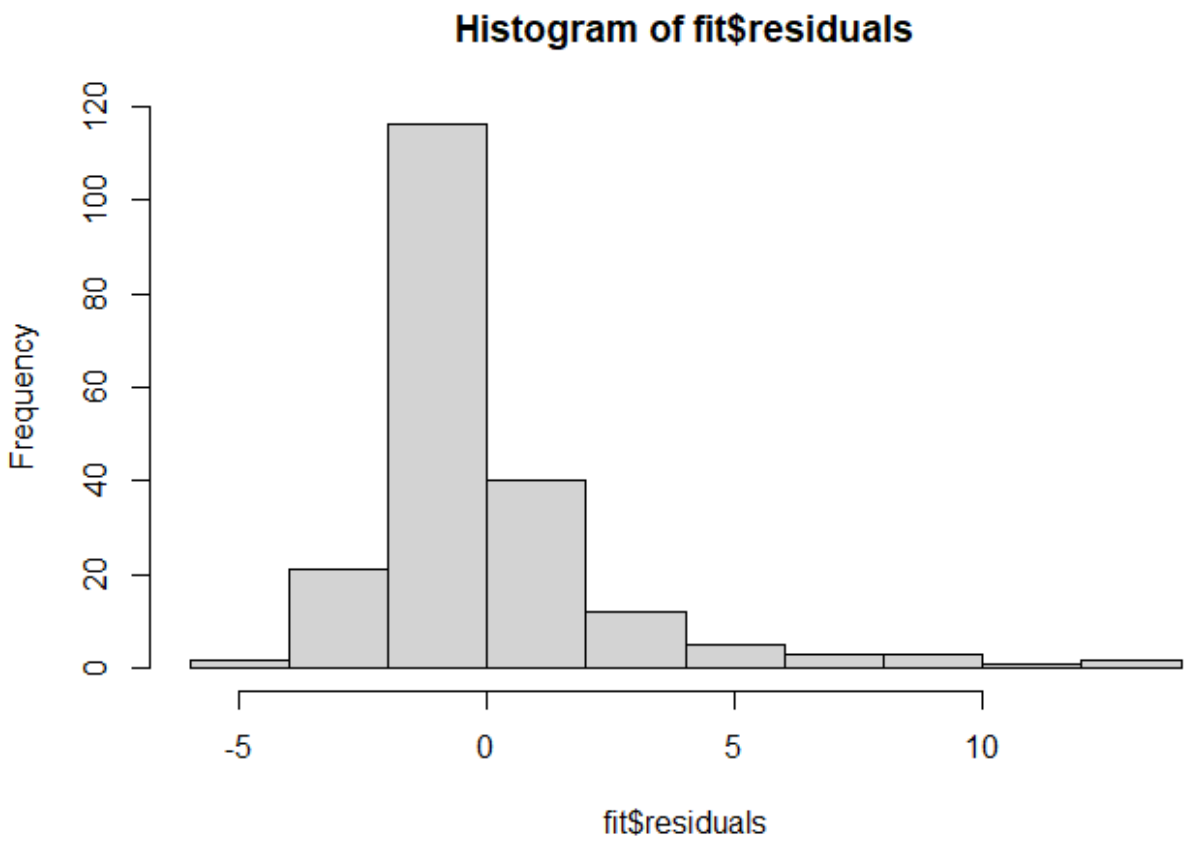
```
## `geom_smooth()` using formula 'y ~ x'
```



