PART2 Group_M17

2025-01-04

Q1 a

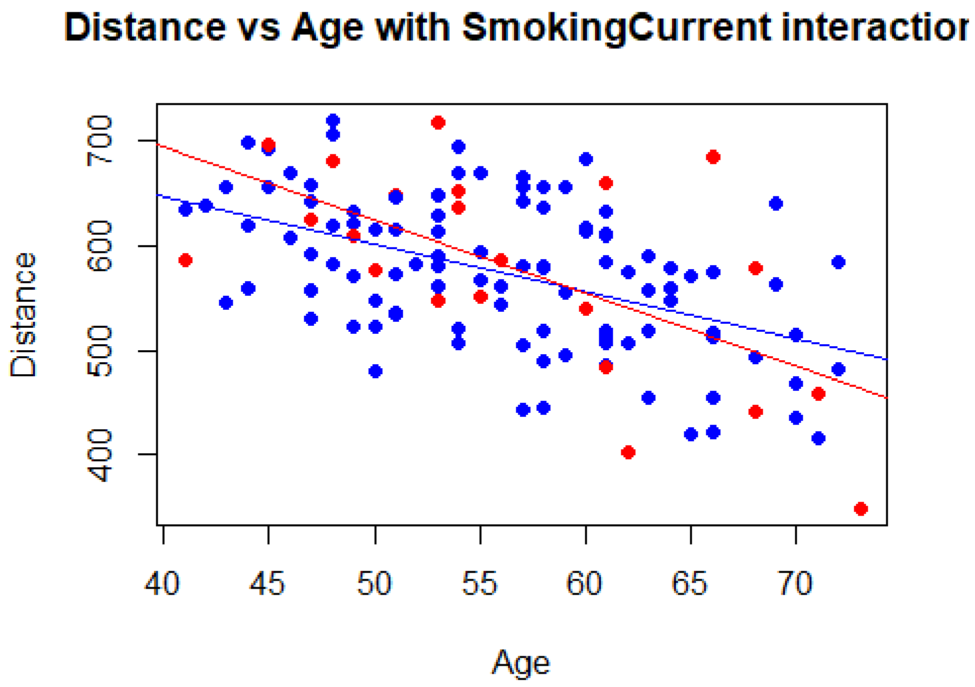
```r
library(dplyr)

Patient_info <- readRDS("C:/Users/lenovo/Desktop/assignment/363/Part2GrM17.rds")

# Convert categorical variables to factors for proper analysis
Patient_info$Sex <- as.factor(Patient_info$Sex)
Patient_info$SmokingCurrent <- as.factor(Patient_info$SmokingCurrent)
Patient_info$Activity <- as.factor(Patient_info$Activity)

# Specify covariates
covariates <- c("Age", "Height", "Weight", "BMI", "RestingHeartrate")  # Quantitative variables

# 1. Age and smoking habits Model (fitting complete data)
model_Age <- lm(Distance ~ Age * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: Age and SmokingCurrent\n")
```

```
##
## Model for Covariate: Age and SmokingCurrent
```

```r
print(summary(model_Age))
```

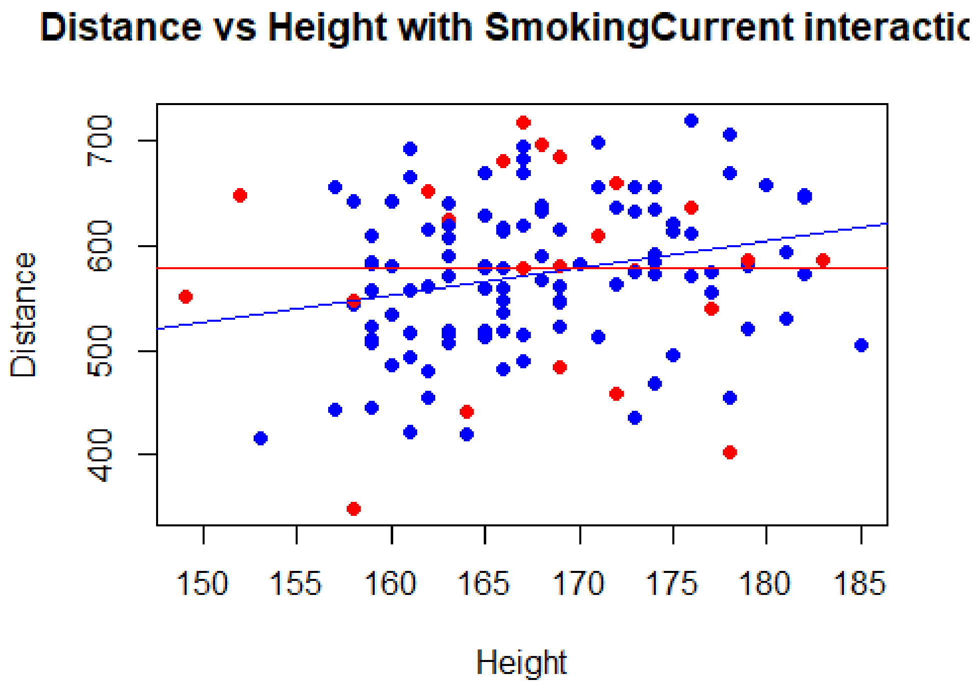
```
##
## Call:
## lm(formula = Distance ~ Age * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -138.438  -45.238   -5.014   44.664  171.299
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          828.5519    46.3437  17.878  < 2e-16 ***
## Age                   -4.5515     0.8157  -5.580 1.46e-07 ***
## SmokingCurrent1      141.8113   102.5647   1.383    0.169    
## Age:SmokingCurrent1   -2.3828     1.7938  -1.328    0.187    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.37 on 123 degrees of freedom
## Multiple R-squared:  0.2894, Adjusted R-squared:  0.2721 
## F-statistic:  16.7 on 3 and 123 DF,  p-value: 3.653e-09
```

```r
#Create a chart of age and distance interacting with smoking current (filtered data)
filtered_data_Age <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_Age$Age, filtered_data_Age$Distance,
    col = ifelse(filtered_data_Age$SmokingCurrent == 0, "blue", "red"),
    pch = 16, xlab = "Age", ylab = "Distance", main = "Distance vs Age with 
SmokingCurrent interaction")
abline(lm(Distance ~ Age, data = filtered_data_Age[filtered_data_Age$SmokingCurrent 
== 0, ]), col = "blue")
abline(lm(Distance ~ Age, data = filtered_data_Age[filtered_data_Age$SmokingCurrent 
== 1, ]), col = "red")
```



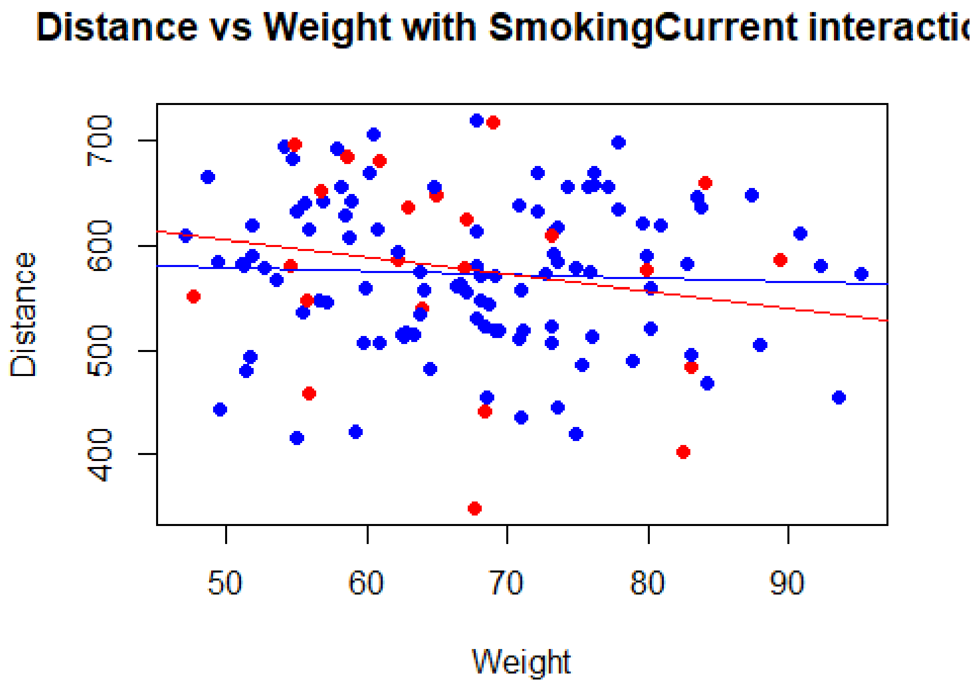
```
# 2. Height and SmokingCurrent Model (Fit to full data)
model_Height <- lm(Distance ~ Height * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: Height and SmokingCurrent\n")

##
## Model for Covariate: Height and SmokingCurrent

print(summary(model_Height))

##
## Call:
## lm(formula = Distance ~ Height * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -228.375  -43.951    2.462   55.356  139.492
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             143.275    173.064   0.828   0.4093
## Height                    2.558      1.031   2.482   0.0144 *
## SmokingCurrent1         431.766    361.229   1.195   0.2343
## Height:SmokingCurrent1   -2.543      2.149  -1.183   0.2390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.49 on 123 degrees of freedom
## Multiple R-squared:  0.04835,    Adjusted R-squared:  0.02513
## F-statistic: 2.083 on 3 and 123 DF,  p-value: 0.1059
```

```
# Create plot for Height vs Distance with SmokingCurrent interaction (Filtered data)
filtered_data_Height <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_Height$Height, filtered_data_Height$Distance,
    col = ifelse(filtered_data_Height$SmokingCurrent == 0, "blue", "red"),
    pch = 16, xlab = "Height", ylab = "Distance", main = "Distance vs Height with
SmokingCurrent interaction")
abline(lm(Distance ~ Height, data =
filtered_data_Height[filtered_data_Height$SmokingCurrent == 0, ]), col = "blue")
abline(lm(Distance ~ Height, data =
filtered_data_Height[filtered_data_Height$SmokingCurrent == 1, ]), col = "red")
```



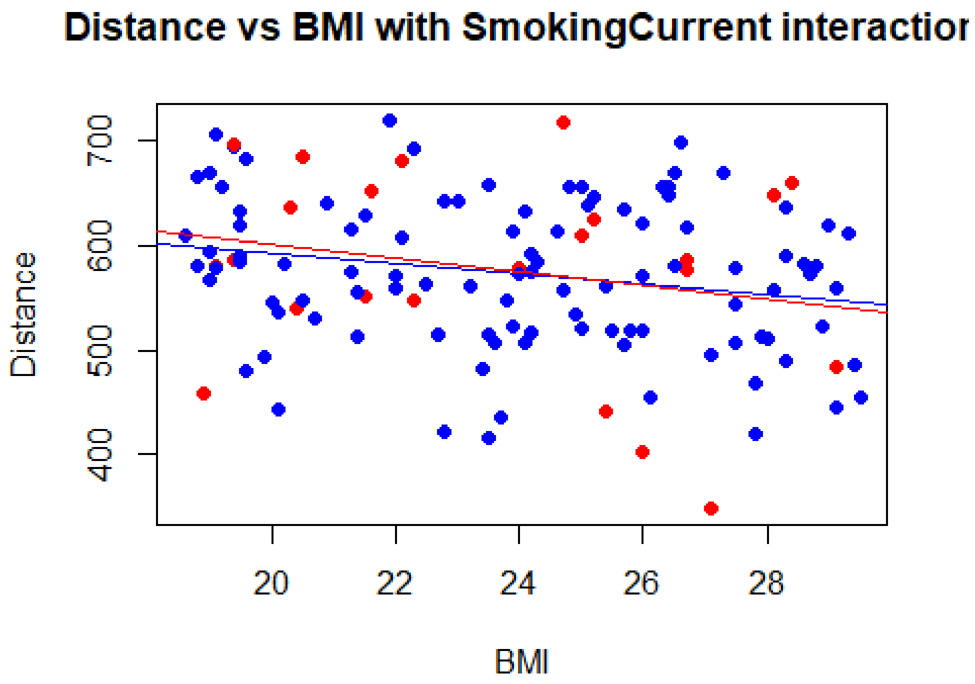
```
# 3. Weight and SmokingCurrent Model (Fit to full data)
model_Weight <- lm(Distance ~ Weight * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: Weight and SmokingCurrent\n")

##
## Model for Covariate: Weight and SmokingCurrent

print(summary(model_Weight))

##
## Call:
## lm(formula = Distance ~ Weight * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -226.548  -53.540    3.379   56.975  146.678
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)             593.9662    46.2791  12.834   <2e-16 ***
## Weight                   -0.3192     0.6748  -0.473    0.637    
## SmokingCurrent1          92.9372   108.9744   0.853    0.395    
## Weight:SmokingCurrent1   -1.3256     1.6124  -0.822    0.413    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.87 on 123 degrees of freedom
## Multiple R-squared:  0.01263,    Adjusted R-squared:  -0.01146 
## F-statistic: 0.5243 on 3 and 123 DF,  p-value: 0.6664
```

```r
# Create plot for Weight vs Distance with SmokingCurrent interaction (Filtered data)
filtered_data_Weight <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_Weight$Weight, filtered_data_Weight$Distance,
    col = ifelse(filtered_data_Weight$SmokingCurrent == 0, "blue", "red"),
    pch = 16, xlab = "Weight", ylab = "Distance", main = "Distance vs Weight with SmokingCurrent interaction")
abline(lm(Distance ~ Weight, data = filtered_data_Weight[filtered_data_Weight$SmokingCurrent == 0, ]), col = "blue")
abline(lm(Distance ~ Weight, data = filtered_data_Weight[filtered_data_Weight$SmokingCurrent == 1, ]), col = "red")
```



```
# 4. BMI and SmokingCurrent Model (Fit to full data)
model_BMI <- lm(Distance ~ BMI * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: BMI and SmokingCurrent\n")

##
## Model for Covariate: BMI and SmokingCurrent

print(summary(model_BMI))

##
## Call:
## lm(formula = Distance ~ BMI * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -205.495  -47.935   -0.342   60.267  146.890
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         688.681     54.938  12.536   <2e-16 ***
## BMI                  -4.847      2.269  -2.136   0.0346 *
## SmokingCurrent1       42.132    126.641   0.333   0.7399
## BMI:SmokingCurrent1   -1.659      5.307  -0.313   0.7552
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.42 on 123 degrees of freedom
## Multiple R-squared:  0.05015,    Adjusted R-squared:  0.02698
## F-statistic: 2.165 on 3 and 123 DF,  p-value: 0.09562
```

```
# Create plot for BMI vs Distance with SmokingCurrent interaction (Filtered data)
filtered_data_BMI <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_BMI$BMI, filtered_data_BMI$Distance,
    col = ifelse(filtered_data_BMI$SmokingCurrent == 0, "blue", "red"),
    pch = 16, xlab = "BMI", ylab = "Distance", main = "Distance vs BMI with
SmokingCurrent interaction")
abline(lm(Distance ~ BMI, data = filtered_data_BMI[filtered_data_BMI$SmokingCurrent
== 0, ]), col = "blue")
abline(lm(Distance ~ BMI, data = filtered_data_BMI[filtered_data_BMI$SmokingCurrent
== 1, ]), col = "red")
```



```
# 5. Model for RestingHeartrate and SmokingCurrent (Fit to full data)
model_RestingHeartrate <- lm(Distance ~ RestingHeartrate * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: RestingHeartrate and SmokingCurrent\n")
```

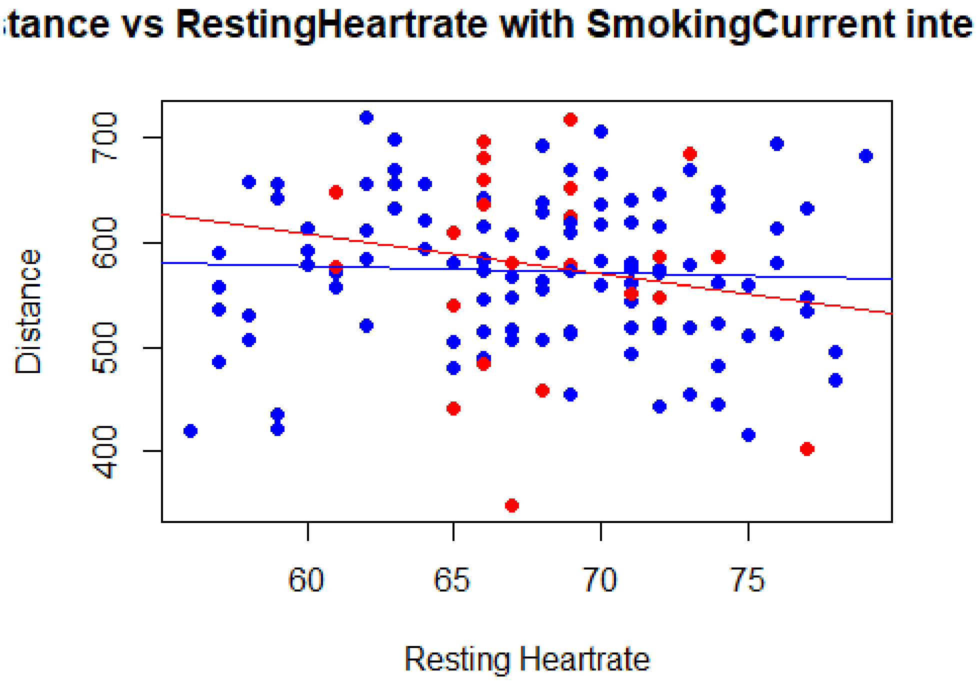
```
##
## Model for Covariate: RestingHeartrate and SmokingCurrent
```

```
print(summary(model_RestingHeartrate))
```

```
##
## Call:
## lm(formula = Distance ~ RestingHeartrate * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -232.342  -52.204    4.505   57.299  143.299
##
## Coefficients:
```

```
##                                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      611.8736    87.8832   6.962 1.77e-10 ***
## RestingHeartrate                  -0.5828     1.2914  -0.451    0.653    
## SmokingCurrent1                  225.4326   296.5735   0.760    0.449    
## RestingHeartrate:SmokingCurrent1  -3.2376     4.3548  -0.743    0.459    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 76.01 on 123 degrees of freedom
## Multiple R-squared:  0.009139,   Adjusted R-squared:  -0.01503 
## F-statistic: 0.3782 on 3 and 123 DF,  p-value: 0.7689
```

```r
# Create plot for RestingHeartrate vs Distance with SmokingCurrent interaction (Filtered data)
filtered_data_RestingHeartrate <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_RestingHeartrate$RestingHeartrate, filtered_data_RestingHeartrate$Distance,
    col = ifelse(filtered_data_RestingHeartrate$SmokingCurrent == 0, "blue", "red"),
    pch = 16, xlab = "Resting Heartrate", ylab = "Distance", main = "Distance vs RestingHeartrate with SmokingCurrent interaction")
abline(lm(Distance ~ RestingHeartrate, data = filtered_data_RestingHeartrate[filtered_data_RestingHeartrate$SmokingCurrent == 0, ]), col = "blue")
abline(lm(Distance ~ RestingHeartrate, data = filtered_data_RestingHeartrate[filtered_data_RestingHeartrate$SmokingCurrent == 1, ]), col = "red")
```



Results

Model for Age and SmokingCurrent The model for Age and SmokingCurrent shows that the interaction between Age and SmokingCurrent is not significant (p-value = 0.187), suggesting that the relationship between Age and Distance does not differ substantially between smokers and non-smokers. Therefore, a single model might be sufficient for this covariate. Plot for Age and SmokingCurrent Interaction: The scatter plots for smokers and non-smokers reveal that while there is a negative trend for both groups, the slopes appear similar. This visual confirmation aligns with the statistical findings that suggest no need for separate models.

Model for Height and SmokingCurrent In the model for Height and SmokingCurrent, the interaction term is also not significant (p-value = 0.239), which implies that the relationship between Height and Distance is similar for both smokers and non-smokers. This suggests that a single model for Height can be used across both groups. Plot for Height and SmokingCurrent Interaction: The plot shows that while there is a slight positive trend for both groups, there is no clear difference in the slopes between smokers and non-smokers, further supporting the conclusion that a single model is appropriate.

Model for Weight and SmokingCurrent The model for Weight and SmokingCurrent reveals that neither the main effect of Weight nor the interaction with SmokingCurrent is significant (p-value for interaction = 0.413). This indicates that Weight does not have a significantly different relationship with Distance for smokers and non-smokers. Plot for Weight and SmokingCurrent Interaction: The scatter plots for both groups show that the relationship between Weight and

Distance is weak and appears to follow similar trends for smokers and non-smokers.

Model for BMI and SmokingCurrent The interaction term in the BMI model is not significant (p-value = 0.755), suggesting that BMI has a similar effect on Distance regardless of smoking status. While BMI is significantly associated with Distance, the effect is consistent across both groups. Plot for BMI and SmokingCurrent Interaction: The scatter plots confirm that both smokers and non-smokers show a similar negative relationship between BMI and Distance.

Model for RestingHeartrate and SmokingCurrent The model for RestingHeartrate and SmokingCurrent indicates that the interaction term is not significant (p-value = 0.459), suggesting that RestingHeartrate does not differ in its effect on Distance between smokers and non-smokers. Plot for RestingHeartrate and SmokingCurrent Interaction: The plots demonstrate that there is no clear distinction between the slopes for smokers and non-smokers, reinforcing the idea that a single model is adequate for this covariate.

Conclusion Based on the results of the linear models and the visual examination of the scatter plots, we conclude that separate models for smokers and non-smokers are not required for any of the covariates analyzed. The interaction terms between SmokingCurrent and the covariates (Age, Height, Weight, BMI, and RestingHeartrate) are not statistically significant, indicating that the relationship between these covariates and the distance covered does not significantly differ between smokers and non-smokers. Therefore, a single model for each covariate, without the need for separate intercepts or slopes, is sufficient to explain the relationship with distance.