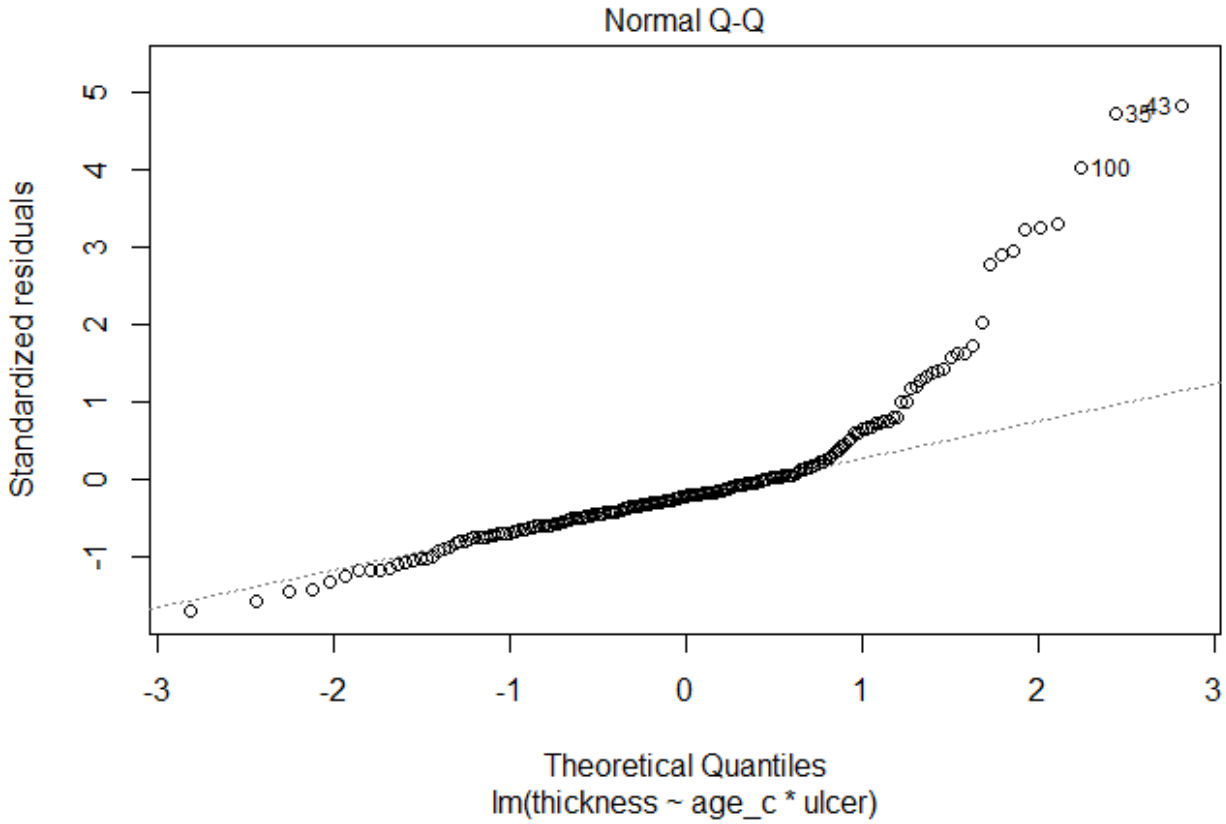

```
# Use scatter plot to check for linearity
plot(fit, which = 1)
```




```
# Use histogram and Q-Q plot to check for normality
hist(fit$residuals)
```




```
plot(fit, which = 2)
```




```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.0.5
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.0.5
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 4.0.5
```

```r
# Use Breusch-Pagan test to check for homoscedasticity
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 7.0503, df = 3, p-value = 0.07031
```

```r
# Robust standard errors
coeftest(fit, vcov = vcovHC(fit))
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)        1.856339   0.221222  8.3913 8.474e-15 ***
## age_c              0.024281   0.013991  1.7354   0.08419 .
## ulcerPresent       2.400900   0.399986  6.0025 8.938e-09 ***
## age_c:ulcerPresent 0.009093   0.024167  0.3763   0.70713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Bootstrap standard errors
samp_SEs <- replicate(5000, {
  # Bootstrap your data (resample observations)
  boot_data <- sample_frac(melanoma, replace = TRUE)
  # Fit regression model
  boot_fit <- lm(thickness ~ age_c * ulcer, data = boot_data)
  # Save the coefficients
  coef(boot_fit)
})

# Estimated SEs
samp_SEs %>%
  # Transpose the obtained matrices
  t %>%
  # Consider the matrix as a data frame
  as.data.frame %>%
  # Compute the standard error (standard deviation of the sampling distribution)
  summarize_all(sd)
```

```
##   (Intercept)     age_c ulcerPresent age_c:ulcerPresent
## 1   0.2170567 0.0140254    0.3892841         0.02350194
```

I used linear regression model to predict tumor thickness from age and presence of ulceration. The intercept, or the tumor thickness, was 1.856mm with mean age and ulcer presence. The coefficient of age was 0.0242 which means that if the average age was to increase by 1 year, the tumor thickness would increase by 0.0242. The coefficient of ulcer presence was 2.4, which means that the tumor thickness would increase by 2.4mm if ulceration was present. The coefficient of both age and ulcer presence was 0.00909, which means that if the average age was to increase by 1 year and ulceration was present, then the average tumor thickness would increase by 0.00909. However, p-value for both age and interaction of age and ulceration was not significant, thus I only considered ulcer presence as a significant explanatory variable. This model would only explain about 20% of the variation in the response.