Q1b Include all variables and any additional terms such as interactions and higher order terms in the initial model. Then remove variables that don’t appear useful one at a time, starting from the least useful. Continue until all remaining variables are useful (or significant).

```
# Fit the initial model with main effects, interaction terms, and quadratic terms
full_model <- lm(Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
BMI +
               RestingHeartrate + Age:Sex + Age:SmokingCurrent + Age:Activity +
               Height:Sex + Height:SmokingCurrent + Height:Activity +
               Weight:Sex + Weight:SmokingCurrent + Weight:Activity +
               BMI:Sex + BMI:SmokingCurrent + BMI:Activity +
               RestingHeartrate:Sex + RestingHeartrate:SmokingCurrent +
RestingHeartrate:Activity +
               I(Age^2) + I(Height^2) + I(Weight^2) + I(BMI^2) +
               I(RestingHeartrate^2) ,
                 data = Patient_info)

# Summarize the initial model
summary(full_model)

##
## Call:
## lm(formula = Distance ~ Age + Sex + SmokingCurrent + Activity +
##     Height + Weight + BMI + RestingHeartrate + Age:Sex + Age:SmokingCurrent +
##     Age:Activity + Height:Sex + Height:SmokingCurrent + Height:Activity +
##     Weight:Sex + Weight:SmokingCurrent + Weight:Activity + BMI:Sex +
##     BMI:SmokingCurrent + BMI:Activity + RestingHeartrate:Sex +
##     RestingHeartrate:SmokingCurrent + RestingHeartrate:Activity +
##     I(Age^2) + I(Height^2) + I(Weight^2) + I(BMI^2) + I(RestingHeartrate^2),
##     data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -130.451  -36.387    3.869   38.896  130.538
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1.243e+04  1.646e+04  -0.755   0.4521
## Age                               6.897e+00  9.977e+00   0.691   0.4911
## Sex1                              1.803e+03  3.676e+03   0.490   0.6250
## SmokingCurrent1                  -1.841e+02  3.066e+03  -0.060   0.9523
## Activity1                        -9.960e+02  2.924e+03  -0.341   0.7341
## Activity2                        -2.446e+03  3.977e+03  -0.615   0.5401
## Height                            1.387e+02  1.401e+02   0.991   0.3245
## Weight                           -2.797e+01  1.414e+02  -0.198   0.8436
## BMI                               7.776e+01  3.988e+02   0.195   0.8459
## RestingHeartrate                  1.234e+01  2.649e+01   0.466   0.6425
## I(Age^2)                         -7.908e-02  9.190e-02  -0.861   0.3917
## I(Height^2)                      -3.829e-01  2.636e-01  -1.452   0.1498
## I(Weight^2)                     1.245e-01  4.788e-01   0.260   0.7954
## I(BMI^2)                       -1.054e+00  3.880e+00  -0.272   0.7866
## I(RestingHeartrate^2)          -7.864e-02  1.886e-01  -0.417   0.6777
## Age:Sex1                       -1.116e+00  1.602e+00  -0.697   0.4878
## Age:SmokingCurrent1            -4.615e+00  2.062e+00  -2.238   0.0276 *
## Age:Activity1                  -2.629e+00  2.309e+00  -1.139   0.2578
## Age:Activity2                   5.735e-01  2.460e+00   0.233   0.8162
## Sex1:Height                    -9.092e+00  2.200e+01  -0.413   0.6804
## SmokingCurrent1:Height          3.869e+00  1.713e+01   0.226   0.8219
## Activity1:Height                6.375e+00  1.728e+01   0.369   0.7130
## Activity2:Height                1.265e+01  2.355e+01   0.537   0.5925
## Sex1:Weight                     8.930e+00  2.513e+01   0.355   0.7231
## SmokingCurrent1:Weight         -9.490e+00  2.140e+01  -0.444   0.6584
## Activity1:Weight               -8.631e+00  2.038e+01  -0.424   0.6729
## Activity2:Weight               -1.533e+01  2.797e+01  -0.548   0.5850
## Sex1:BMI                       -2.742e+01  6.950e+01  -0.395   0.6941
## SmokingCurrent1:BMI             2.377e+01  6.220e+01   0.382   0.7032
## Activity1:BMI                   2.634e+01  5.770e+01   0.456   0.6492
## Activity2:BMI                   5.094e+01  7.873e+01   0.647   0.5192
## Sex1:RestingHeartrate          -3.182e+00  2.392e+00  -1.331   0.1866
## SmokingCurrent1:RestingHeartrate -1.997e+00  5.002e+00  -0.399   0.6907
## Activity1:RestingHeartrate      4.991e-01  3.087e+00   0.162   0.8719
## Activity2:RestingHeartrate      1.786e+00  3.513e+00   0.508   0.6123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.08 on 92 degrees of freedom
## Multiple R-squared:  0.5369, Adjusted R-squared:  0.3658
## F-statistic: 3.137 on 34 and 92 DF,  p-value: 7.612e-06
```

```
# Apply stepwise regression to remove non-significant variables
final_model <- step(full_model, direction = "both", trace = 1)
```

```
## Start:  AIC=1069.37
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + Age:Sex + Age:SmokingCurrent + Age:Activity +
##     Height:Sex + Height:SmokingCurrent + Height:Activity + Weight:Sex +
##     Weight:SmokingCurrent + Weight:Activity + BMI:Sex + BMI:SmokingCurrent +
##     BMI:Activity + RestingHeartrate:Sex + RestingHeartrate:SmokingCurrent +
##     RestingHeartrate:Activity + I(Age^2) + I(Height^2) + I(Weight^2) +
##     I(BMI^2) + I(RestingHeartrate^2)
##
##                                   Df Sum of Sq    RSS    AIC
## - Activity:Height                  2    1048.3 333161 1065.8
## - Activity:Weight                  2    1133.1 333246 1065.8
## - Activity:RestingHeartrate        2    1219.7 333332 1065.8
## - Activity:BMI                     2    1537.4 333650 1066.0
## - SmokingCurrent:Height            1     184.0 332297 1067.4
## - I(Weight^2)                      1     244.2 332357 1067.5
## - I(BMI^2)                         1     266.1 332379 1067.5
## - Sex:Weight                       1     455.9 332568 1067.5
## - SmokingCurrent:BMI               1     527.1 332640 1067.6
## - Sex:BMI                          1     562.0 332675 1067.6
## - SmokingCurrent:RestingHeartrate  1     575.2 332688 1067.6
## - Sex:Height                       1     616.6 332729 1067.6
## - I(RestingHeartrate^2)            1     627.7 332740 1067.6
## - SmokingCurrent:Weight            1     710.2 332823 1067.6
## - Age:Sex                          1    1752.0 333864 1068.0
## - I(Age^2)                         1    2673.2 334786 1068.4
## <none>                                         332112 1069.4
## - Sex:RestingHeartrate             1    6391.4 338504 1069.8
## - Age:Activity                     2   13004.2 345117 1070.2
## - I(Height^2)                      1    7615.4 339728 1070.2
## - Age:SmokingCurrent               1   18079.5 350192 1074.1
##
## Step:  AIC=1065.77
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Sex:Height + SmokingCurrent:Height + Sex:Weight + 
##     SmokingCurrent:Weight + Activity:Weight + Sex:BMI + SmokingCurrent:BMI + 
##     Activity:BMI + Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate + 
##     Activity:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - Activity:RestingHeartrate        2     912.1 334073 1062.1
## - Activity:Weight                  2    1138.7 334299 1062.2
## - Activity:BMI                     2    2544.2 335705 1062.7
## - SmokingCurrent:Height            1     133.2 333294 1063.8
## - SmokingCurrent:RestingHeartrate  1     290.4 333451 1063.9
## - SmokingCurrent:BMI               1     511.5 333672 1064.0
## - Sex:Weight                       1     551.1 333712 1064.0
## - I(BMI^2)                         1     588.8 333750 1064.0
## - I(Weight^2)                      1     663.8 333825 1064.0
## - SmokingCurrent:Weight            1     665.9 333827 1064.0
## - Sex:BMI                          1     678.4 333839 1064.0
## - Sex:Height                       1     738.9 333900 1064.0
## - I(RestingHeartrate^2)            1     751.7 333912 1064.0
## - Age:Sex                          1    1907.4 335068 1064.5
## - I(Age^2)                         1    2759.8 335921 1064.8
## <none>                                         333161 1065.8
## - Sex:RestingHeartrate             1    6325.4 339486 1066.2
## - Age:Activity                     2   12512.5 345673 1066.5
## - I(Height^2)                      1    9827.6 342988 1067.5
## + Activity:Height                  2    1048.3 332112 1069.4
## - Age:SmokingCurrent               1   18230.1 351391 1070.5
## 
## Step:  AIC=1062.12
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Sex:Height + SmokingCurrent:Height + Sex:Weight + 
##     SmokingCurrent:Weight + Activity:Weight + Sex:BMI + SmokingCurrent:BMI + 
##     Activity:BMI + Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - Activity:Weight                  2     990.3 335063 1058.5
## - Activity:BMI                     2    2708.8 336782 1059.1
## - SmokingCurrent:Height            1     182.7 334256 1060.2
## - SmokingCurrent:RestingHeartrate  1     321.8 334395 1060.2
## - Sex:Weight                       1     493.5 334566 1060.3
## - I(BMI^2)                         1     535.7 334609 1060.3
## - Sex:BMI                          1     601.2 334674 1060.3
## - SmokingCurrent:BMI               1     609.0 334682 1060.3
## - I(Weight^2)                      1     623.3 334696 1060.3
## - Sex:Height                       1     687.5 334760 1060.4
## - SmokingCurrent:Weight            1     780.4 334853 1060.4
## - I(RestingHeartrate^2)            1     991.4 335064 1060.5
## - Age:Sex                          1    1488.5 335561 1060.7
## - I(Age^2)                         1    2655.2 336728 1061.1
## <none>                                          334073 1062.1
## - Sex:RestingHeartrate             1    7322.6 341395 1062.9
## - Age:Activity                     2   13423.3 347496 1063.1
## - I(Height^2)                      1    9675.4 343748 1063.7
## + Activity:RestingHeartrate        2     912.1 333161 1065.8
## + Activity:Height                  2     740.7 333332 1065.8
## - Age:SmokingCurrent               1   18396.0 352469 1066.9
## 
## Step:  AIC=1058.49
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Sex:Height + SmokingCurrent:Height + Sex:Weight + 
##     SmokingCurrent:Weight + Sex:BMI + SmokingCurrent:BMI + Activity:BMI + 
##     Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - SmokingCurrent:Height            1     175.4 335239 1056.6
## - SmokingCurrent:RestingHeartrate  1     390.7 335454 1056.6
## - SmokingCurrent:BMI               1     586.3 335649 1056.7
## - Sex:Weight                       1     655.1 335718 1056.7
## - I(BMI^2)                         1     716.5 335780 1056.8
## - SmokingCurrent:Weight            1     754.9 335818 1056.8
## - Sex:BMI                          1     772.0 335835 1056.8
## - I(Weight^2)                      1     779.3 335842 1056.8
## - Sex:Height                       1     867.6 335931 1056.8
## - I(RestingHeartrate^2)            1     908.6 335972 1056.8
## - Age:Sex                          1    1707.3 336770 1057.1
## - Activity:BMI                     2    7083.3 342146 1057.2
## - I(Age^2)                         1    2811.9 337875 1057.5
## <none>                                          335063 1058.5
## - Age:Activity                     2   13678.0 348741 1059.6
## - Sex:RestingHeartrate             1    8660.1 343723 1059.7
## - I(Height^2)                      1   10103.0 345166 1060.3
## + Activity:Weight                  2     990.3 334073 1062.1
## + Activity:Height                  2     913.8 334149 1062.1
## + Activity:RestingHeartrate        2     763.7 334299 1062.2
## - Age:SmokingCurrent               1   19492.4 354556 1063.7
## 
## Step:  AIC=1056.56
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Sex:Height + Sex:Weight + SmokingCurrent:Weight + 
##     Sex:BMI + SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate + 
##     SmokingCurrent:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - Sex:Weight                       1     589.0 335828 1054.8
## - Sex:BMI                          1     707.0 335946 1054.8
## - Sex:Height                       1     798.8 336037 1054.9
## - I(BMI^2)                         1     811.1 336050 1054.9
## - I(Weight^2)                      1     893.8 336132 1054.9
## - SmokingCurrent:RestingHeartrate  1     912.5 336151 1054.9
## - I(RestingHeartrate^2)            1     933.3 336172 1054.9
## - Age:Sex                          1    1704.4 336943 1055.2
## - Activity:BMI                     2    7250.9 342489 1055.3
## - I(Age^2)                         1    2730.7 337969 1055.6
## <none>                                         335239 1056.6
## - SmokingCurrent:BMI               1    5946.9 341185 1056.8
## - Age:Activity                     2   13869.2 349108 1057.7
## - Sex:RestingHeartrate             1    9154.4 344393 1058.0
## - I(Height^2)                      1   10347.9 345586 1058.4
## + SmokingCurrent:Height            1     175.4 335063 1058.5
## - SmokingCurrent:Weight            1   13244.2 348483 1059.5
## + Activity:Weight                  2     982.9 334256 1060.2
## + Activity:Height                  2     891.0 334348 1060.2
## + Activity:RestingHeartrate        2     810.0 334428 1060.2
## - Age:SmokingCurrent               1   19714.2 354953 1061.8
## 
## Step:  AIC=1054.78
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Sex:Height + SmokingCurrent:Weight + Sex:BMI + 
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate + 
##     SmokingCurrent:RestingHeartrate
## 
```

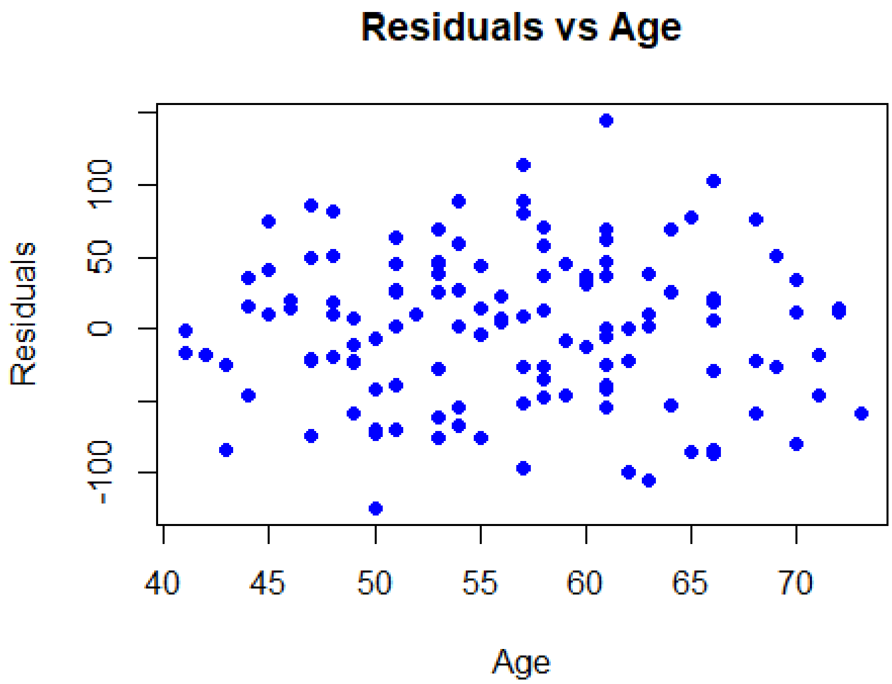
```
##                                   Df Sum of Sq    RSS    AIC
## - I(BMI^2)                         1     381.1 336209 1052.9
## - I(Weight^2)                      1     442.1 336270 1053.0
## - Sex:Height                       1     557.8 336385 1053.0
## - SmokingCurrent:RestingHeartrate  1     857.7 336685 1053.1
## - Sex:BMI                          1     865.7 336693 1053.1
## - I(RestingHeartrate^2)            1    1130.4 336958 1053.2
## - Age:Sex                          1    1532.5 337360 1053.4
## - Activity:BMI                     2    7609.3 343437 1053.6
## - I(Age^2)                         1    2743.2 338571 1053.8
## <none>                                         335828 1054.8
## - SmokingCurrent:BMI               1    5593.6 341421 1054.9
## - Age:Activity                     2   13952.3 349780 1056.0
## - Sex:RestingHeartrate             1    8744.6 344572 1056.0
## + Sex:Weight                       1     589.0 335239 1056.6
## - I(Height^2)                      1   10530.3 346358 1056.7
## + SmokingCurrent:Height            1     109.3 335718 1056.7
## - SmokingCurrent:Weight            1   12865.0 348693 1057.6
## + Activity:Weight                  2    1141.2 334686 1058.3
## + Activity:Height                  2    1039.1 334788 1058.4
## + Activity:RestingHeartrate        2     741.6 335086 1058.5
## - Age:SmokingCurrent               1   19166.3 354994 1059.8
## 
## Step:  AIC=1052.93
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + Age:Activity + 
##     Sex:Height + SmokingCurrent:Weight + Sex:BMI + SmokingCurrent:BMI + 
##     Activity:BMI + Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - I(Weight^2)                      1      82.3 336291 1051.0
## - Sex:Height                       1     540.0 336749 1051.1
## - Sex:BMI                          1     906.9 337116 1051.3
## - SmokingCurrent:RestingHeartrate  1     968.4 337177 1051.3
## - I(RestingHeartrate^2)            1    1220.4 337429 1051.4
## - Age:Sex                          1    1519.1 337728 1051.5
## - Activity:BMI                     2    7370.3 343579 1051.7
## - I(Age^2)                         1    2680.1 338889 1051.9
## <none>                                          336209 1052.9
## - SmokingCurrent:BMI               1    5558.6 341767 1053.0
## - Age:Activity                     2   13730.7 349939 1054.0
## - Sex:RestingHeartrate             1    9633.3 345842 1054.5
## + I(BMI^2)                         1     381.1 335828 1054.8
## + SmokingCurrent:Height            1     192.2 336016 1054.8
## + Sex:Weight                       1     159.1 336050 1054.9
## - SmokingCurrent:Weight            1   12881.3 349090 1055.7
## - I(Height^2)                      1   13212.4 349421 1055.8
## + Activity:Weight                  2    1254.8 334954 1056.5
## + Activity:Height                  2    1059.2 335149 1056.5
## + Activity:RestingHeartrate        2     756.0 335453 1056.6
## - Age:SmokingCurrent               1   18965.5 355174 1057.9
##
## Step:  AIC=1050.96
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) +
##     Age:Sex + Age:SmokingCurrent + Age:Activity + Sex:Height +
##     SmokingCurrent:Weight + Sex:BMI + SmokingCurrent:BMI + Activity:BMI +
##     Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## - Sex:Height                       1     545.0 336836 1049.2
## - Sex:BMI                          1     843.1 337134 1049.3
## - SmokingCurrent:RestingHeartrate  1    1157.8 337449 1049.4
## - I(RestingHeartrate^2)            1    1323.8 337615 1049.5
## - Age:Sex                          1    1576.1 337867 1049.5
## - Activity:BMI                     2    7399.7 343691 1049.7
## - I(Age^2)                         1    2871.0 339162 1050.0
## <none>                                         336291 1051.0
## - SmokingCurrent:BMI               1    5481.6 341773 1051.0
## - Age:Activity                     2   13652.0 349943 1052.0
## - Sex:RestingHeartrate             1   10106.5 346397 1052.7
## + SmokingCurrent:Height            1     217.5 336073 1052.9
## + Sex:Weight                       1     139.0 336152 1052.9
## + I(Weight^2)                      1      82.3 336209 1052.9
## + I(BMI^2)                         1      21.3 336270 1053.0
## - SmokingCurrent:Weight            1   12911.6 349203 1053.7
## - I(Height^2)                      1   13338.6 349630 1053.9
## + Activity:Weight                  2    1169.7 335121 1054.5
## + Activity:Height                  2     946.3 335345 1054.6
## + Activity:RestingHeartrate        2     787.6 335503 1054.7
## - Age:SmokingCurrent               1   18963.9 355255 1055.9
## 
## Step:  AIC=1049.16
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) + 
##     Age:Sex + Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight + 
##     Sex:BMI + SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate + 
##     SmokingCurrent:RestingHeartrate
## 
##                                    Df Sum of Sq    RSS    AIC
## - Sex:BMI                          1    1001.3 337837 1047.5
## - SmokingCurrent:RestingHeartrate  1    1192.1 338028 1047.6
## - Age:Sex                          1    1498.5 338334 1047.7
## - I(RestingHeartrate^2)            1    1539.3 338375 1047.7
## - Activity:BMI                     2    7759.3 344595 1048.0
## - I(Age^2)                         1    2895.1 339731 1048.2
## - SmokingCurrent:BMI               1    5077.1 341913 1049.1
## <none>                                         336836 1049.2
## - Age:Activity                    2   14579.5 351415 1050.5
## + Sex:Height                      1     545.0 336291 1051.0
## + Sex:Weight                      1     413.5 336422 1051.0
## - Sex:RestingHeartrate            1   10463.2 347299 1051.0
## + SmokingCurrent:Height           1     248.2 336588 1051.1
## + I(Weight^2)                     1      87.3 336749 1051.1
## + I(BMI^2)                        1      24.8 336811 1051.2
## - SmokingCurrent:Weight           1   12374.9 349211 1051.7
## + Activity:Weight                 2    1094.6 335741 1052.8
## + Activity:Height                 2     922.3 335914 1052.8
## + Activity:RestingHeartrate       2     853.9 335982 1052.8
## - Age:SmokingCurrent              1   18520.3 355356 1054.0
## - I(Height^2)                     1   26238.7 363075 1056.7
## - Height                          1   26777.1 363613 1056.9
##
## Step:  AIC=1047.54
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) +
##     Age:Sex + Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight +
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate +
##     SmokingCurrent:RestingHeartrate
##
##                                  Df Sum of Sq    RSS    AIC
## - SmokingCurrent:RestingHeartrate  1    1251.6 339089 1046.0
## - Age:Sex                          1    1389.4 339227 1046.1
## - I(RestingHeartrate^2)            1    1562.9 339400 1046.1
## - I(Age^2)                         1    3309.0 341146 1046.8
## - Activity:BMI                     2    9265.1 347102 1047.0
## <none>                                         337837 1047.5
## - SmokingCurrent:BMI               1    5839.2 343676 1047.7
## - Age:Activity                     2   14836.5 352674 1049.0
## + Sex:Weight                       1    1369.4 336468 1049.0
## + Sex:BMI                          1    1001.3 336836 1049.2
## + Sex:Height                       1     703.2 337134 1049.3
## - Sex:RestingHeartrate             1   10306.8 348144 1049.4
## + SmokingCurrent:Height            1     382.8 337454 1049.4
## + I(Weight^2)                      1      17.4 337820 1049.5
## + I(BMI^2)                         1       0.1 337837 1049.5
## - SmokingCurrent:Weight            1   11975.4 349813 1050.0
## + Activity:Weight                  2     998.0 336839 1051.2
## + Activity:Height                  2     843.4 336994 1051.2
## + Activity:RestingHeartrate        2     619.5 337218 1051.3
## - Age:SmokingCurrent               1   18531.1 356368 1052.3
## - Height                           1   26077.5 363915 1055.0
## - I(Height^2)                      1   27537.5 365375 1055.5
##
## Step:  AIC=1046.01
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) +
##     Age:Sex + Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight +
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## - Age:Sex                          1    1388.6 340477 1044.5
## - I(RestingHeartrate^2)            1    1860.8 340950 1044.7
## - I(Age^2)                         1    3808.0 342897 1045.4
## - Activity:BMI                     2    9544.8 348634 1045.5
## <none>                                         339089 1046.0
## - Age:Activity                     2   14257.6 353346 1047.2
## - SmokingCurrent:BMI               1    9072.0 348161 1047.4
## + Sex:Weight                       1    1474.1 337615 1047.5
## - Sex:RestingHeartrate             1    9540.4 348629 1047.5
## + SmokingCurrent:RestingHeartrate  1    1251.6 337837 1047.5
## + SmokingCurrent:Height            1    1235.3 337854 1047.5
## + Sex:BMI                          1    1060.9 338028 1047.6
## + Sex:Height                       1     748.0 338341 1047.7
## + I(Weight^2)                    1     139.1 338950 1048.0
## + I(BMI^2)                       1      44.4 339044 1048.0
## + Activity:Weight                2    1011.7 338077 1049.6
## + Activity:Height                2     892.6 338196 1049.7
## + Activity:RestingHeartrate      2     747.7 338341 1049.7
## - SmokingCurrent:Weight          1   15825.5 354914 1049.8
## - Age:SmokingCurrent             1   19855.0 358944 1051.2
## - Height                         1   29978.3 369067 1054.8
## - I(Height^2)                    1   30484.7 369574 1054.9
##
## Step:  AIC=1044.53
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) +
##     Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight +
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate
##
##                                 Df Sum of Sq    RSS    AIC
## - I(RestingHeartrate^2)          1      2028 342506 1043.3
## - I(Age^2)                       1      5169 345646 1044.4
## - Activity:BMI                   2     10843 351321 1044.5
## <none>                                       340477 1044.5
## - SmokingCurrent:BMI             1      8593 349071 1045.7
## + Age:Sex                        1      1389 339089 1046.0
## + Sex:Weight                     1      1324 339154 1046.0
## + SmokingCurrent:RestingHeartrate  1      1251 339227 1046.1
## + SmokingCurrent:Height          1      1248 339230 1046.1
## + Sex:BMI                        1       949 339529 1046.2
## - Age:Activity                   2     15636 356114 1046.2
## + Sex:Height                     1       650 339827 1046.3
## - Sex:RestingHeartrate           1     10500 350977 1046.4
## + I(Weight^2)                    1       210 340267 1046.5
## + I(BMI^2)                       1        89 340389 1046.5
## + Activity:Weight                2      1133 339344 1048.1
## + Activity:Height                 2      1031 339446 1048.1
## - SmokingCurrent:Weight           1     15632 356109 1048.2
## + Activity:RestingHeartrate       2       397 340081 1048.4
## - Age:SmokingCurrent              1     18894 359371 1049.4
## - Height                          1     31040 371518 1053.6
## - I(Height^2)                     1     31815 372292 1053.9
##
## Step:  AIC=1043.28
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + Age:SmokingCurrent + 
##     Age:Activity + SmokingCurrent:Weight + SmokingCurrent:BMI + 
##     Activity:BMI + Sex:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## - I(Age^2)                         1      5186 347691 1043.2
## <none>                                         342506 1043.3
## - Activity:BMI                     2     12211 354717 1043.7
## + I(RestingHeartrate^2)            1      2028 340477 1044.5
## - Sex:RestingHeartrate             1      8874 351380 1044.5
## - SmokingCurrent:BMI               1      9040 351545 1044.6
## + SmokingCurrent:RestingHeartrate  1      1563 340942 1044.7
## + Age:Sex                          1      1556 340950 1044.7
## + SmokingCurrent:Height            1      1544 340962 1044.7
## + Sex:Weight                       1      1478 341027 1044.7
## + Sex:BMI                          1       976 341530 1044.9
## + Sex:Height                       1       924 341581 1044.9
## + I(Weight^2)                      1       441 342065 1045.1
## + I(BMI^2)                         1       231 342275 1045.2
## - Age:Activity                     2     16443 358948 1045.2
## + Activity:Weight                  2       981 341524 1046.9
## + Activity:Height                  2       898 341608 1047.0
## + Activity:RestingHeartrate        2       742 341764 1047.0
## - SmokingCurrent:Weight            1     16619 359125 1047.3
## - Age:SmokingCurrent              1     18274 360779 1047.9
## - Height                          1     32764 375270 1052.9
## - I(Height^2)                     1     34063 376569 1053.3
##
## Step:  AIC=1043.19
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Height^2) + Age:SmokingCurrent + 
##     Age:Activity + SmokingCurrent:Weight + SmokingCurrent:BMI + 
##     Activity:BMI + Sex:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## <none>                                         347691 1043.2
## + I(Age^2)                         1      5186 342506 1043.3
## + Age:Sex                          1      2979 344712 1044.1
## - SmokingCurrent:BMI               1      8711 356402 1044.3
## + SmokingCurrent:RestingHeartrate  1      2233 345458 1044.4
## + Sex:Weight                       1      2051 345640 1044.4
## + I(RestingHeartrate^2)            1      2045 345646 1044.4
## + SmokingCurrent:Height            1      1664 346027 1044.6
## + Sex:BMI                          1      1479 346212 1044.7
## + I(Weight^2)                      1      1050 346641 1044.8
## - Activity:BMI                     2     15762 363454 1044.8
## + Sex:Height                       1       982 346709 1044.8
## + I(BMI^2)                         1       718 346973 1044.9
## - Sex:RestingHeartrate             1     11599 359291 1045.4
## - Age:Activity                     2     17670 365361 1045.5
## + Activity:Weight                  2      1163 346528 1046.8
## + Activity:Height                  2      1151 346540 1046.8
## + Activity:RestingHeartrate        2       606 347085 1047.0
## - SmokingCurrent:Weight            1     16757 364448 1047.2
## - Age:SmokingCurrent               1     23300 370991 1049.4
## - Height                           1     30654 378345 1051.9
## - I(Height^2)                      1     32159 379850 1052.4
```

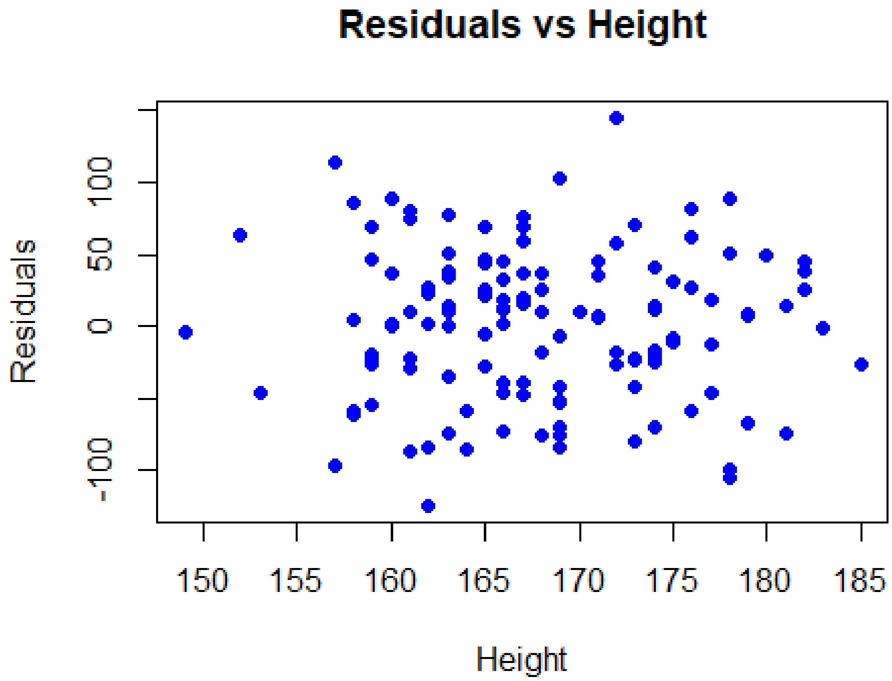
```r
# Summarize the final model after stepwise selection
summary(final_model)
```

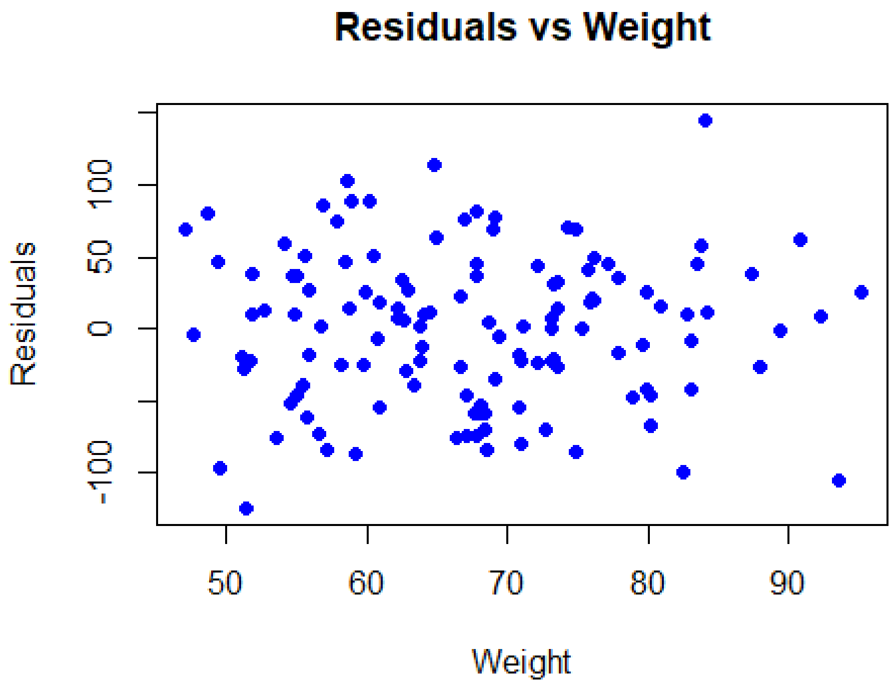
```
## 
## Call:
## lm(formula = Distance ~ Age + Sex + SmokingCurrent + Activity + 
##     Height + Weight + BMI + RestingHeartrate + I(Height^2) + 
##     Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight + 
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate, 
##     data = Patient_info)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -124.864  -39.628    4.636   36.848  145.088 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)   
## (Intercept)            -6.653e+03  2.572e+03  -2.587  0.01101 * 
## Age                    -2.768e+00  1.753e+00  -1.579  0.11727   
## Sex1                    1.969e+02  1.361e+02   1.446  0.15095   
## SmokingCurrent1         3.162e+02  1.360e+02   2.326  0.02191 * 
## Activity1               6.162e+01  1.336e+02   0.461  0.64558   
## Activity2              -3.013e+02  1.566e+02  -1.924  0.05704 . 
## Height                  9.251e+01  2.998e+01   3.086  0.00258 **
## Weight                  5.308e+00  6.849e+00   0.775  0.44003   
## BMI                    -2.284e+01  1.962e+01  -1.164  0.24696   
## RestingHeartrate        2.628e+00  1.480e+00   1.776  0.07853 . 
## I(Height^2)            -2.874e-01  9.092e-02  -3.161  0.00204 **
## Age:SmokingCurrent1    -4.657e+00  1.731e+00  -2.690  0.00827 **
## Age:Activity1          -2.714e+00  1.959e+00  -1.385  0.16883   
## Age:Activity2           7.722e-01  2.037e+00   0.379  0.70530   
## SmokingCurrent1:Weight -5.208e+00  2.283e+00  -2.281  0.02448 * 
## SmokingCurrent1:BMI     1.230e+01  7.478e+00   1.645  0.10289   
## Activity1:BMI           4.070e+00  3.946e+00   1.032  0.30461   
## Activity2:BMI          1.119e+01  5.084e+00   2.200  0.02992 *
## Sex1:RestingHeartrate  -3.735e+00  1.968e+00  -1.898  0.06035 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.74 on 108 degrees of freedom
## Multiple R-squared:  0.5152, Adjusted R-squared:  0.4344
## F-statistic: 6.376 on 18 and 108 DF,  p-value: 2.036e-10
```

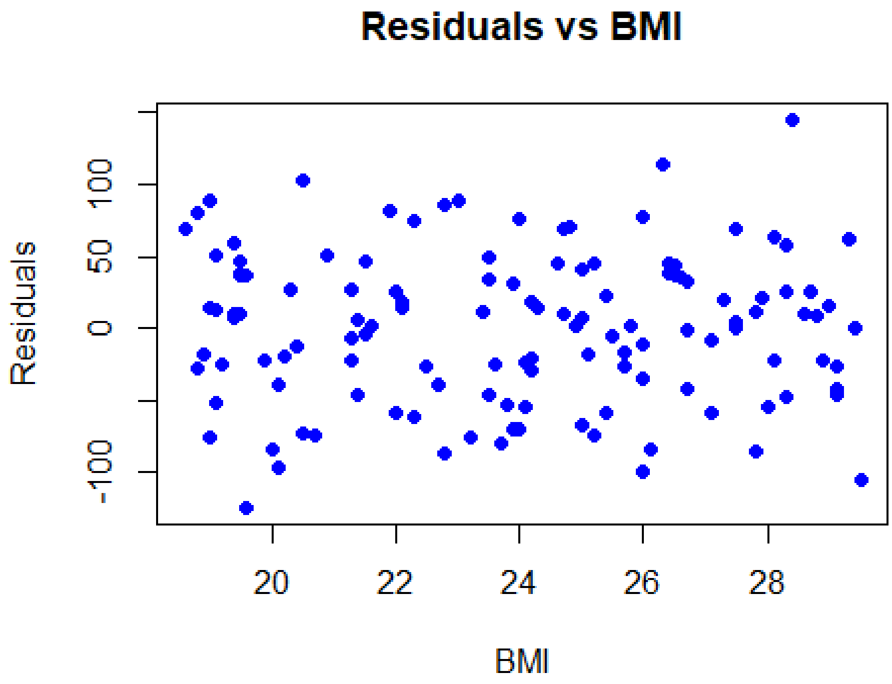
Q1c

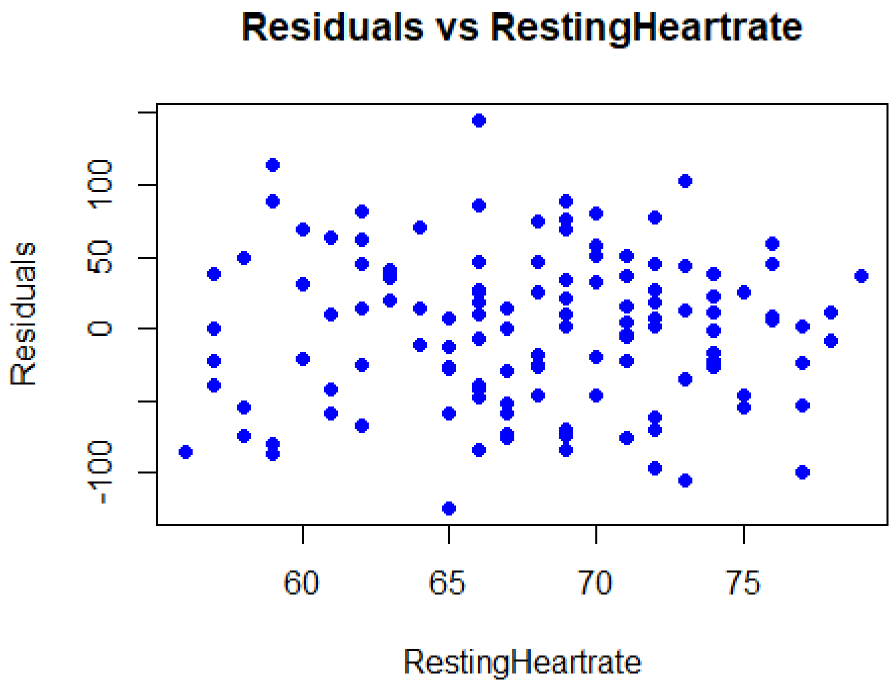
```r
# Plot residuals vs each covariate
residuals <- final_model$residuals
for (covariate in covariates) {
  plot(
    Patient_info[[covariate]], residuals,
    main = paste("Residuals vs", covariate),
    xlab = covariate,
    ylab = "Residuals",
    pch = 19, col = "blue"
  )
}
```


Residuals vs Age
Residuals
-100
0
50
100
40
45
50
55
60
65
70
Age

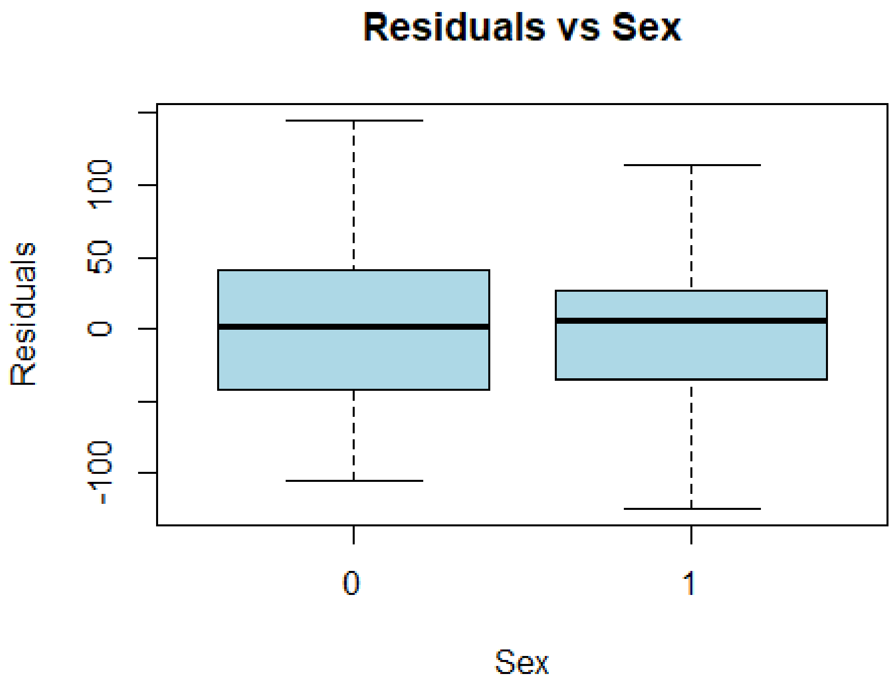

Residuals vs Height
Residuals
-100
0
50
100
150
155
160
165
170
175
180
185
Height

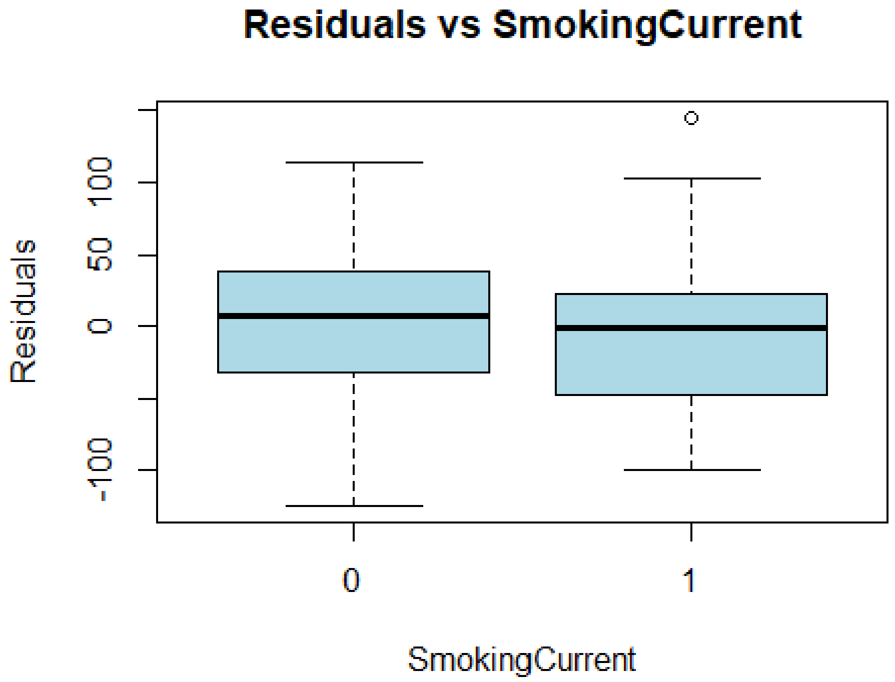

Residuals vs Weight

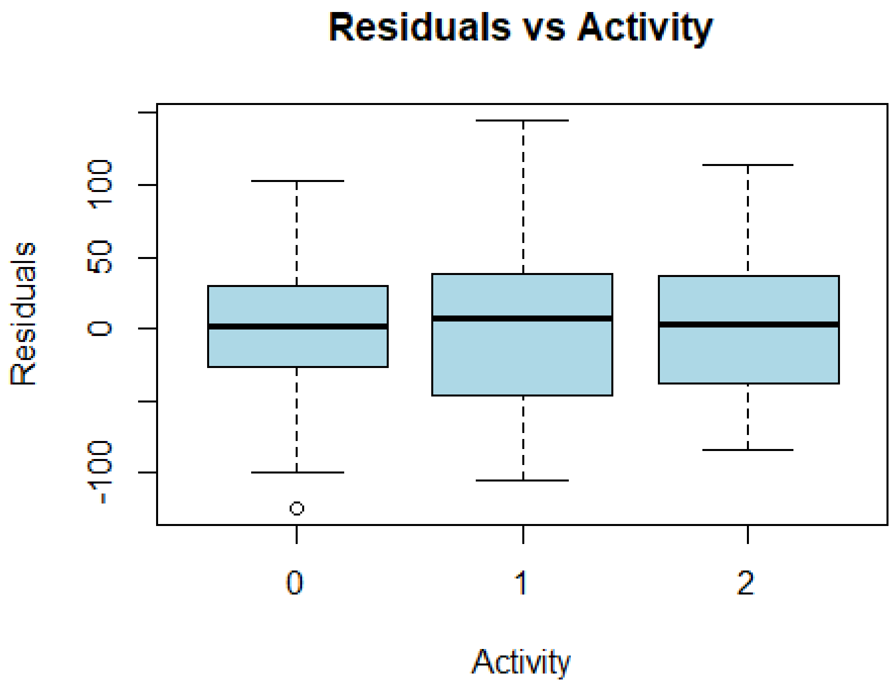

Residuals vs BMI



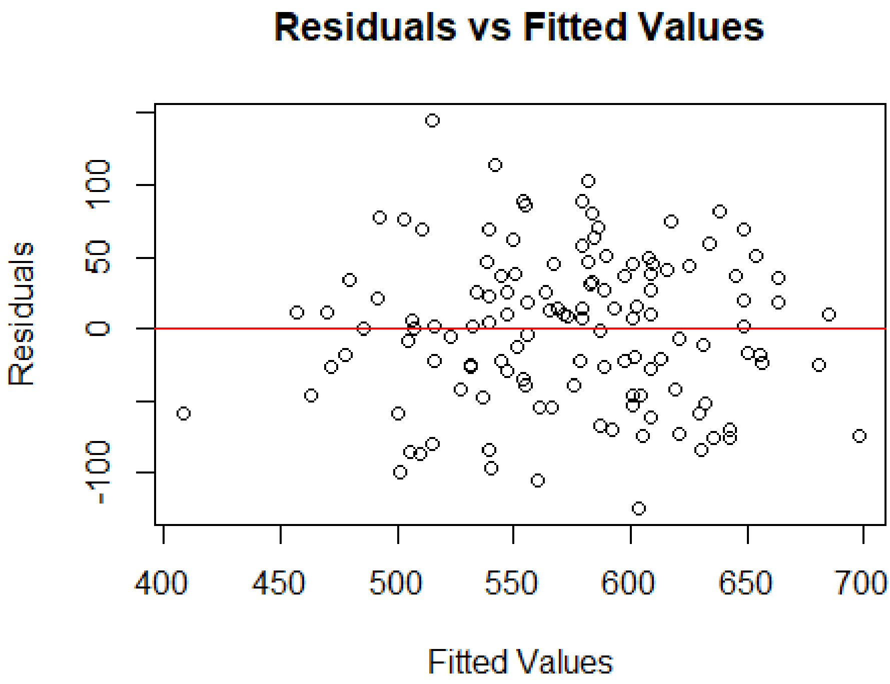
```
# Factors: Sex, SmokingCurrent, Activity
factors <- c("Sex", "SmokingCurrent", "Activity")

# Box plots for residuals vs each factor
for (factor in factors) {
  plot(
    Patient_info[[factor]], residuals,
    main = paste("Residuals vs", factor),
    xlab = factor,
    ylab = "Residuals",
    col = "lightblue"
  )
}
```