Assumptions of linearity was checked using a scatter plot, which was not met because the scatter plot had a funnel-like shape. Assumptions of normality was checked using both a histogram and a Q-Q plot, which was somewhat met because both of them had somewhat normal distribution. Assumptions of homoscedasticity was checked through Breusch-Pagan test, which was not met because the p-value was 0.0703, which was greater than 0.05.

Although some test results were not significant, the regression results were ran again using robust standard errors and bootstrapped standard errors. Robust standard errors resulted in larger SE for age and smaller SE for ulcer presence and interaction of age and ulcer presence, but eventually gave the same result where only ulcer presence was significant. Bootstrapped standard errors resulted in smaller SE for age and larger SE for ulcer presence and interaction of age and ulcer presence, but also gave the same result where only ulcer presence was significant.

#The interaction graph shows that the Absent and Present lines are parallel. This indicates that there is no interaction between the two and confirms that test result.

Logistic Regression

```
# Fit a new logistic regression model
fit2 <- glm(UlcerPresence ~ thickness + age, data = melanoma_new, family = binomial(l
summary(fit2)
```

```
##
## Call:
## glm(formula = UlcerPresence ~ thickness + age, family = binomial(link = "logit"),
##     data = melanoma_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2250  -0.8396  -0.6914   1.0479   1.7663
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.708254   0.544589  -3.137  0.00171 **
## thickness    0.448208   0.087807   5.104 3.32e-07 ***
## age          0.004885   0.009687   0.504  0.61405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 281.13  on 204  degrees of freedom
## Residual deviance: 235.51  on 202  degrees of freedom
## AIC: 241.51
##
## Number of Fisher Scoring iterations: 5
```

```
# Add predicted probabilities to the dataset
melanoma_new$prob <- predict(fit2, type = "response")

# Predicted outcome is based on the probability of UlcerPresence
# if the probability is greater than 0.5, the clump is found to be present
melanoma_new$predicted <- ifelse(melanoma_new$prob > 0.5, "Present", "Absent")

# Confusion matrix
table(truth = melanoma_new$ulcer, prediction = melanoma_new$predicted)
```

```
##          prediction
## truth     Absent Present
##   Absent     102      13
##   Present     37      53
```

```
# Accuracy (correctly classified cases)
(102 + 53)/205
```

```
## [1] 0.7560976
```

```
# Sensitivity (True Positive Rate, TPR)
53/90
```

```
## [1] 0.5888889
```

```
# Specificity (True Negative Rate, TNR)
102/115
```

```
## [1] 0.8869565
```

```
# Precision (Positive Predictive Value, PPV)
53/68
```

```
## [1] 0.7794118
```