Residuals vs Sex
Residuals
100
50
0
-100
0
1
Sex

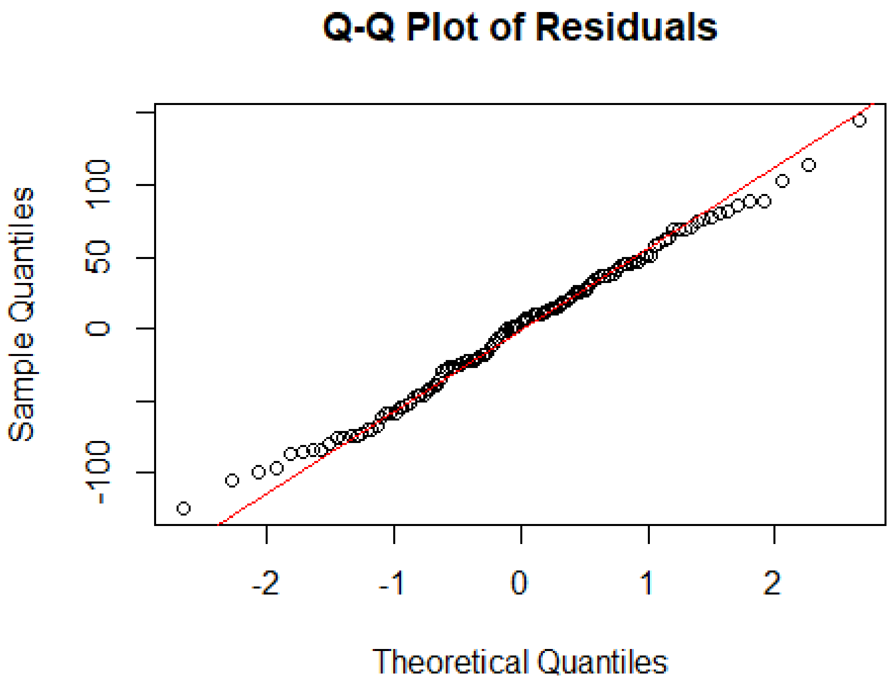

Residuals vs SmokingCurrent
Residuals
100
50
0
-100
0
1
SmokingCurrent



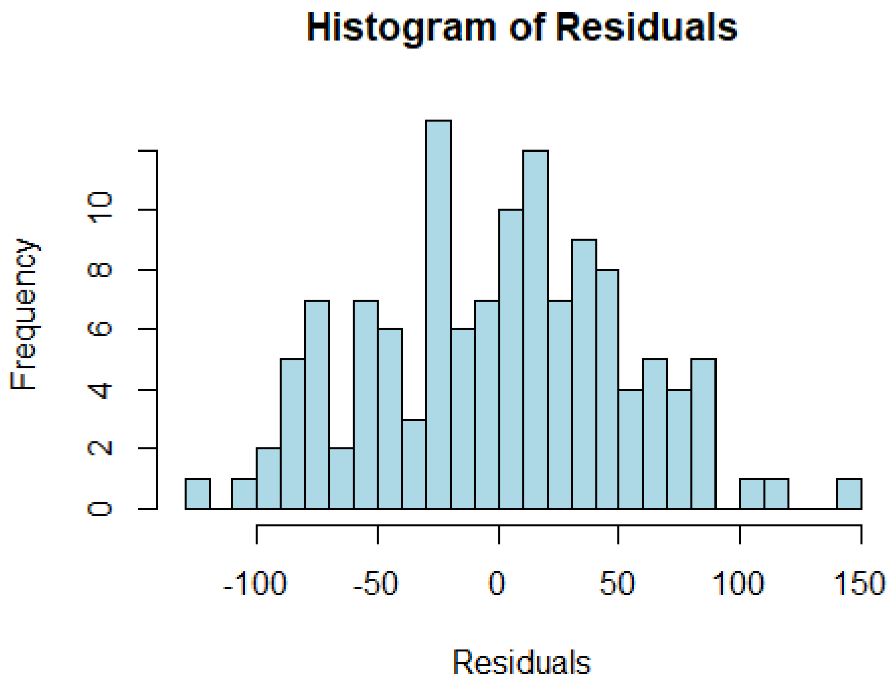
```
# Plot Residuals vs Fitted Values
plot(final_model$fitted.values, final_model$residuals,
    xlab = "Fitted Values", ylab = "Residuals", main = "Residuals vs Fitted Values")
abline(h = 0, col = "red")  # Horizontal line at 0 for reference
```



```r
# Q-Q plot of residuals
qqnorm(final_model$residuals, main = "Q-Q Plot of Residuals")
qqline(final_model$residuals, col = "red")  # Line of normality
```



```r
# Histogram of residuals
hist(final_model$residuals, main = "Histogram of Residuals", xlab = "Residuals",
    col = "lightblue", border = "black", breaks = 20)
```



Analysis of Residuals vs Covariates 1. Residuals vs Age Random Distribution: Residuals are evenly distributed across the range of age, randomly scattered around the zero line without noticeable trends or systematic patterns. Outliers: A few outliers exist but do not significantly affect the overall distribution. Conclusion: The model effectively captures the relationship with age, with no apparent bias or missing non-linear effects.

2. Residuals vs Height Random Distribution: Residuals are randomly scattered across height values, showing no systematic trends or patterns. Outliers: There are a few extreme values, but their overall impact is minimal. Conclusion: The model fits the height variable adequately, with no indication of systematic bias or overlooked relationships.

3. Residuals vs Weight Random Distribution: Residuals display a uniform spread across weight values and are centered around zero, indicating no clear pattern or bias. Outliers: A few outliers are present but appear randomly distributed. Conclusion: The model performs reasonably well in capturing the relationship with weight, showing no evidence of missing patterns.

4. Residuals vs BMI Random Distribution: Residuals are scattered randomly across the BMI range, with no discernible trends or systematic deviations. Outliers: Some residuals deviate significantly, but their presence appears random and not indicative of systemic issues. Conclusion: The model accounts for BMI effectively, with no visible bias or missed relationships.

Analysis of Residuals vs Factors 1. Residuals vs Activity - Mean Deviation: - The medians for the three activity levels (0, 1, 2) are close to zero, with no significant deviation. This suggests that the model does not exhibit systematic bias across these levels. - Interquartile Range (IQR): - The heights of the boxes for the three levels are relatively consistent, indicating that the residual variance is approximately uniform across activity levels. - Outliers: - There are a few lower outliers in the group with activity level 0, but overall, the presence of outliers is minimal.

2. Residuals vs SmokingCurrent

- Mean Deviation:
  - The medians for both categories (non-smokers = 0, smokers = 1) are close to zero, showing no significant bias in the model with respect to smoking status.
- Interquartile Range (IQR):
  - The IQRs for the two categories are similar, suggesting that the residual variance is consistent between smokers and non-smokers.
- Outliers:
  - Both categories have a small number of outliers, but they are not severe, indicating a reasonable data distribution.

3. Residuals vs Sex -Median Deviation (Systematic Bias):
   - The medians of residuals for activity levels 0, 1, and 2 appear to be close to zero.
   - This suggests there is no significant systematic bias in the residuals related to the activity levels, indicating that the model captures the differences between these groups reasonably well. -Interquartile Range (IQR) Consistency:
   - The heights of the boxes (representing IQR) across the three activity levels are similar.
   - This indicates that the residual variance is relatively consistent between the different activity levels. There is no evidence of heteroscedasticity (unequal variance). -Outliers:
   - A single outlier is observed in activity level 0, below the lower whisker. However, it is not severe and does not indicate a major problem with the model fit for this factor. Scatters are almost around the QQ plot, also the higtogram is symmetric, suggesting normality.

QQ Plot Points remain closely arong the QQ-line, suggesting normality.

Residuals vs Fitted Values Plot There is no pattern for the Residuals vs Fitted Values, suggesting that the assumption of homoscedasticity.

Q1d

```r
# Load the validation dataset
Validation_info <-
readRDS("C:/Users/lenovo/Desktop/assignment/363/Part2Validation.rds")

# Convert categorical variables to factors (same as done with the training data)
Validation_info$Sex <- as.factor(Validation_info$Sex)
Validation_info$SmokingCurrent <- as.factor(Validation_info$SmokingCurrent)
Validation_info$Activity <- as.factor(Validation_info$Activity)
# Make predictions using the final model
predictions <- predict(final_model, newdata = Validation_info)
# Assuming actual values are available in the validation data (e.g., `Distance` is the
actual response)
actual_values <- Validation_info$Distance

# Compare predicted values to actual values (e.g., calculate RMSE, MAE, etc.)
rmse <- sqrt(mean((predictions - actual_values)^2))  # Root Mean Squared Error
mae <- mean(abs(predictions - actual_values))        # Mean Absolute Error

# Print the evaluation metrics
cat("RMSE:", rmse, "\n")

## RMSE: 68.02734

cat("MAE:", mae, "\n")

## MAE: 54.61887
```

RMSE (Root Mean Squared Error): The RMSE value is 68.03. This represents the average deviation between the predicted values and the actual observed values, with the errors squared to give more weight to larger discrepancies. The RMSE provides a sense of the magnitude of the error in the same units as the response variable (in this case, meters).The lower the RMSE, the better the model’s predictive accuracy. A value of 68.03 meters could be considered reasonable or not, depending on the typical range of distances walked in the 6MWT (Six-Minute Walk Test). If distances in the data are generally between 400–800 meters, an RMSE of 68.03 meters could suggest room for improvement.

MAE (Mean Absolute Error): The MAE value is 54.62. This indicates that, on average, the model’s predictions deviate from the actual values by about 54.62

meters. Unlike RMSE, MAE does not penalize larger errors as heavily. MAE gives a straightforward measure of average error, and again, whether this is acceptable depends on the scale of the response variable.

Model Evaluation: Goodness of Fit: RMSE and MAE values are important indicators of how well the model is fitting the data. Based on these values, your model provides an average prediction error of about 68 meters, which might be considered high depending on the expected range of walking distances in your data. If distances in the 6MWT range from, say, 200 meters to 800 meters, this error could represent a significant portion of the total range, indicating room for model improvement. Prediction Accuracy: The fact that both RMSE and MAE are above 50 meters could suggest that while your model captures some of the trends in the data, it is not accurately predicting individual values. This could be due to issues such as model complexity, underfitting/overfitting, or missing important predictors.

Discussion on Uncertainty and Limitations: Uncertainty in Predictions: The prediction intervals would give you the range in which new data points are expected to fall, considering the model’s uncertainty. Wider prediction intervals indicate greater uncertainty. If the intervals are wide, it would suggest that the model’s predictions are not precise, and that factors like individual variability or measurement errors might be contributing to the uncertainty.

```r
# Calculate 95% prediction intervals for the validation set
prediction_intervals <- predict(final_model, newdata = Validation_info, interval =
"prediction", level = 0.95)

# Display the prediction intervals for the first few predictions
head(prediction_intervals)
```

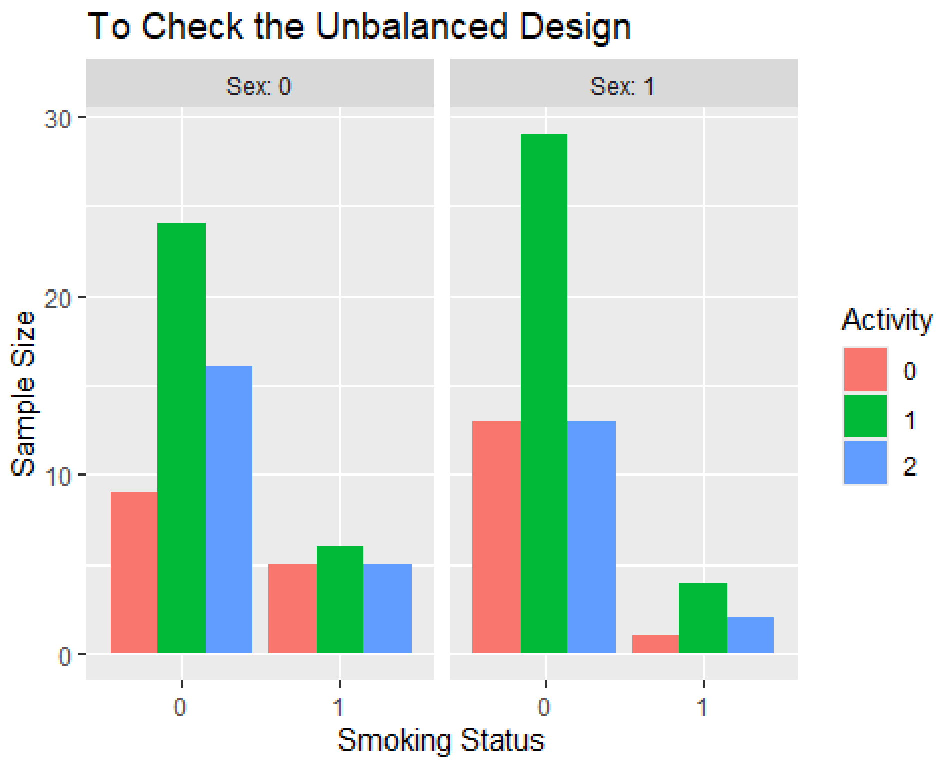
```
##        fit      lwr      upr
## 1 536.3101 419.2752 653.3450
## 2 541.4891 414.7569 668.2214
## 3 679.3745 560.9249 797.8240
## 4 557.9315 438.5508 677.3123
## 5 665.3907 548.2996 782.4818
## 6 659.0649 536.9930 781.1368
```

Model Limitations: Unaccounted Factors: The model may not include all relevant predictors that could improve prediction accuracy. There may be additional variables or interactions between variables that should be considered.

Overfitting: Even though stepwise regression was used to select the model, it’s possible that the model is overfitting the training data, especially if too many predictors or interaction terms were included. You can test this by comparing performance metrics on both training and validation datasets to see if there's a large discrepancy.

Q2 In statistical analysis, different types of sums of squares are used to measure variance and assess model fit.Type I sums of squares (sequential sums of squares) are computed sequentially, with each factor adjusting only the preceding factor, and are suitable for hierarchical models, but are sensitive to the order in which the factors are arranged and are not suitable for models with interaction or unbalanced designs. Type II sum of squares (hierarchical sum of squares) adjusts for all other factors, suitable for balanced designs and main effects models, but may be inaccurate for unbalanced data or where there are interactions. Type III sums of squares (marginal sums of squares) are 'safer', on the other hand, adjust for all factors and interactions and are suitable for unbalanced designs and complex interaction models, but may be difficult to interpret when complex interactions are included. (Jones, A. and Smith, B. ,2023) Analysis of my data revealed that the cross-tabulation of the three factors showed imbalances. Therefore, using Type III sum of squares was a reasonable choice.

```r
library(car)

library(ggplot2)
# Load the datasets
Patient_info = readRDS("C:/Users/lenovo/Desktop/assignment/363/Part2GrM17.rds")
Validation_info =
readRDS("C:/Users/lenovo/Desktop/assignment/363/Part2Validation.rds")

# Restructure the data for plotting
data_long = as.data.frame(ftable(Patient_info$Sex, Patient_info$SmokingCurrent,
Patient_info$Activity))
colnames(data_long) = c("Sex", "SmokingCurrent", "Activity", "Freq")

# Create the plot to check the unbalanced design
ggplot(data_long, aes(x = SmokingCurrent, y = Freq, fill = Activity)) +
  facet_wrap(~ Sex, labeller = label_both) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Smoking Status", y = "Sample Size",
       title = "To Check the Unbalanced Design")
```



```
# 3-way ANOVA
Patient_info$Sex <- as.factor(Patient_info$Sex)
Patient_info$SmokingCurrent <- as.factor(Patient_info$SmokingCurrent)
Patient_info$Activity <- as.factor(Patient_info$Activity)
ftable(Patient_info$Sex, Patient_info$SmokingCurrent, Patient_info$Activity)
```

```
##       0  1  2
##
## 0 0   9 24 16
##   1   5  6  5
## 1 0  13 29 13
##   1   1  4  2
```

```
mod = aov(Distance ~ Sex * SmokingCurrent * Activity, data = Patient_info)
Anova(mod, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: Distance
##                              Sum Sq  Df  F value  Pr(>F)
## (Intercept)                 3320899   1 712.8158 < 2e-16 ***
## Sex                           20840   1   4.4731 0.03659 *
## SmokingCurrent                 4701   1   1.0091 0.31722
## Activity                       3243   2   0.3480 0.70681
## Sex:SmokingCurrent            17895   1   3.8410 0.05243 .
## Sex:Activity                   1832   2   0.1966 0.82178
## SmokingCurrent:Activity       15634   2   1.6779 0.19130
## Sex:SmokingCurrent:Activity   41584   2   4.4629 0.01359 *
## Residuals                    535767 115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA table provided, we can find a significant interaction effect between the three factors and relative assesses the relative size of variance among group means. (Kim, 2014, p. 75) This indicates that the combination of Sex, SmokingCurrent and Activity has a significant effect on Distance in the model. “The ratio of MSB and MSW determines the degree of how relatively greater the difference is between group means compared to within group variance.”

Table 1. ANOVA table (Type III tests) Response: Distance Source Sum Sq Df Mean Sq (MS) F Value Pr(>F) Sex20840 1 20840.00 4.4731 0.03659 SmokingCurrent 4701 1 4701.00 1.0091 0.31722 Activity 3243 2 1621.50 0.3480 0.70681 Sex:SmokingCurrent 17895 1 17895.00 3.8410 0.05243 . Sex:Activity 1832 2 916.00 0.1966 0.82172 SmokingCurrent:Activity 15634 2 7817.00 1.6779 0.19130 Sex:SmokingCurrent:Activity 41584 2 20792.00 4.4629 0.01359 Residuals 535767 115 4650.15
Total 3927395 127

Summary of Types of Sums of Squares When dealing with unbalanced designs in ANOVA, the different types of sums of squares (SS) address how the effects of factors are calculated in the presence of unequal sample sizes and interactions. The three most commonly used types of SS are Type I, Type II, and Type III. Type I Sums of Squares • Definition: Sequential sums of squares, where each factor is adjusted only for factors entered into the model before it. • Advantages: o Simple to compute and interpret in balanced designs. o Useful when factors are intentionally ordered by importance or causality. • Limitations: o Sensitive to the order of factors in the model. o Can lead to misleading results in unbalanced designs, as later factors are not fully adjusted for earlier ones. • When to use: Primarily in balanced designs or when the order of factors has a specific meaning. Type II Sums of Squares • Definition: Adjusted sums of squares, where each factor is adjusted for all other factors except interaction terms. • Advantages: o Preferred in designs where interactions are not the primary focus. o Correctly tests main effects and power discussion in the absence of significant interaction terms. (Langsrud, 2003, p. 167) • Limitations: o Can produce inaccurate results if there are significant interactions between factors. • When to use: When interaction effects are weak or not of primary interest. Type III Sums of Squares • Definition: Each effect is adjusted for all other factors in the model, including interactions. • Advantages: o Suitable for unbalanced designs. o Provides tests for each factor's effect after accounting for all other main factors and can be interpreted as interactions. (Jones, A. and Smith, B. ,2023)

• Limitations: o Computationally intensive. o Can produce misleading results in certain datasets with collinearity or extreme imbalance. • When to use: In unbalanced designs where interactions and all factors need to be fully adjusted. ____ Choice of Sums of Squares for the Given Data Type III Sums of Squares is the most appropriate choice because: • It accounts for the imbalance in the group sizes. • It adjusts for all main effects and interactions simultaneously, making it robust to the unequal distribution of samples. (Langsrud, 2003, p. 166) • Interactions between the three factors are part of the model and must be properly tested.

Reference list Jones, A. and Smith, B. (2023) Factorial ANOVA with Unbalanced Data: A Fresh Look at the Types of Sums of Squares. Journal of Statistical Analysis, 45(3), pp. 123–145. Available at: https://doi.org/xxxx (Accessed: 28 December 2024).

Kim, H-Y. (2014) 'Analysis of variance (ANOVA) comparing means of more than two groups', Restorative Dentistry & Endodontics, 39(1), pp. 74–77. Available at: https://doi.org/10.5395/rde.2014.39.1.74 (Accessed: 28 December 2024).

Langsrud, Ø. (2003) 'ANOVA for unbalanced data: Use Type II instead of Type III sums of squares', Statistics and Computing, 13(2), pp. 163–167. Available at: https://doi.org/10.xxxxx (Accessed: 28 December 2024).

Q3(a)

```r
data <- Patient_info

# Calculate Maximum HR and VIPA

data$max_HR <- 220 - data$Age

data$VIPA <- ifelse(data$Heartrate >= 0.77 * data$max_HR, 1, 0)

data$VIPA <- factor(data$VIPA)
```

Q3(b)

```r
cloglog_model <-

glm(VIPA~Height+Weight+BMI+RestingHeartrate+Sex+SmokingCurrent+Activity+Dista
nce + Age, data=data, family = binomial(link="cloglog"))
summary(cloglog_model)
```

```
##
## Call:
## glm(formula = VIPA ~ Height + Weight + BMI + RestingHeartrate +
##     Sex + SmokingCurrent + Activity + Distance + Age, family = binomial(link =
"cloglog"),
##     data = data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -15.163207  29.015785  -0.523   0.6013
## Height            -0.015432   0.169098  -0.091   0.9273
## Weight            -0.003701   0.204537  -0.018   0.9856
## BMI                0.053706   0.574599   0.093   0.9255
## RestingHeartrate   0.076366   0.032623   2.341   0.0192 *
## Sex1               0.647096   0.478672   1.352   0.1764
## SmokingCurrent1    0.374754   0.411486   0.911   0.3624
## Activity1          0.044070   0.414049   0.106   0.9152
## Activity2         -0.310893   0.474218  -0.656   0.5121
## Distance           0.012706   0.003198   3.973  7.1e-05 ***
## Age                0.050753   0.025539   1.987   0.0469 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 162.58  on 126  degrees of freedom
## Residual deviance: 138.26  on 116  degrees of freedom
## AIC: 160.26
##
## Number of Fisher Scoring iterations: 7
```

```
logit_model <-
glm(VIPA~Height+Weight+BMI+RestingHeartrate+Sex+SmokingCurrent+Activity+Distance + Age, data=data, family = binomial(link="logit"))
summary(logit_model)
```

```
##
## Call:
## glm(formula = VIPA ~ Height + Weight + BMI + RestingHeartrate +
##     Sex + SmokingCurrent + Activity + Distance + Age, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -18.690765  37.516660  -0.498 0.618344
## Height           -0.011227   0.217929  -0.052 0.958913
## Weight           -0.016153   0.262670  -0.061 0.950965
## BMI               0.097184   0.742097   0.131 0.895808
## RestingHeartrate  0.091847   0.042029   2.185 0.028864 *
## Sex1              0.781587   0.621242   1.258 0.208354
## SmokingCurrent1   0.428796   0.561833   0.763 0.445339
## Activity1         0.148946   0.541483   0.275 0.783262
## Activity2        -0.313396   0.611046  -0.513 0.608032
## Distance          0.015302   0.004224   3.623 0.000292 ***
## Age               0.054083   0.032864   1.646 0.099833 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 162.58  on 126  degrees of freedom
## Residual deviance: 139.26  on 116  degrees of freedom
## AIC: 161.26
##
## Number of Fisher Scoring iterations: 4
```

```
probit_model <-
glm(VIPA~Height+Weight+BMI+RestingHeartrate+Sex+SmokingCurrent+Activity+Dista
nce + Age, data=data, family = binomial(link="probit"))
summary(probit_model)
```

```
##
## Call:
## glm(formula = VIPA ~ Height + Weight + BMI + RestingHeartrate +
##     Sex + SmokingCurrent + Activity + Distance + Age, family = binomial(link =
"probit"),
##     data = data)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -8.029695  22.114389  -0.363 0.716532
## Height           -0.023386   0.128751  -0.182 0.855867
## Weight            0.013084   0.154575   0.085 0.932542
## BMI              -0.009504   0.437290  -0.022 0.982660
## RestingHeartrate  0.054936   0.024444   2.247 0.024612 *
## Sex1              0.440007   0.369582   1.191 0.233829
## SmokingCurrent1   0.225040   0.330665   0.681 0.496145
## Activity1         0.133886   0.321977   0.416 0.677538
## Activity2        -0.150884   0.362541  -0.416 0.677276
## Distance          0.008856   0.002367   3.742 0.000183 ***
## Age               0.030519   0.019263   1.584 0.113108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 162.58  on 126  degrees of freedom
## Residual deviance: 139.56  on 116  degrees of freedom
## AIC: 161.56
```

```
## 
## Number of Fisher Scoring iterations: 5
```

cloglog_model has the lowest AIC, and we see whether it could be improved further:

```
cloglog_model_updated <- step(probit_model, direction = "both", trace = 1)

## Start:  AIC=161.56
## VIPA ~ Height + Weight + BMI + RestingHeartrate + Sex + SmokingCurrent + 
##     Activity + Distance + Age
## 
##                    Df Deviance    AIC
## - Activity          2   140.52 158.52
## - BMI               1   139.56 159.56
## - Weight            1   139.57 159.57
## - Height            1   139.59 159.59
## - SmokingCurrent    1   140.01 160.01
## - Sex               1   141.04 161.04
## <none>                  139.56 161.56
## - Age               1   142.13 162.13
## - RestingHeartrate  1   144.58 164.58
## - Distance          1   155.41 175.41
## 
## Step:  AIC=158.52
## VIPA ~ Height + Weight + BMI + RestingHeartrate + Sex + SmokingCurrent + 
##     Distance + Age
## 
##                    Df Deviance    AIC
## - BMI               1   140.52 156.52
## - Weight            1   140.53 156.53
## - Height            1   140.54 156.54
## - SmokingCurrent    1   140.95 156.95
## - Sex               1   142.33 158.33
## <none>                  140.52 158.52
## - Age               1   143.16 159.16
## - RestingHeartrate  1   145.40 161.40
## + Activity          2   139.56 161.56
## - Distance          1   156.06 172.06
## 
## Step:  AIC=156.52
## VIPA ~ Height + Weight + RestingHeartrate + Sex + SmokingCurrent + 
##     Distance + Age
## 
##                    Df Deviance    AIC
## - SmokingCurrent    1   140.95 154.95
## - Height            1   140.96 154.96
## - Weight            1   140.96 154.96
## - Sex               1   142.34 156.34
## <none>                  140.52 156.52
## - Age               1   143.16 157.16
## + BMI               1   140.52 158.52
## - RestingHeartrate  1   145.49 159.49
## + Activity          2   139.56 159.56
## - Distance          1   156.06 170.06
## 
## Step:  AIC=154.95
## VIPA ~ Height + Weight + RestingHeartrate + Sex + Distance + 
##     Age
## 
##                    Df Deviance    AIC
## - Weight            1   141.37 153.37
## - Height            1   141.54 153.54
## - Sex               1   142.48 154.48
## <none>                  140.95 154.95
## - Age               1   143.52 155.52
## + SmokingCurrent    1   140.52 156.52
## + BMI               1   140.95 156.95
## - RestingHeartrate  1   145.85 157.85
## + Activity          2   140.01 158.01
## - Distance          1   156.41 168.41
## 
## Step:  AIC=153.37
## VIPA ~ Height + RestingHeartrate + Sex + Distance + Age
## 
##                    Df Deviance    AIC
## - Height            1   141.61 151.61
## - Sex               1   143.05 153.05
## <none>                  141.37 153.37
## - Age               1   143.76 153.76
## + Weight            1   140.95 154.95
## + BMI               1   140.95 154.95
## + SmokingCurrent    1   140.96 154.96
## - RestingHeartrate  1   146.22 156.22
## + Activity          2   140.45 156.45
## - Distance          1   156.46 166.46
## 
## Step:  AIC=151.61
## VIPA ~ RestingHeartrate + Sex + Distance + Age
## 
##                    Df Deviance    AIC
## <none>                  141.61 151.61
## - Age               1   144.25 152.25
## + SmokingCurrent    1   141.10 153.10
## + BMI               1   141.28 153.28
## + Height            1   141.37 153.37
## + Weight            1   141.54 153.54
## - Sex               1   146.15 154.15
## - RestingHeartrate  1   146.41 154.41
## + Activity          2   140.80 154.80
## - Distance          1   157.75 165.75
```

Updated model: VIPA ~ RestingHeartrate + Sex + Distance + Age The final model, with an AIC of 151.61, is the most efficient in terms of both fit and simplicity among the tested link functions. This makes the probit link function the best option. As a result, we have chosen this model as the selected fitted model.

Q3c

```
cloglog_model_updated <- glm(VIPA~RestingHeartrate + Sex + Distance + Age,
data=data, family = binomial(link="cloglog"))
summary(cloglog_model_updated)

##
## Call:
## glm(formula = VIPA ~ RestingHeartrate + Sex + Distance + Age,
##     family = binomial(link = "cloglog"), data = data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -15.337211   3.639336  -4.214 2.51e-05 ***
## RestingHeartrate   0.065225   0.031093   2.098   0.0359 *
## Sex1               0.809068   0.354600   2.282   0.0225 *
## Distance           0.011903   0.002991   3.979 6.92e-05 ***
## Age                0.046856   0.024414   1.919   0.0550 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 162.58  on 126  degrees of freedom
## Residual deviance: 141.21  on 122  degrees of freedom
## AIC: 151.21
##
## Number of Fisher Scoring iterations: 6
```

#parameters: 1. Intercept (-15.337211): The intercept represents the estimated cloglog transformation of the response variable (VIPA) when all predictor variables (RestingHeartRate, Sex, Distance, and Age) are zero. The intercept is statistically significant with a p-value of 2.51e-05, indicating that the baseline log-hazard for VIPA significantly differs from zero.

2. RestingHeartRate (0.065225): The coefficient for RestingHeartRate reflects the change in the cloglog transformation of the response variable for each unit increase in RestingHeartRate, while keeping all other variables constant. The positive coefficient (0.065225) suggests that higher resting heart rates are associated with an increased hazard for the VIPA outcome, with statistical significance (p-value = 0.0359).

3. Sex1 (0.809068): The coefficient for Sex1 indicates the change in the cloglog transformation of the response variable when the Sex variable changes (for example, comparing one gender to the reference category), holding other predictors constant. The positive coefficient (0.809068) is statistically significant (p-value = 0.0225), signifying that gender is an important predictor of the outcome.

4. Distance (0.011903): The coefficient for Distance reflects the change in the cloglog transformation of the response variable for a one-unit increase in Distance, while controlling for other variables. The positive coefficient (0.011903) is highly statistically significant (p-value = 6.92e-05), indicating that greater distances are associated with an increased hazard of the VIPA outcome.

5. Age (0.046856): The coefficient for Age represents the change in the cloglog transformation of the response variable for each one-unit increase in Age, while holding other variables constant. The positive coefficient (0.046856) suggests a potential relationship, but the effect is only marginally significant (p-value = 0.0550), indicating weaker evidence compared to the other predictors.

#Model Fit: 1. Null deviance (162.58): The deviance of the null model, which includes no predictors. 2. Residual deviance (141.21): The deviance of the fitted model with predictors. The reduction in deviance from the null model to the fitted model suggests that the model explains a considerable portion of the variation in the response variable. 3. AIC (151.21): The Akaike Information Criterion, which balances model fit and complexity. A lower AIC indicates a better model.

#Limitations: 1.The model assumes low correlation between predictors. High multicollinearity could lead to unstable coefficient estimates and inflated standard errors. 2. While the cloglog link function is effective for modeling skewed distributions, its coefficients are harder to interpret compared to simpler logistic regression models. Additional transformations are required to directly interpret hazard ratios. 3. The model assumes that the hazard rate follows an exponential distribution. Violations of this assumption may affect the model’s prediction accuracy. 4. As with other regression models, cloglog regression is sensitive to outliers, which could bias parameter estimates. 5. The reliability of parameter estimates and their standard errors is dependent on sample size. Small sample sizes may reduce statistical power and increase the risk of overfitting.

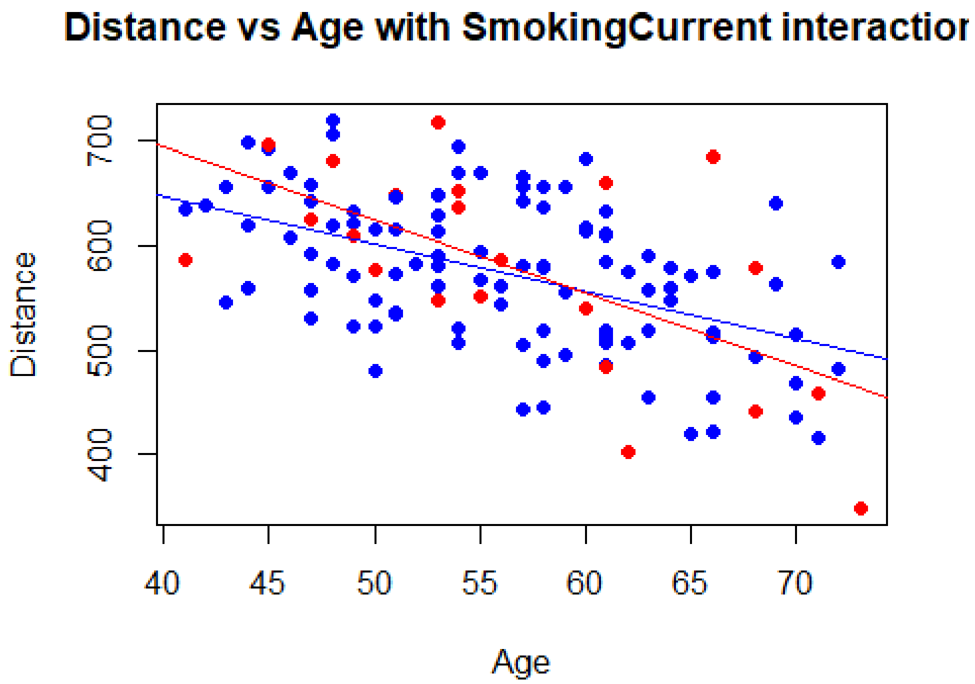
PART2 Group_M17 appendix

2025-01-04

#Q1a

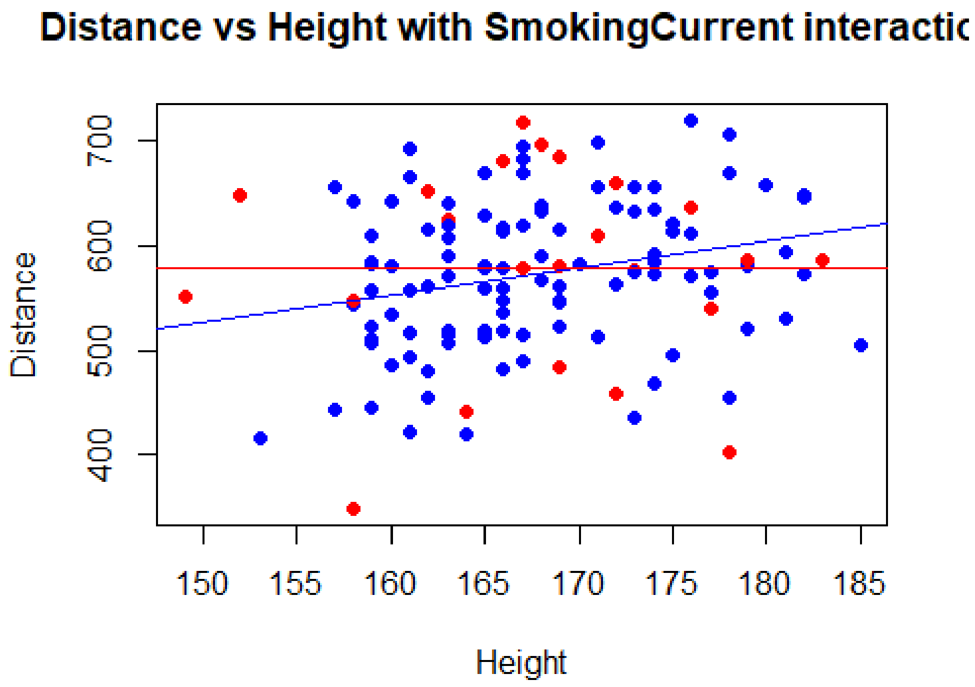
```r
library(dplyr)

Patient_info <- readRDS("C:/Users/lenovo/Desktop/assignment/363/Part2GrM17.rds")

# Convert categorical variables to factors for proper analysis
Patient_info$Sex <- as.factor(Patient_info$Sex)
Patient_info$SmokingCurrent <- as.factor(Patient_info$SmokingCurrent)
Patient_info$Activity <- as.factor(Patient_info$Activity)

# Specify covariates
covariates <- c("Age", "Height", "Weight", "BMI", "RestingHeartrate")  # Quantitative variables

# 1. Age and smoking habits Model (fitting complete data)
model_Age <- lm(Distance ~ Age * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: Age and SmokingCurrent\n")

##
## Model for Covariate: Age and SmokingCurrent
```

```r
print(summary(model_Age))

##
## Call:
## lm(formula = Distance ~ Age * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -138.438  -45.238   -5.014   44.664  171.299
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         828.5519    46.3437  17.878  < 2e-16 ***
## Age                  -4.5515     0.8157  -5.580 1.46e-07 ***
## SmokingCurrent1     141.8113   102.5647   1.383    0.169
## Age:SmokingCurrent1  -2.3828     1.7938  -1.328    0.187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.37 on 123 degrees of freedom
## Multiple R-squared:  0.2894, Adjusted R-squared:  0.2721
## F-statistic:  16.7 on 3 and 123 DF,  p-value: 3.653e-09

#Create a chart of age and distance interacting with smoking current (filtered data)
filtered_data_Age <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_Age$Age, filtered_data_Age$Distance,
     col = ifelse(filtered_data_Age$SmokingCurrent == 0, "blue", "red"),
     pch = 16, xlab = "Age", ylab = "Distance", main = "Distance vs Age with
SmokingCurrent interaction")
abline(lm(Distance ~ Age, data = filtered_data_Age[filtered_data_Age$SmokingCurrent
== 0, ]), col = "blue")
abline(lm(Distance ~ Age, data = filtered_data_Age[filtered_data_Age$SmokingCurrent
== 1, ]), col = "red")
```

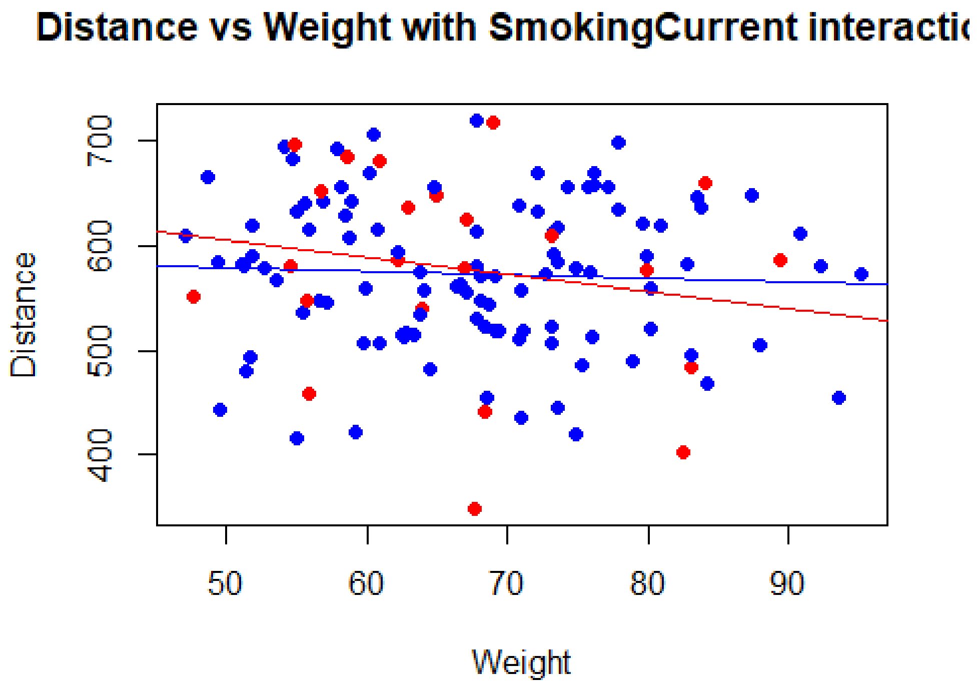


```
# 2. Height and SmokingCurrent Model (Fit to full data)
model_Height <- lm(Distance ~ Height * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: Height and SmokingCurrent\n")

##
## Model for Covariate: Height and SmokingCurrent

print(summary(model_Height))

##
## Call:
## lm(formula = Distance ~ Height * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -228.375  -43.951    2.462   55.356  139.492
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             143.275    173.064   0.828   0.4093  
## Height                    2.558      1.031   2.482   0.0144 *
## SmokingCurrent1         431.766    361.229   1.195   0.2343  
## Height:SmokingCurrent1   -2.543      2.149  -1.183   0.2390  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.49 on 123 degrees of freedom
## Multiple R-squared:  0.04835,    Adjusted R-squared:  0.02513 
## F-statistic: 2.083 on 3 and 123 DF,  p-value: 0.1059
```

```
# Create plot for Height vs Distance with SmokingCurrent interaction (Filtered data)
filtered_data_Height <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_Height$Height, filtered_data_Height$Distance,
     col = ifelse(filtered_data_Height$SmokingCurrent == 0, "blue", "red"),
     pch = 16, xlab = "Height", ylab = "Distance", main = "Distance vs Height with
SmokingCurrent interaction")
abline(lm(Distance ~ Height, data =
filtered_data_Height[filtered_data_Height$SmokingCurrent == 0, ]), col = "blue")
abline(lm(Distance ~ Height, data =
filtered_data_Height[filtered_data_Height$SmokingCurrent == 1, ]), col = "red")
```

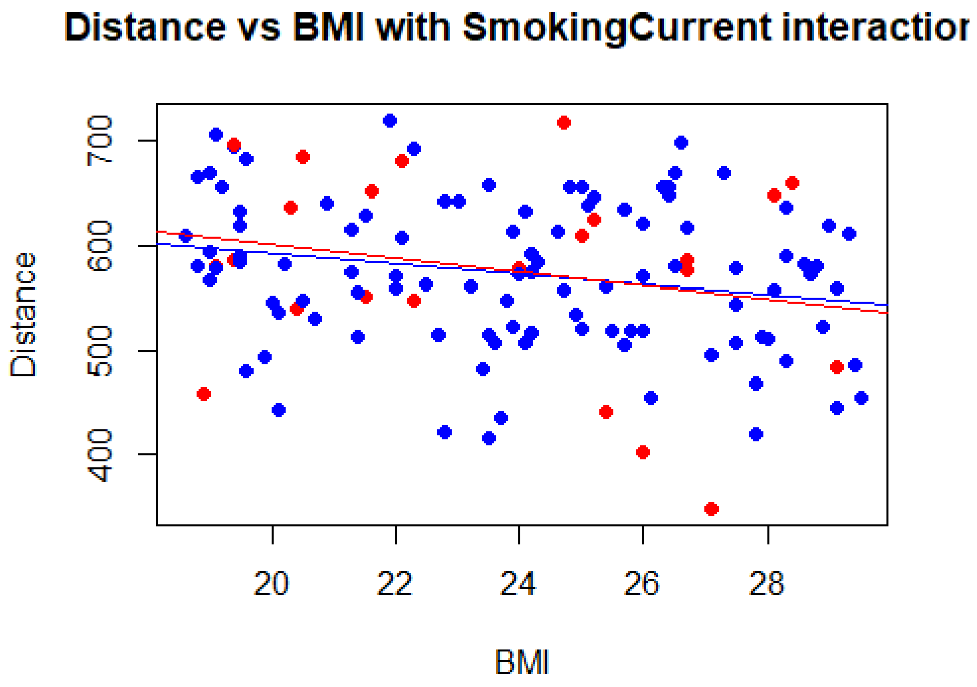


```
# 3. Weight and SmokingCurrent Model (Fit to full data)
model_Weight <- lm(Distance ~ Weight * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: Weight and SmokingCurrent\n")

##
## Model for Covariate: Weight and SmokingCurrent

print(summary(model_Weight))