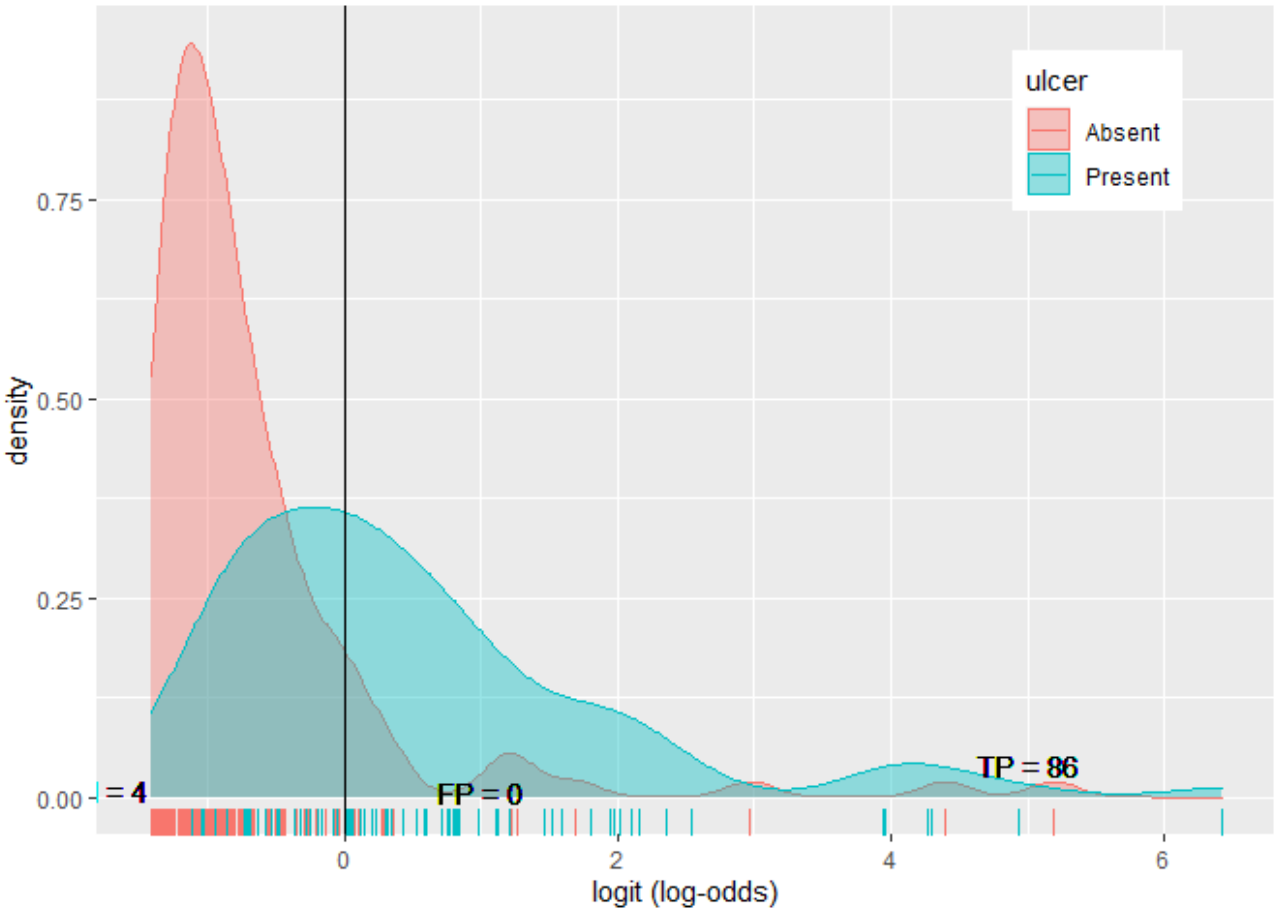
```
# Predicted log odds
melanoma_new$logit <- predict(fit2, type = "link")

# Exponentiate coefficients
exp(coef(fit2))
```

```
## (Intercept)   thickness         age
##   0.1811819   1.5655049   1.0048972
```

```
# Density plot of log-odds for each outcome
melanoma_new %>%
  ggplot() +
  geom_density(aes(logit, color = ulcer, fill = ulcer), alpha = .4) +
    geom_rug(aes(logit, color = ulcer)) +
  geom_text(x = -5, y = .07, label = "TN = 115") +
  geom_text(x = -1.75, y = .008, label = "FN = 4") +
  geom_text(x = 1, y = .006, label = "FP = 0") +
  geom_text(x = 5, y = .04, label = "TP = 86") +
  theme(legend.position = c(.85,.85)) +
```