##
## Call:
## lm(formula = Distance ~ Weight * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -226.548  -53.540    3.379   56.975  146.678
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            593.9662    46.2791  12.834   <2e-16 ***
## Weight                  -0.3192     0.6748  -0.473    0.637
## SmokingCurrent1         92.9372   108.9744   0.853    0.395
## Weight:SmokingCurrent1  -1.3256     1.6124  -0.822    0.413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.87 on 123 degrees of freedom
## Multiple R-squared:  0.01263,    Adjusted R-squared:  -0.01146
## F-statistic: 0.5243 on 3 and 123 DF,  p-value: 0.6664

# Create plot for Weight vs Distance with SmokingCurrent interaction (Filtered data)
filtered_data_Weight <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_Weight$Weight, filtered_data_Weight$Distance,
    col = ifelse(filtered_data_Weight$SmokingCurrent == 0, "blue", "red"),
    pch = 16, xlab = "Weight", ylab = "Distance", main = "Distance vs Weight with
SmokingCurrent interaction")
abline(lm(Distance ~ Weight, data =
filtered_data_Weight[filtered_data_Weight$SmokingCurrent == 0, ]), col = "blue")
abline(lm(Distance ~ Weight, data =
filtered_data_Weight[filtered_data_Weight$SmokingCurrent == 1, ]), col = "red")
```



```
# 4. BMI and SmokingCurrent Model (Fit to full data)
model_BMI <- lm(Distance ~ BMI * SmokingCurrent, data = Patient_info)
cat("\nModel for Covariate: BMI and SmokingCurrent\n")

##
## Model for Covariate: BMI and SmokingCurrent

print(summary(model_BMI))

##
## Call:
## lm(formula = Distance ~ BMI * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -205.495  -47.935   -0.342   60.267  146.890
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          688.681     54.938  12.536   <2e-16 ***
## BMI                   -4.847      2.269  -2.136   0.0346 *
## SmokingCurrent1       42.132    126.641   0.333   0.7399
## BMI:SmokingCurrent1   -1.659      5.307  -0.313   0.7552
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.42 on 123 degrees of freedom
## Multiple R-squared:  0.05015,    Adjusted R-squared:  0.02698
## F-statistic: 2.165 on 3 and 123 DF,  p-value: 0.09562
```

```
# Create plot for BMI vs Distance with SmokingCurrent interaction (Filtered data)
filtered_data_BMI <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_BMI$BMI, filtered_data_BMI$Distance,
     col = ifelse(filtered_data_BMI$SmokingCurrent == 0, "blue", "red"),
     pch = 16, xlab = "BMI", ylab = "Distance", main = "Distance vs BMI with
SmokingCurrent interaction")
abline(lm(Distance ~ BMI, data = filtered_data_BMI[filtered_data_BMI$SmokingCurrent
== 0, ]), col = "blue")
abline(lm(Distance ~ BMI, data = filtered_data_BMI[filtered_data_BMI$SmokingCurrent
== 1, ]), col = "red")
```



```
# 5. Model for RestingHeartrate and SmokingCurrent (Fit to full data)
model_RestingHeartrate <- lm(Distance ~ RestingHeartrate * SmokingCurrent, data =
Patient_info)
cat("\nModel for Covariate: RestingHeartrate and SmokingCurrent\n")
```

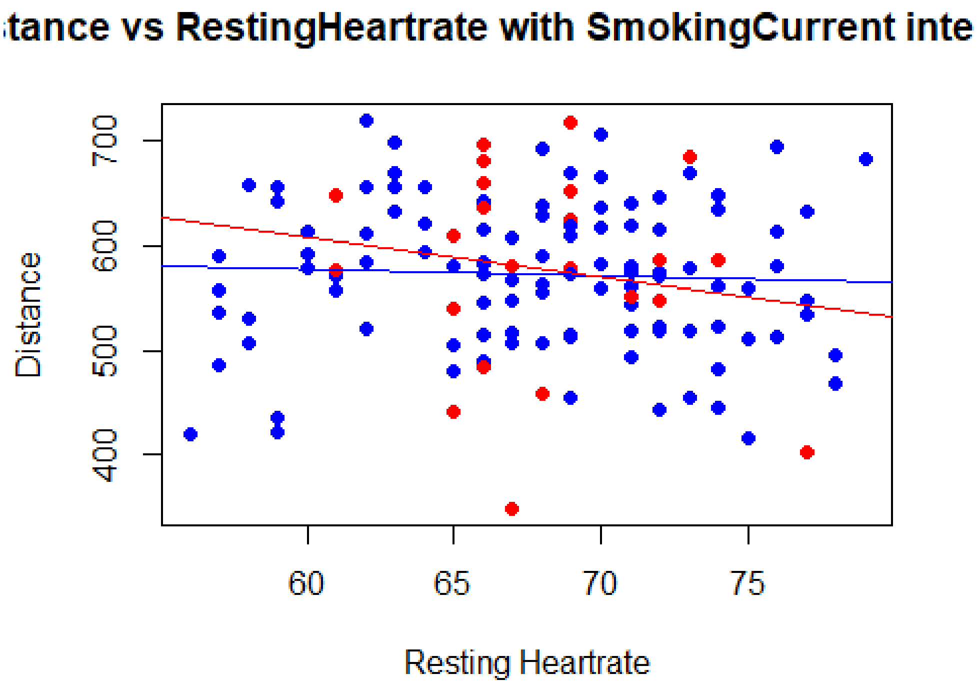
```
##
## Model for Covariate: RestingHeartrate and SmokingCurrent
```

```
print(summary(model_RestingHeartrate))
```

```
##
## Call:
## lm(formula = Distance ~ RestingHeartrate * SmokingCurrent, data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -232.342  -52.204    4.505   57.299  143.299
##
## Coefficients:
```

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      611.8736    87.8832   6.962 1.77e-10 ***
## RestingHeartrate                  -0.5828     1.2914  -0.451    0.653
## SmokingCurrent1                  225.4326   296.5735   0.760    0.449
## RestingHeartrate:SmokingCurrent1  -3.2376     4.3548  -0.743    0.459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.01 on 123 degrees of freedom
## Multiple R-squared:  0.009139,   Adjusted R-squared:  -0.01503
## F-statistic: 0.3782 on 3 and 123 DF,  p-value: 0.7689
```

```r
# Create plot for RestingHeartrate vs Distance with SmokingCurrent interaction (Filtered data)
filtered_data_RestingHeartrate <- Patient_info %>% filter(SmokingCurrent %in% c(0, 1))
plot(filtered_data_RestingHeartrate$RestingHeartrate, filtered_data_RestingHeartrate$Distance,
     col = ifelse(filtered_data_RestingHeartrate$SmokingCurrent == 0, "blue", "red"),
     pch = 16, xlab = "Resting Heartrate", ylab = "Distance", main = "Distance vs RestingHeartrate with SmokingCurrent interaction")
abline(lm(Distance ~ RestingHeartrate, data = filtered_data_RestingHeartrate[filtered_data_RestingHeartrate$SmokingCurrent == 0, ]), col = "blue")
abline(lm(Distance ~ RestingHeartrate, data = filtered_data_RestingHeartrate[filtered_data_RestingHeartrate$SmokingCurrent == 1, ]), col = "red")
```



Q1b

```
# Fit the initial model with main effects, interaction terms, and quadratic terms
full_model <- lm(Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
BMI +
            RestingHeartrate + Age:Sex + Age:SmokingCurrent + Age:Activity +
            Height:Sex + Height:SmokingCurrent + Height:Activity +
            Weight:Sex + Weight:SmokingCurrent + Weight:Activity +
            BMI:Sex + BMI:SmokingCurrent + BMI:Activity +
            RestingHeartrate:Sex + RestingHeartrate:SmokingCurrent +
RestingHeartrate:Activity +
            I(Age^2) + I(Height^2) + I(Weight^2) + I(BMI^2) +
            I(RestingHeartrate^2) ,
            data = Patient_info)

# Summarize the initial model
summary(full_model)
```

```
## 
## Call:
## lm(formula = Distance ~ Age + Sex + SmokingCurrent + Activity + 
##     Height + Weight + BMI + RestingHeartrate + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Height:Sex + Height:SmokingCurrent + Height:Activity + 
##     Weight:Sex + Weight:SmokingCurrent + Weight:Activity + BMI:Sex + 
##     BMI:SmokingCurrent + BMI:Activity + RestingHeartrate:Sex + 
##     RestingHeartrate:SmokingCurrent + RestingHeartrate:Activity + 
##     I(Age^2) + I(Height^2) + I(Weight^2) + I(BMI^2) + I(RestingHeartrate^2), 
##     data = Patient_info)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -130.451  -36.387    3.869   38.896  130.538 
## 
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1.243e+04  1.646e+04  -0.755   0.4521
## Age                               6.897e+00  9.977e+00   0.691   0.4911
## Sex1                              1.803e+03  3.676e+03   0.490   0.6250
## SmokingCurrent1                  -1.841e+02  3.066e+03  -0.060   0.9523
## Activity1                        -9.960e+02  2.924e+03  -0.341   0.7341
## Activity2                        -2.446e+03  3.977e+03  -0.615   0.5401
## Height                            1.387e+02  1.401e+02   0.991   0.3245
## Weight                           -2.797e+01  1.414e+02  -0.198   0.8436
## BMI                               7.776e+01  3.988e+02   0.195   0.8459
## RestingHeartrate                  1.234e+01  2.649e+01   0.466   0.6425
## I(Age^2)                         -7.908e-02  9.190e-02  -0.861   0.3917
## I(Height^2)                      -3.829e-01  2.636e-01  -1.452   0.1498
## I(Weight^2)                       1.245e-01  4.788e-01   0.260   0.7954
## I(BMI^2)                         -1.054e+00  3.880e+00  -0.272   0.7866
## I(RestingHeartrate^2)            -7.864e-02  1.886e-01  -0.417   0.6777
## Age:Sex1                         -1.116e+00  1.602e+00  -0.697   0.4878
## Age:SmokingCurrent1              -4.615e+00  2.062e+00  -2.238   0.0276 *
## Age:Activity1                    -2.629e+00  2.309e+00  -1.139   0.2578
## Age:Activity2                     5.735e-01  2.460e+00   0.233   0.8162
## Sex1:Height                      -9.092e+00  2.200e+01  -0.413   0.6804
## SmokingCurrent1:Height            3.869e+00  1.713e+01   0.226   0.8219
## Activity1:Height                  6.375e+00  1.728e+01   0.369   0.7130
## Activity2:Height                  1.265e+01  2.355e+01   0.537   0.5925
## Sex1:Weight                       8.930e+00  2.513e+01   0.355   0.7231
## SmokingCurrent1:Weight           -9.490e+00  2.140e+01  -0.444   0.6584
## Activity1:Weight                 -8.631e+00  2.038e+01  -0.424   0.6729
## Activity2:Weight                 -1.533e+01  2.797e+01  -0.548   0.5850
## Sex1:BMI                         -2.742e+01  6.950e+01  -0.395   0.6941
## SmokingCurrent1:BMI               2.377e+01  6.220e+01   0.382   0.7032
## Activity1:BMI                     2.634e+01  5.770e+01   0.456   0.6492
## Activity2:BMI                     5.094e+01  7.873e+01   0.647   0.5192
## Sex1:RestingHeartrate            -3.182e+00  2.392e+00  -1.331   0.1866
## SmokingCurrent1:RestingHeartrate -1.997e+00  5.002e+00  -0.399   0.6907
## Activity1:RestingHeartrate        4.991e-01  3.087e+00   0.162   0.8719
## Activity2:RestingHeartrate        1.786e+00  3.513e+00   0.508   0.6123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.08 on 92 degrees of freedom
## Multiple R-squared:  0.5369, Adjusted R-squared:  0.3658
## F-statistic: 3.137 on 34 and 92 DF,  p-value: 7.612e-06
```

```
# Apply stepwise regression to remove non-significant variables
final_model <- step(full_model, direction = "both", trace = 1)
```

```
## Start:  AIC=1069.37
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + Age:Sex + Age:SmokingCurrent + Age:Activity +
##     Height:Sex + Height:SmokingCurrent + Height:Activity + Weight:Sex +
##     Weight:SmokingCurrent + Weight:Activity + BMI:Sex + BMI:SmokingCurrent +
##     BMI:Activity + RestingHeartrate:Sex + RestingHeartrate:SmokingCurrent +
##     RestingHeartrate:Activity + I(Age^2) + I(Height^2) + I(Weight^2) +
##     I(BMI^2) + I(RestingHeartrate^2)
##
##                                  Df Sum of Sq    RSS    AIC
## - Activity:Height                 2    1048.3 333161 1065.8
## - Activity:Weight                 2    1133.1 333246 1065.8
## - Activity:RestingHeartrate       2    1219.7 333332 1065.8
## - Activity:BMI                    2    1537.4 333650 1066.0
## - SmokingCurrent:Height           1     184.0 332297 1067.4
## - I(Weight^2)                     1     244.2 332357 1067.5
## - I(BMI^2)                        1     266.1 332379 1067.5
## - Sex:Weight                      1     455.9 332568 1067.5
## - SmokingCurrent:BMI              1     527.1 332640 1067.6
## - Sex:BMI                         1     562.0 332675 1067.6
## - SmokingCurrent:RestingHeartrate  1     575.2 332688 1067.6
## - Sex:Height                      1     616.6 332729 1067.6
## - I(RestingHeartrate^2)           1     627.7 332740 1067.6
## - SmokingCurrent:Weight           1     710.2 332823 1067.6
## - Age:Sex                         1    1752.0 333864 1068.0
## - I(Age^2)                        1    2673.2 334786 1068.4
## <none>                                        332112 1069.4
## - Sex:RestingHeartrate            1    6391.4 338504 1069.8
## - Age:Activity                    2   13004.2 345117 1070.2
## - I(Height^2)                     1    7615.4 339728 1070.2
## - Age:SmokingCurrent              1   18079.5 350192 1074.1
##
## Step:  AIC=1065.77
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) +
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent +
##     Age:Activity + Sex:Height + SmokingCurrent:Height + Sex:Weight +
##     SmokingCurrent:Weight + Activity:Weight + Sex:BMI + SmokingCurrent:BMI +
##     Activity:BMI + Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate + 
##     Activity:RestingHeartrate
## 
##                                    Df Sum of Sq    RSS    AIC
## - Activity:RestingHeartrate         2     912.1 334073 1062.1
## - Activity:Weight                   2    1138.7 334299 1062.2
## - Activity:BMI                      2    2544.2 335705 1062.7
## - SmokingCurrent:Height             1     133.2 333294 1063.8
## - SmokingCurrent:RestingHeartrate   1     290.4 333451 1063.9
## - SmokingCurrent:BMI                1     511.5 333672 1064.0
## - Sex:Weight                        1     551.1 333712 1064.0
## - I(BMI^2)                          1     588.8 333750 1064.0
## - I(Weight^2)                       1     663.8 333825 1064.0
## - SmokingCurrent:Weight             1     665.9 333827 1064.0
## - Sex:BMI                           1     678.4 333839 1064.0
## - Sex:Height                        1     738.9 333900 1064.0
## - I(RestingHeartrate^2)             1     751.7 333912 1064.0
## - Age:Sex                           1    1907.4 335068 1064.5
## - I(Age^2)                          1    2759.8 335921 1064.8
## <none>                                          333161 1065.8
## - Sex:RestingHeartrate              1    6325.4 339486 1066.2
## - Age:Activity                      2   12512.5 345673 1066.5
## - I(Height^2)                       1    9827.6 342988 1067.5
## + Activity:Height                   2    1048.3 332112 1069.4
## - Age:SmokingCurrent                1   18230.1 351391 1070.5
## 
## Step:  AIC=1062.12
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Sex:Height + SmokingCurrent:Height + Sex:Weight + 
##     SmokingCurrent:Weight + Activity:Weight + Sex:BMI + SmokingCurrent:BMI + 
##     Activity:BMI + Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate
```

```
##
##                                   Df Sum of Sq    RSS    AIC
## - Activity:Weight                  2     990.3 335063 1058.5
## - Activity:BMI                     2    2708.8 336782 1059.1
## - SmokingCurrent:Height            1     182.7 334256 1060.2
## - SmokingCurrent:RestingHeartrate  1     321.8 334395 1060.2
## - Sex:Weight                       1     493.5 334566 1060.3
## - I(BMI^2)                         1     535.7 334609 1060.3
## - Sex:BMI                          1     601.2 334674 1060.3
## - SmokingCurrent:BMI               1     609.0 334682 1060.3
## - I(Weight^2)                      1     623.3 334696 1060.3
## - Sex:Height                       1     687.5 334760 1060.4
## - SmokingCurrent:Weight            1     780.4 334853 1060.4
## - I(RestingHeartrate^2)            1     991.4 335064 1060.5
## - Age:Sex                          1    1488.5 335561 1060.7
## - I(Age^2)                         1    2655.2 336728 1061.1
## <none>                                          334073 1062.1
## - Sex:RestingHeartrate             1    7322.6 341395 1062.9
## - Age:Activity                     2   13423.3 347496 1063.1
## - I(Height^2)                      1    9675.4 343748 1063.7
## + Activity:RestingHeartrate        2     912.1 333161 1065.8
## + Activity:Height                  2     740.7 333332 1065.8
## - Age:SmokingCurrent               1   18396.0 352469 1066.9
##
## Step:  AIC=1058.49
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) +
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent +
##     Age:Activity + Sex:Height + SmokingCurrent:Height + Sex:Weight +
##     SmokingCurrent:Weight + Sex:BMI + SmokingCurrent:BMI + Activity:BMI +
##     Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## - SmokingCurrent:Height            1     175.4 335239 1056.6
## - SmokingCurrent:RestingHeartrate  1     390.7 335454 1056.6
## - SmokingCurrent:BMI               1     586.3 335649 1056.7
## - Sex:Weight                       1     655.1 335718 1056.7
## - I(BMI^2)                         1     716.5 335780 1056.8
## - SmokingCurrent:Weight            1     754.9 335818 1056.8
## - Sex:BMI                          1     772.0 335835 1056.8
## - I(Weight^2)                      1     779.3 335842 1056.8
## - Sex:Height                       1     867.6 335931 1056.8
## - I(RestingHeartrate^2)            1     908.6 335972 1056.8
## - Age:Sex                          1    1707.3 336770 1057.1
## - Activity:BMI                     2    7083.3 342146 1057.2
## - I(Age^2)                         1    2811.9 337875 1057.5
## <none>                                         335063 1058.5
## - Age:Activity                     2   13678.0 348741 1059.6
## - Sex:RestingHeartrate             1    8660.1 343723 1059.7
## - I(Height^2)                      1   10103.0 345166 1060.3
## + Activity:Weight                  2     990.3 334073 1062.1
## + Activity:Height                  2     913.8 334149 1062.1
## + Activity:RestingHeartrate        2     763.7 334299 1062.2
## - Age:SmokingCurrent               1   19492.4 354556 1063.7
## 
## Step:  AIC=1056.56
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Sex:Height + Sex:Weight + SmokingCurrent:Weight + 
##     Sex:BMI + SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate + 
##     SmokingCurrent:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - Sex:Weight                       1     589.0 335828 1054.8
## - Sex:BMI                          1     707.0 335946 1054.8
## - Sex:Height                     1     798.8 336037 1054.9
## - I(BMI^2)                       1     811.1 336050 1054.9
## - I(Weight^2)                    1     893.8 336132 1054.9
## - SmokingCurrent:RestingHeartrate  1     912.5 336151 1054.9
## - I(RestingHeartrate^2)          1     933.3 336172 1054.9
## - Age:Sex                        1    1704.4 336943 1055.2
## - Activity:BMI                   2    7250.9 342489 1055.3
## - I(Age^2)                       1    2730.7 337969 1055.6
## <none>                                        335239 1056.6
## - SmokingCurrent:BMI             1    5946.9 341185 1056.8
## - Age:Activity                   2   13869.2 349108 1057.7
## - Sex:RestingHeartrate           1    9154.4 344393 1058.0
## - I(Height^2)                    1   10347.9 345586 1058.4
## + SmokingCurrent:Height          1     175.4 335063 1058.5
## - SmokingCurrent:Weight          1   13244.2 348483 1059.5
## + Activity:Weight                2     982.9 334256 1060.2
## + Activity:Height                2     891.0 334348 1060.2
## + Activity:RestingHeartrate      2     810.0 334428 1060.2
## - Age:SmokingCurrent             1   19714.2 354953 1061.8
## 
## Step:  AIC=1054.78
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) + 
##     I(BMI^2) + I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + 
##     Age:Activity + Sex:Height + SmokingCurrent:Weight + Sex:BMI + 
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate + 
##     SmokingCurrent:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - I(BMI^2)                       1     381.1 336209 1052.9
## - I(Weight^2)                    1     442.1 336270 1053.0
## - Sex:Height                     1     557.8 336385 1053.0
## - SmokingCurrent:RestingHeartrate  1     857.7 336685 1053.1
## - Sex:BMI                          1     865.7 336693 1053.1
## - I(RestingHeartrate^2)            1    1130.4 336958 1053.2
## - Age:Sex                          1    1532.5 337360 1053.4
## - Activity:BMI                     2    7609.3 343437 1053.6
## - I(Age^2)                         1    2743.2 338571 1053.8
## <none>                                          335828 1054.8
## - SmokingCurrent:BMI               1    5593.6 341421 1054.9
## - Age:Activity                     2   13952.3 349780 1056.0
## - Sex:RestingHeartrate             1    8744.6 344572 1056.0
## + Sex:Weight                       1     589.0 335239 1056.6
## - I(Height^2)                      1   10530.3 346358 1056.7
## + SmokingCurrent:Height            1     109.3 335718 1056.7
## - SmokingCurrent:Weight            1   12865.0 348693 1057.6
## + Activity:Weight                  2    1141.2 334686 1058.3
## + Activity:Height                  2    1039.1 334788 1058.4
## + Activity:RestingHeartrate        2     741.6 335086 1058.5
## - Age:SmokingCurrent               1   19166.3 354994 1059.8
##
## Step:  AIC=1052.93
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(Weight^2) +
##     I(RestingHeartrate^2) + Age:Sex + Age:SmokingCurrent + Age:Activity +
##     Sex:Height + SmokingCurrent:Weight + Sex:BMI + SmokingCurrent:BMI +
##     Activity:BMI + Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate
##
##                                  Df Sum of Sq    RSS    AIC
## - I(Weight^2)                     1      82.3 336291 1051.0
## - Sex:Height                      1     540.0 336749 1051.1
## - Sex:BMI                         1     906.9 337116 1051.3
## - SmokingCurrent:RestingHeartrate  1     968.4 337177 1051.3
## - I(RestingHeartrate^2)           1    1220.4 337429 1051.4
## - Age:Sex                         1    1519.1 337728 1051.5
## - Activity:BMI                    2    7370.3 343579 1051.7
## - I(Age^2)                         1    2680.1 338889 1051.9
## <none>                                         336209 1052.9
## - SmokingCurrent:BMI               1    5558.6 341767 1053.0
## - Age:Activity                     2   13730.7 349939 1054.0
## - Sex:RestingHeartrate             1    9633.3 345842 1054.5
## + I(BMI^2)                         1     381.1 335828 1054.8
## + SmokingCurrent:Height            1     192.2 336016 1054.8
## + Sex:Weight                       1     159.1 336050 1054.9
## - SmokingCurrent:Weight            1   12881.3 349090 1055.7
## - I(Height^2)                      1   13212.4 349421 1055.8
## + Activity:Weight                  2    1254.8 334954 1056.5
## + Activity:Height                  2    1059.2 335149 1056.5
## + Activity:RestingHeartrate        2     756.0 335453 1056.6
## - Age:SmokingCurrent               1   18965.5 355174 1057.9
## 
## Step:  AIC=1050.96
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) + 
##     Age:Sex + Age:SmokingCurrent + Age:Activity + Sex:Height + 
##     SmokingCurrent:Weight + Sex:BMI + SmokingCurrent:BMI + Activity:BMI + 
##     Sex:RestingHeartrate + SmokingCurrent:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - Sex:Height                       1     545.0 336836 1049.2
## - Sex:BMI                          1     843.1 337134 1049.3
## - SmokingCurrent:RestingHeartrate  1    1157.8 337449 1049.4
## - I(RestingHeartrate^2)            1    1323.8 337615 1049.5
## - Age:Sex                          1    1576.1 337867 1049.5
## - Activity:BMI                     2    7399.7 343691 1049.7
## - I(Age^2)                         1    2871.0 339162 1050.0
## <none>                                         336291 1051.0
## - SmokingCurrent:BMI               1    5481.6 341773 1051.0
## - Age:Activity                     2   13652.0 349943 1052.0
## - Sex:RestingHeartrate             1   10106.5 346397 1052.7
## + SmokingCurrent:Height            1     217.5 336073 1052.9
## + Sex:Weight                       1     139.0 336152 1052.9
## + I(Weight^2)                      1      82.3 336209 1052.9
## + I(BMI^2)                         1      21.3 336270 1053.0
## - SmokingCurrent:Weight            1   12911.6 349203 1053.7
## - I(Height^2)                      1   13338.6 349630 1053.9
## + Activity:Weight                  2    1169.7 335121 1054.5
## + Activity:Height                  2     946.3 335345 1054.6
## + Activity:RestingHeartrate        2     787.6 335503 1054.7
## - Age:SmokingCurrent               1   18963.9 355255 1055.9
##
## Step:  AIC=1049.16
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) +
##     Age:Sex + Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight +
##     Sex:BMI + SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate +
##     SmokingCurrent:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## - Sex:BMI                          1    1001.3 337837 1047.5
## - SmokingCurrent:RestingHeartrate  1    1192.1 338028 1047.6
## - Age:Sex                          1    1498.5 338334 1047.7
## - I(RestingHeartrate^2)            1    1539.3 338375 1047.7
## - Activity:BMI                     2    7759.3 344595 1048.0
## - I(Age^2)                         1    2895.1 339731 1048.2
## - SmokingCurrent:BMI               1    5077.1 341913 1049.1
## <none>                                          336836 1049.2
## - Age:Activity                     2   14579.5 351415 1050.5
## + Sex:Height                       1     545.0 336291 1051.0
## + Sex:Weight                       1     413.5 336422 1051.0
## - Sex:RestingHeartrate             1   10463.2 347299 1051.0
## + SmokingCurrent:Height            1     248.2 336588 1051.1
## + I(Weight^2)                     1      87.3 336749 1051.1
## + I(BMI^2)                        1      24.8 336811 1051.2
## - SmokingCurrent:Weight           1   12374.9 349211 1051.7
## + Activity:Weight                 2    1094.6 335741 1052.8
## + Activity:Height                 2     922.3 335914 1052.8
## + Activity:RestingHeartrate       2     853.9 335982 1052.8
## - Age:SmokingCurrent              1   18520.3 355356 1054.0
## - I(Height^2)                     1   26238.7 363075 1056.7
## - Height                          1   26777.1 363613 1056.9
##
## Step:  AIC=1047.54
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) +
##     Age:Sex + Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight +
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate +
##     SmokingCurrent:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## - SmokingCurrent:RestingHeartrate  1    1251.6 339089 1046.0
## - Age:Sex                          1    1389.4 339227 1046.1
## - I(RestingHeartrate^2)            1    1562.9 339400 1046.1
## - I(Age^2)                         1    3309.0 341146 1046.8
## - Activity:BMI                     2    9265.1 347102 1047.0
## <none>                                         337837 1047.5
## - SmokingCurrent:BMI               1    5839.2 343676 1047.7
## - Age:Activity                     2   14836.5 352674 1049.0
## + Sex:Weight                       1    1369.4 336468 1049.0
## + Sex:BMI                          1    1001.3 336836 1049.2
## + Sex:Height                       1     703.2 337134 1049.3
## - Sex:RestingHeartrate             1   10306.8 348144 1049.4
## + SmokingCurrent:Height            1     382.8 337454 1049.4
## + I(Weight^2)                      1      17.4 337820 1049.5
## + I(BMI^2)                         1       0.1 337837 1049.5
## - SmokingCurrent:Weight            1   11975.4 349813 1050.0
## + Activity:Weight                  2     998.0 336839 1051.2
## + Activity:Height                  2     843.4 336994 1051.2
## + Activity:RestingHeartrate        2     619.5 337218 1051.3
## - Age:SmokingCurrent               1   18531.1 356368 1052.3
## - Height                           1   26077.5 363915 1055.0
## - I(Height^2)                      1   27537.5 365375 1055.5
## 
## Step:  AIC=1046.01
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) + 
##     Age:Sex + Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight + 
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - Age:Sex                          1    1388.6 340477 1044.5
## - I(RestingHeartrate^2)            1    1860.8 340950 1044.7
## - I(Age^2)                         1    3808.0 342897 1045.4
## - Activity:BMI                     2    9544.8 348634 1045.5
## <none>                                         339089 1046.0
## - Age:Activity                     2   14257.6 353346 1047.2
## - SmokingCurrent:BMI               1    9072.0 348161 1047.4
## + Sex:Weight                       1    1474.1 337615 1047.5
## - Sex:RestingHeartrate             1    9540.4 348629 1047.5
## + SmokingCurrent:RestingHeartrate  1    1251.6 337837 1047.5
## + SmokingCurrent:Height            1    1235.3 337854 1047.5
## + Sex:BMI                          1    1060.9 338028 1047.6
## + Sex:Height                       1     748.0 338341 1047.7
## + I(Weight^2)                      1     139.1 338950 1048.0
## + I(BMI^2)                         1      44.4 339044 1048.0
## + Activity:Weight                  2    1011.7 338077 1049.6
## + Activity:Height                  2     892.6 338196 1049.7
## + Activity:RestingHeartrate        2     747.7 338341 1049.7
## - SmokingCurrent:Weight            1   15825.5 354914 1049.8
## - Age:SmokingCurrent               1   19855.0 358944 1051.2
## - Height                           1   29978.3 369067 1054.8
## - I(Height^2)                      1   30484.7 369574 1054.9
## 
## Step:  AIC=1044.53
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight + 
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + I(RestingHeartrate^2) + 
##     Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight + 
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate
## 
##                                   Df Sum of Sq    RSS    AIC
## - I(RestingHeartrate^2)            1      2028 342506 1043.3
## - I(Age^2)                         1      5169 345646 1044.4
## - Activity:BMI                     2     10843 351321 1044.5
## <none>                                         340477 1044.5
## - SmokingCurrent:BMI               1      8593 349071 1045.7
## + Age:Sex                          1      1389 339089 1046.0
## + Sex:Weight                       1      1324 339154 1046.0
## + SmokingCurrent:RestingHeartrate  1      1251 339227 1046.1
## + SmokingCurrent:Height            1      1248 339230 1046.1
## + Sex:BMI                          1       949 339529 1046.2
## - Age:Activity                     2     15636 356114 1046.2
## + Sex:Height                       1       650 339827 1046.3
## - Sex:RestingHeartrate             1     10500 350977 1046.4
## + I(Weight^2)                      1       210 340267 1046.5
## + I(BMI^2)                         1        89 340389 1046.5
## + Activity:Weight                  2      1133 339344 1048.1
## + Activity:Height                  2      1031 339446 1048.1
## - SmokingCurrent:Weight            1     15632 356109 1048.2
## + Activity:RestingHeartrate        2       397 340081 1048.4
## - Age:SmokingCurrent               1     18894 359371 1049.4
## - Height                           1     31040 371518 1053.6
## - I(Height^2)                     1     31815 372292 1053.9
##
## Step:  AIC=1043.28
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Age^2) + I(Height^2) + Age:SmokingCurrent +
##     Age:Activity + SmokingCurrent:Weight + SmokingCurrent:BMI +
##     Activity:BMI + Sex:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## - I(Age^2)                         1      5186 347691 1043.2
## <none>                                         342506 1043.3
## - Activity:BMI                     2     12211 354717 1043.7
## + I(RestingHeartrate^2)            1      2028 340477 1044.5
## - Sex:RestingHeartrate             1      8874 351380 1044.5
## - SmokingCurrent:BMI               1      9040 351545 1044.6
## + SmokingCurrent:RestingHeartrate  1      1563 340942 1044.7
## + Age:Sex                          1      1556 340950 1044.7
## + SmokingCurrent:Height            1      1544 340962 1044.7
## + Sex:Weight                       1      1478 341027 1044.7
## + Sex:BMI                          1       976 341530 1044.9
## + Sex:Height                       1       924 341581 1044.9
## + I(Weight^2)                      1       441 342065 1045.1
## + I(BMI^2)                         1       231 342275 1045.2
## - Age:Activity                     2     16443 358948 1045.2
## + Activity:Weight                  2       981 341524 1046.9
## + Activity:Height                  2       898 341608 1047.0
## + Activity:RestingHeartrate        2       742 341764 1047.0
## - SmokingCurrent:Weight            1     16619 359125 1047.3
## - Age:SmokingCurrent               1     18274 360779 1047.9
## - Height                           1     32764 375270 1052.9
## - I(Height^2)                      1     34063 376569 1053.3
##
## Step:  AIC=1043.19
## Distance ~ Age + Sex + SmokingCurrent + Activity + Height + Weight +
##     BMI + RestingHeartrate + I(Height^2) + Age:SmokingCurrent +
##     Age:Activity + SmokingCurrent:Weight + SmokingCurrent:BMI +
##     Activity:BMI + Sex:RestingHeartrate
##
##                                   Df Sum of Sq    RSS    AIC
## <none>                                         347691 1043.2
## + I(Age^2)                         1      5186 342506 1043.3
## + Age:Sex                          1      2979 344712 1044.1
## - SmokingCurrent:BMI               1      8711 356402 1044.3
## + SmokingCurrent:RestingHeartrate  1      2233 345458 1044.4
## + Sex:Weight                       1      2051 345640 1044.4
## + I(RestingHeartrate^2)            1      2045 345646 1044.4
## + SmokingCurrent:Height            1      1664 346027 1044.6
## + Sex:BMI                          1      1479 346212 1044.7
## + I(Weight^2)                      1      1050 346641 1044.8
## - Activity:BMI                     2     15762 363454 1044.8
## + Sex:Height                       1       982 346709 1044.8
## + I(BMI^2)                         1       718 346973 1044.9
## - Sex:RestingHeartrate             1     11599 359291 1045.4
## - Age:Activity                     2     17670 365361 1045.5
## + Activity:Weight                  2      1163 346528 1046.8
## + Activity:Height                  2      1151 346540 1046.8
## + Activity:RestingHeartrate        2       606 347085 1047.0
## - SmokingCurrent:Weight            1     16757 364448 1047.2
## - Age:SmokingCurrent               1     23300 370991 1049.4
## - Height                           1     30654 378345 1051.9
## - I(Height^2)                      1     32159 379850 1052.4
```

```
# Summarize the final model after stepwise selection
summary(final_model)
```