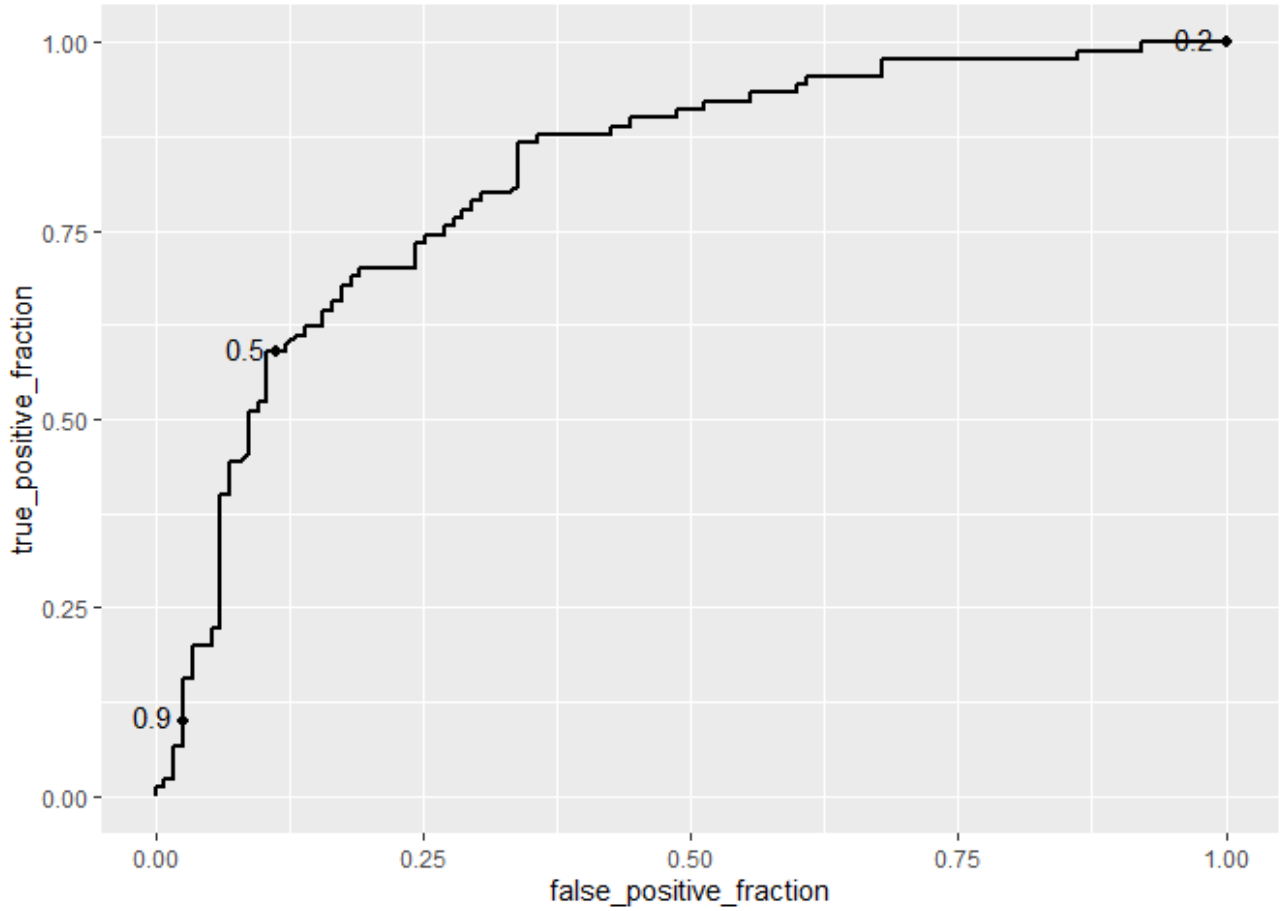
```
  geom_vline(xintercept = 0) +
  xlab("logit (log-odds)")
```




```
library(plotROC)
```

```
## Warning: package 'plotROC' was built under R version 4.0.5
```

```
# Plot ROC depending on values of y and its probabilities displaying some cutoff valu
ROCplot_melanoma <- ggplot(melanoma_new) +
  geom_roc(aes(d = UlcerPresence, m = prob),cutoffs.at = list(0.1, 0.5, 0.9))
ROCplot_melanoma
```




```
# Calculate AUC value
calc_auc(ROCplot_melanoma)
```

```
##   PANEL group       AUC
## 1     1    -1 0.8201449
```

I used the logistic regression model to predict presence of ulcer using tumor thickness and age of the patients. According to the model, 1 mm increase in tumor thickness will increase ulceration by 0.448 and 1 year increase in age will increase ulceration by 0.00488. However, only thickness would be significant because its p-value was less than 0.05.

A confusion matrix was also created for this model. This model has accuracy of 0.756, which means about 75% of this model is accurate. It had sensitivity of 0.588 and specificity of 0.886. Although sensitivity was not very high, specificity was high which means this model predicts true negative models better than the true positive models. PPV was 0.779, which means about 78% of true positive predictions were correctly identified.

A ROC plot was created as well to look at the results in a graphic way. The plot will predict true positive models slightly better than the false positive models because the curve is

closer to the left. The AUC value is 0.82, which is pretty high, so this model has a pretty good prediction power.

#While holding thickness constant, for every 1 year increase in age, the odds of tumor thickness change by a factor of 1.56. This means that the tumor thickness will increase by 56%. While holding age constant, for every 1 mm increase in thickness, the odds of age changed by a factor of 1.00489. This means that the age will increase by 0.489%.

#All adjustments were made regarding the comments. I got points off from EDA section for missing summary statistics, which was included. I then got points off from MANOVA section for incorrect Pr and missing assumptions, which was recalculated and then mentioned. I also got points off from Randomization test and Linear Regression section for missing some explanations and interpretations, which was added. Lastly, I got points off from Logistic Regression section for not exponentiating coefficients and doing wrong interpretations. Cofficients were exponentiated and interpreted while holding other predictors constant.