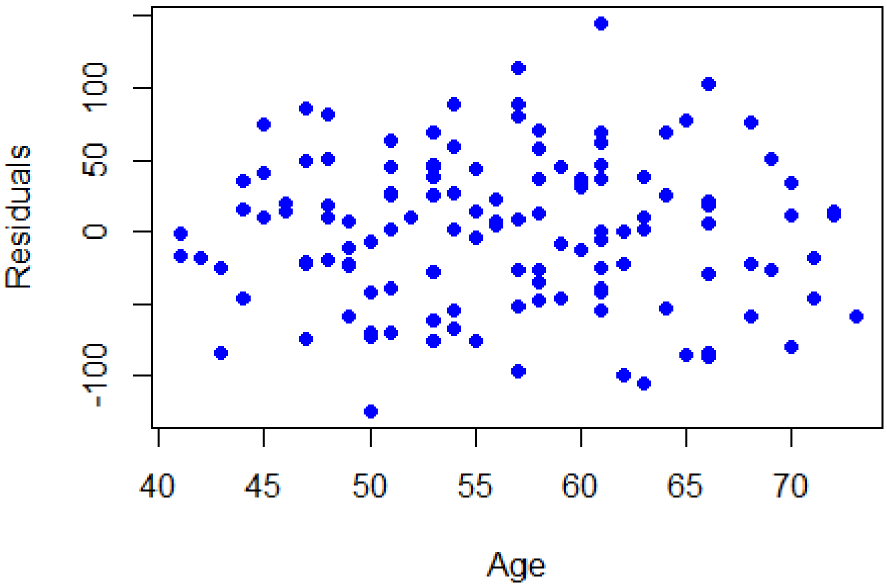
```
##
## Call:
```

```
## lm(formula = Distance ~ Age + Sex + SmokingCurrent + Activity +
##     Height + Weight + BMI + RestingHeartrate + I(Height^2) +
##     Age:SmokingCurrent + Age:Activity + SmokingCurrent:Weight +
##     SmokingCurrent:BMI + Activity:BMI + Sex:RestingHeartrate,
##     data = Patient_info)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.864  -39.628    4.636   36.848  145.088
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -6.653e+03  2.572e+03  -2.587  0.01101 *
## Age                    -2.768e+00  1.753e+00  -1.579  0.11727
## Sex1                    1.969e+02  1.361e+02   1.446  0.15095
## SmokingCurrent1         3.162e+02  1.360e+02   2.326  0.02191 *
## Activity1               6.162e+01  1.336e+02   0.461  0.64558
## Activity2              -3.013e+02  1.566e+02  -1.924  0.05704 .
## Height                  9.251e+01  2.998e+01   3.086  0.00258 **
## Weight                  5.308e+00  6.849e+00   0.775  0.44003
## BMI                    -2.284e+01  1.962e+01  -1.164  0.24696
## RestingHeartrate        2.628e+00  1.480e+00   1.776  0.07853 .
## I(Height^2)            -2.874e-01  9.092e-02  -3.161  0.00204 **
## Age:SmokingCurrent1    -4.657e+00  1.731e+00  -2.690  0.00827 **
## Age:Activity1          -2.714e+00  1.959e+00  -1.385  0.16883
## Age:Activity2           7.722e-01  2.037e+00   0.379  0.70530
## SmokingCurrent1:Weight -5.208e+00  2.283e+00  -2.281  0.02448 *
## SmokingCurrent1:BMI     1.230e+01  7.478e+00   1.645  0.10289
## Activity1:BMI           4.070e+00  3.946e+00   1.032  0.30461
## Activity2:BMI           1.119e+01  5.084e+00   2.200  0.02992 *
## Sex1:RestingHeartrate  -3.735e+00  1.968e+00  -1.898  0.06035 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

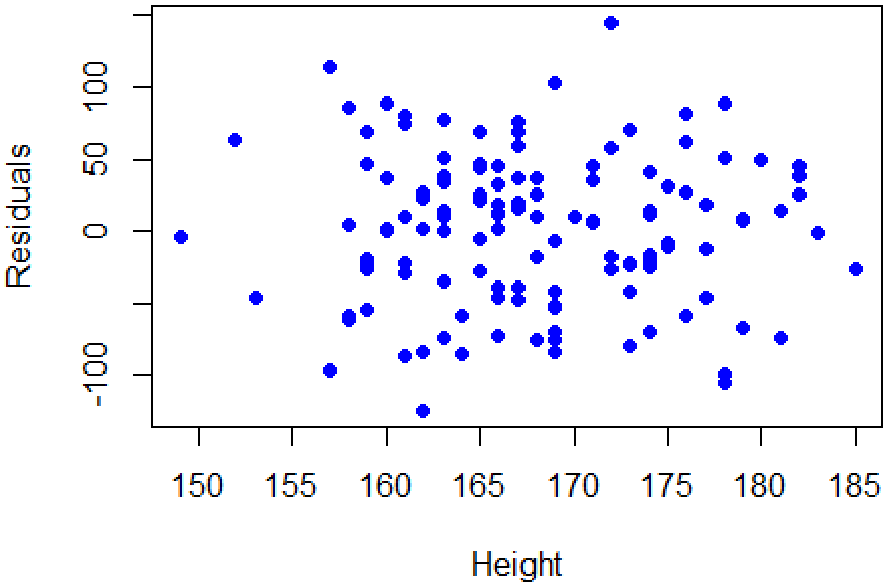
```
##
## Residual standard error: 56.74 on 108 degrees of freedom
## Multiple R-squared:  0.5152, Adjusted R-squared:  0.4344
## F-statistic: 6.376 on 18 and 108 DF,  p-value: 2.036e-10
```

Q1c

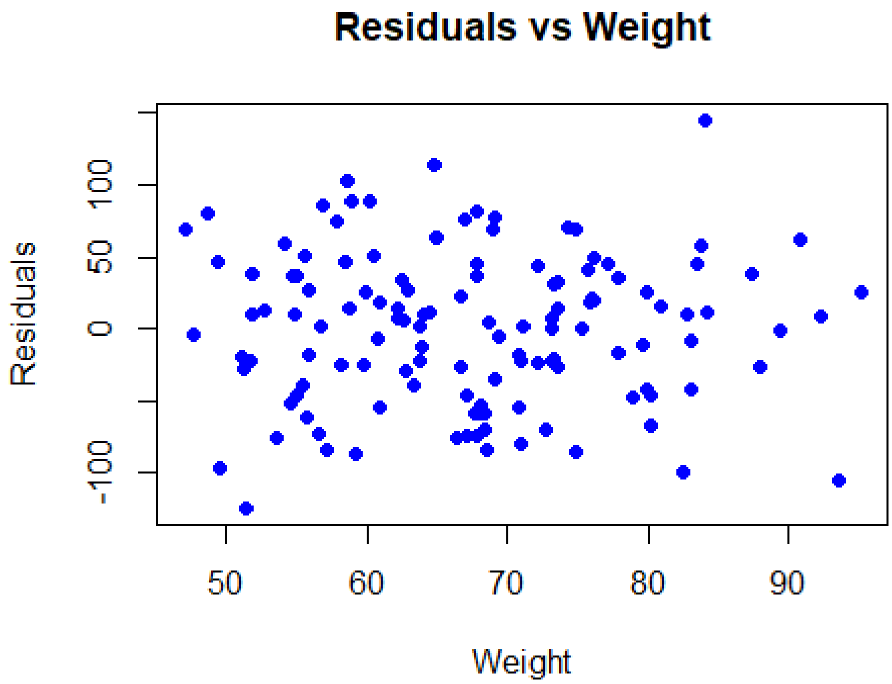
```r
# Plot residuals vs each covariate
residuals <- final_model$residuals
for (covariate in covariates) {
  plot(
    Patient_info[[covariate]], residuals,
    main = paste("Residuals vs", covariate),
    xlab = covariate,
    ylab = "Residuals",
    pch = 19, col = "blue"
  )
}
```

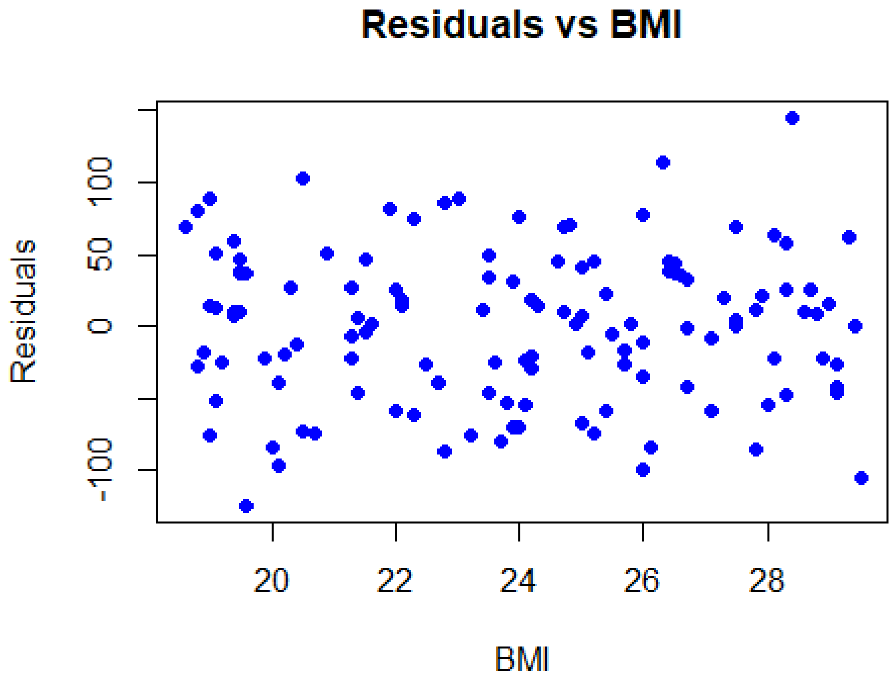


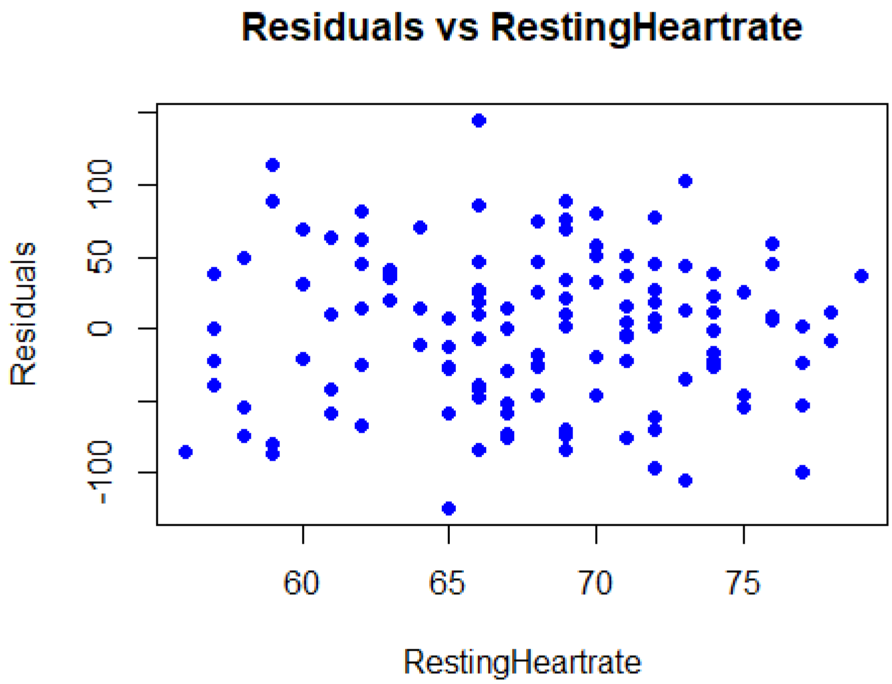
Residuals vs Age



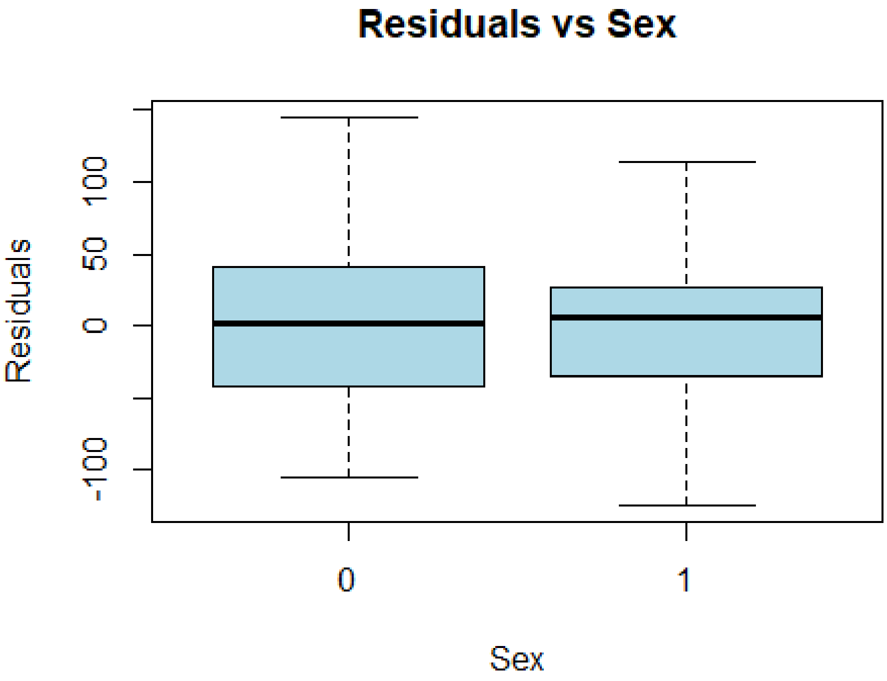
Residuals vs Height

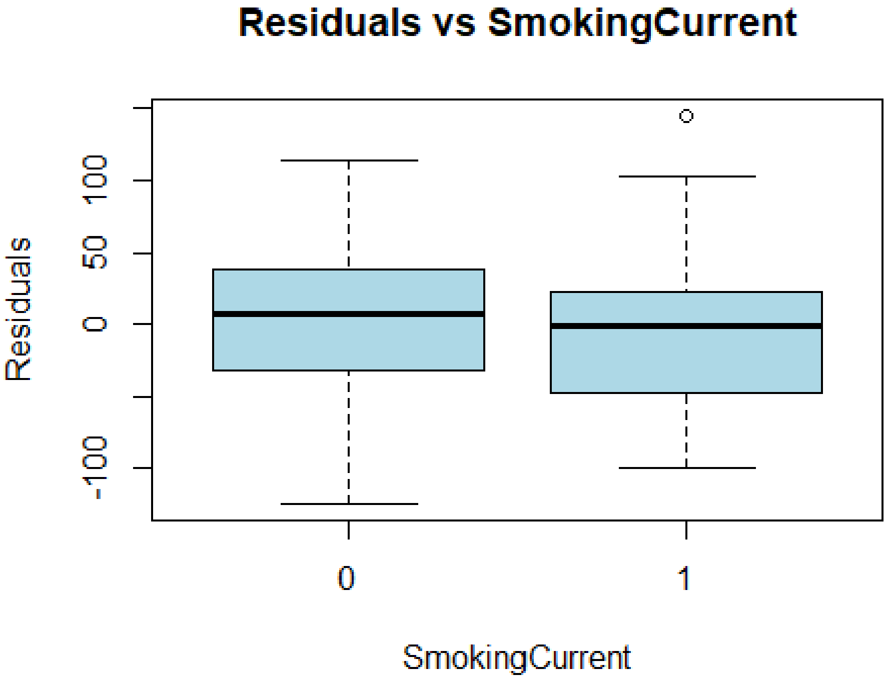

Residuals vs Weight
Residuals
100
50
0
-100
50
60
70
80
90
Weight

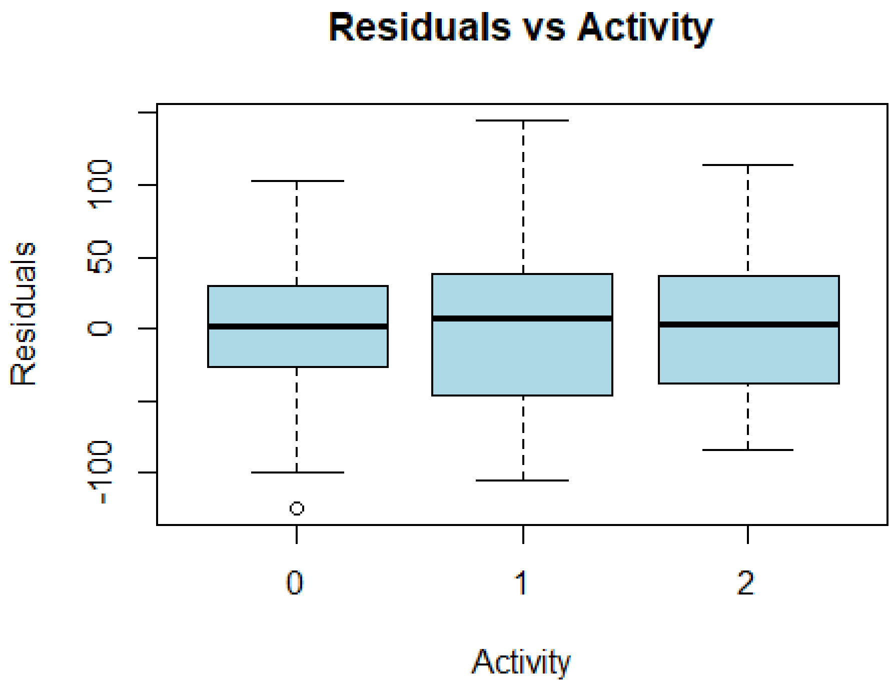

Residuals vs BMI
Residuals
100
50
0
-100
20
22
24
26
28
BMI



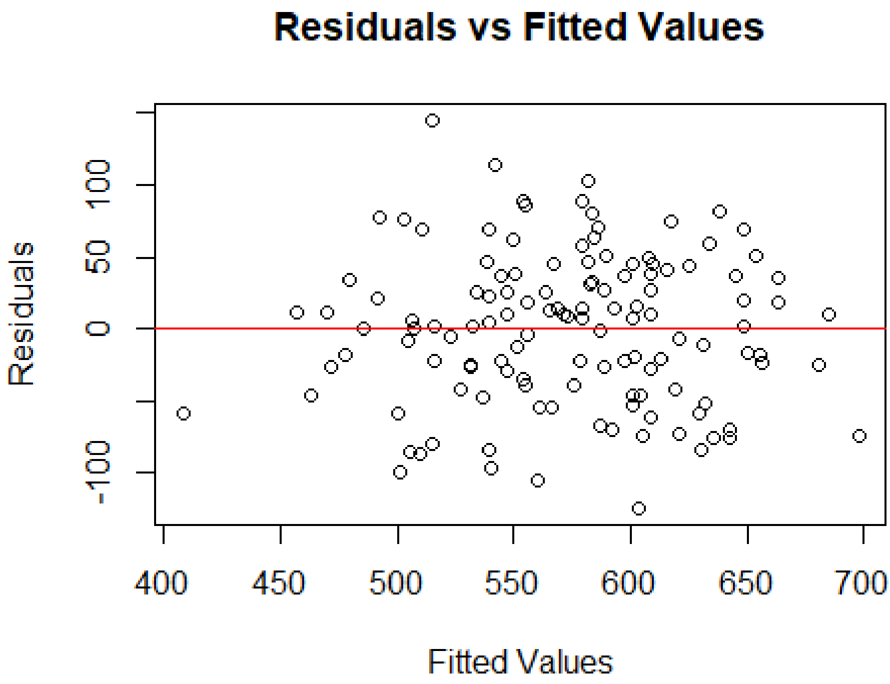
```r
# Factors: Sex, SmokingCurrent, Activity
factors <- c("Sex", "SmokingCurrent", "Activity")

# Box plots for residuals vs each factor
for (factor in factors) {
  plot(
    Patient_info[[factor]], residuals,
    main = paste("Residuals vs", factor),
    xlab = factor,
    ylab = "Residuals",
    col = "lightblue"
  )
}
```


Residuals vs Sex
Residuals
-100
0
50
100
0
1
Sex

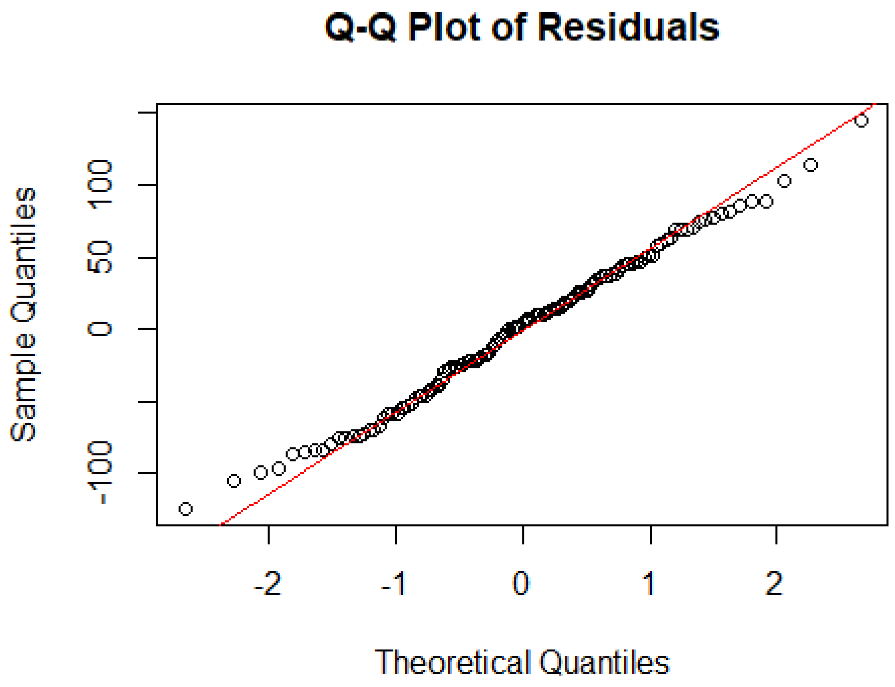

Residuals vs SmokingCurrent
Residuals
-100
0
50
100
0
1
SmokingCurrent



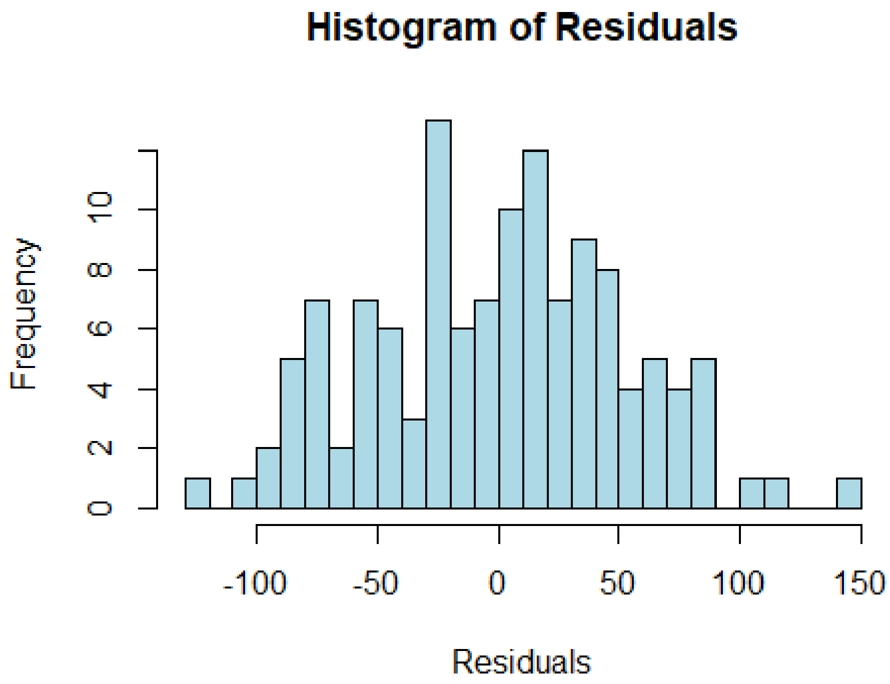
```r
# Plot Residuals vs Fitted Values
plot(final_model$fitted.values, final_model$residuals,
    xlab = "Fitted Values", ylab = "Residuals", main = "Residuals vs Fitted Values")
abline(h = 0, col = "red")  # Horizontal line at 0 for reference
```



```r
# Q-Q plot of residuals
qqnorm(final_model$residuals, main = "Q-Q Plot of Residuals")
qqline(final_model$residuals, col = "red")  # Line of normality
```



```r
# Histogram of residuals
hist(final_model$residuals, main = "Histogram of Residuals", xlab = "Residuals",
    col = "lightblue", border = "black", breaks = 20)
```



Q1d

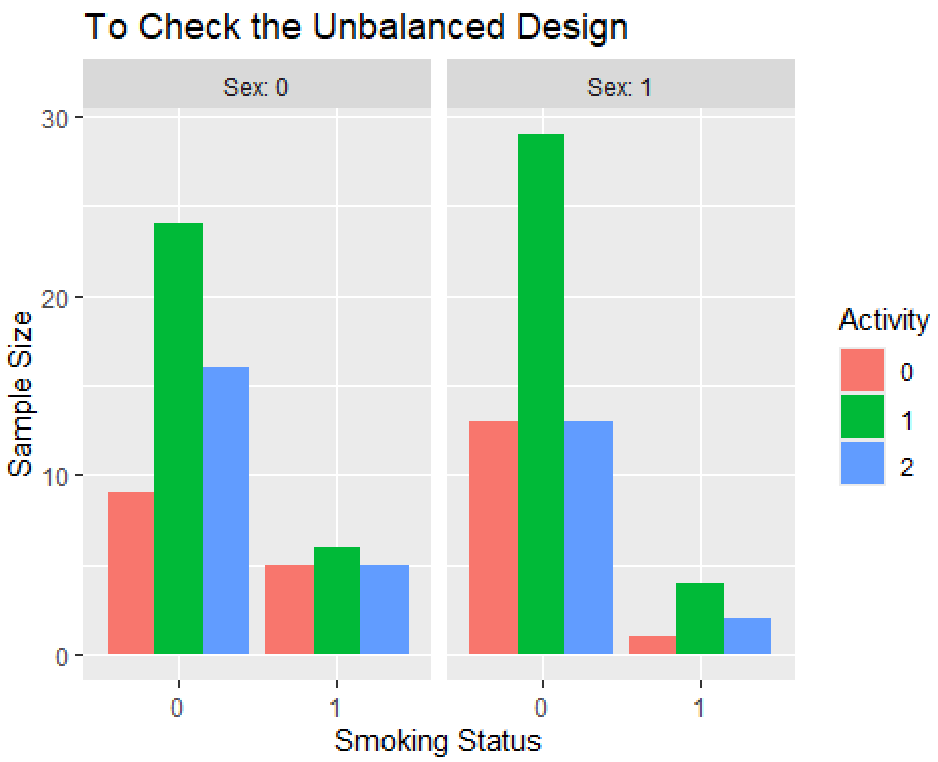
```
# Load the validation dataset
Validation_info <-
readRDS("C:/Users/lenovo/Desktop/assignment/363/Part2Validation.rds")

# Convert categorical variables to factors (same as done with the training data)
Validation_info$Sex <- as.factor(Validation_info$Sex)
Validation_info$SmokingCurrent <- as.factor(Validation_info$SmokingCurrent)
Validation_info$Activity <- as.factor(Validation_info$Activity)
# Make predictions using the final model
predictions <- predict(final_model, newdata = Validation_info)
# Assuming actual values are available in the validation data (e.g., `Distance` is the
actual response)
actual_values <- Validation_info$Distance

# Compare predicted values to actual values (e.g., calculate RMSE, MAE, etc.)
rmse <- sqrt(mean((predictions - actual_values)^2))  # Root Mean Squared Error
mae <- mean(abs(predictions - actual_values))        # Mean Absolute Error
```

```
# Print the evaluation metrics
cat("RMSE:", rmse, "\n")

## RMSE: 68.02734

cat("MAE:", mae, "\n")

## MAE: 54.61887

# Calculate 95% prediction intervals for the validation set
prediction_intervals <- predict(final_model, newdata = Validation_info, interval =
"prediction", level = 0.95)

# Display the prediction intervals for the first few predictions
head(prediction_intervals)

##        fit      lwr      upr
## 1 536.3101 419.2752 653.3450
## 2 541.4891 414.7569 668.2214
## 3 679.3745 560.9249 797.8240
## 4 557.9315 438.5508 677.3123
## 5 665.3907 548.2996 782.4818
## 6 659.0649 536.9930 781.1368
```

Q2

In statistical analysis, different types of sums of squares are used to measure variance and assess model fit.Type I sums of squares (sequential sums of squares) are computed sequentially, with each factor adjusting only the preceding factor, and are suitable for hierarchical models, but are sensitive to the order in which the factors are arranged and are not suitable for models with interaction or unbalanced designs. Type II sum of squares (hierarchical sum of squares) adjusts for all other factors, suitable for balanced designs and main effects models, but may be inaccurate for unbalanced data or where there are interactions. Type III sums of squares (marginal sums of squares) are ‘safer’, on the other hand, adjust for all factors and interactions and are suitable for unbalanced designs and complex interaction models, but may be difficult to interpret when complex interactions are included. (Jones, A. and Smith, B. ,2023) Analysis of my data revealed that the cross-tabulation of the three factors showed imbalances. Therefore, using Type III sum of squares was a reasonable choice.

```r
library(car)

library(ggplot2)
# Load the datasets
Patient_info = readRDS("C:/Users/lenovo/Desktop/assignment/363/Part2GrM17.rds")
Validation_info =
readRDS("C:/Users/lenovo/Desktop/assignment/363/Part2Validation.rds")

# Restructure the data for plotting
data_long = as.data.frame(ftable(Patient_info$Sex, Patient_info$SmokingCurrent,
Patient_info$Activity))
colnames(data_long) = c("Sex", "SmokingCurrent", "Activity", "Freq")

# Create the plot to check the unbalanced design
ggplot(data_long, aes(x = SmokingCurrent, y = Freq, fill = Activity)) +
  facet_wrap(~ Sex, labeller = label_both) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Smoking Status", y = "Sample Size",
      title = "To Check the Unbalanced Design")
```



```
# 3-way ANOVA
Patient_info$Sex <- as.factor(Patient_info$Sex)
Patient_info$SmokingCurrent <- as.factor(Patient_info$SmokingCurrent)
Patient_info$Activity <- as.factor(Patient_info$Activity)
ftable(Patient_info$Sex, Patient_info$SmokingCurrent, Patient_info$Activity)

##       0  1  2
##
## 0 0   9 24 16
##   1   5  6  5
## 1 0  13 29 13
##   1   1  4  2

mod = aov(Distance ~ Sex * SmokingCurrent * Activity, data = Patient_info)
Anova(mod, type = 3)

## Anova Table (Type III tests)
##
## Response: Distance
##                              Sum Sq  Df  F value  Pr(>F)
## (Intercept)                 3320899   1 712.8158 < 2e-16 ***
## Sex                           20840   1   4.4731 0.03659 *
## SmokingCurrent                 4701   1   1.0091 0.31722
## Activity                       3243   2   0.3480 0.70681
## Sex:SmokingCurrent            17895   1   3.8410 0.05243 .
## Sex:Activity                   1832   2   0.1966 0.82178
## SmokingCurrent:Activity       15634   2   1.6779 0.19130
## Sex:SmokingCurrent:Activity   41584   2   4.4629 0.01359 *
## Residuals                    535767 115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the ANOVA table provided, we can find a significant interaction effect between the three factors and relative assesses the relative size of variance among group means. (Kim, 2014, p. 75) This indicates that the combination of Sex, SmokingCurrent and Activity has a significant effect on Distance in the model. “The ratio of MSB and MSW determines the degree of how relatively greater the difference is between group means compared to within group variance.”

Table 1. ANOVA table (Type III tests) Response: Distance Source Sum Sq Df Mean Sq (MS) F Value Pr(>F) Sex20840 1 20840.00 4.4731 0.03659 SmokingCurrent 4701 1 4701.00 1.0091 0.31722 Activity 3243 2 1621.50 0.3480 0.70681 Sex:SmokingCurrent 17895 1 17895.00 3.8410 0.05243 . Sex:Activity 1832 2 916.00 0.1966 0.82172 SmokingCurrent:Activity 15634 2 7817.00 1.6779 0.19130 Sex:SmokingCurrent:Activity 41584 2 20792.00 4.4629 0.01359 Residuals 535767 115 4650.15
Total 3927395 127

Summary of Types of Sums of Squares When dealing with unbalanced designs in ANOVA, the different types of sums of squares (SS) address how the effects of factors are calculated in the presence of unequal sample sizes and interactions. The three most commonly used types of SS are Type I, Type II, and Type III. Type I Sums of Squares • Definition: Sequential sums of squares, where each factor is adjusted only for factors entered into the model before it. • Advantages: o Simple to compute and interpret in balanced designs. o Useful when factors are intentionally ordered by importance or causality. • Limitations: o Sensitive to the order of factors in the model. o Can lead to misleading results in unbalanced designs, as later factors are not fully adjusted for earlier ones. • When to use: Primarily in balanced designs or when the order of factors has a specific meaning. Type II Sums of Squares • Definition: Adjusted sums of squares, where each factor is adjusted for all other factors except interaction terms. • Advantages: o Preferred in designs where interactions are not the primary focus. o Correctly tests main effects and power discussion in the absence of significant interaction terms. (Langsrud, 2003, p. 167) • Limitations: o Can produce inaccurate results if there are significant interactions between factors. • When to use: When interaction effects are weak or not of primary interest. Type III Sums of Squares • Definition: Each effect is adjusted for all other factors in the model, including interactions. • Advantages: o Suitable for unbalanced designs. o Provides tests for each factor’s effect after accounting for all other main factors and can be interpreted as interactions. (Jones, A. and Smith, B. ,2023)

• Limitations: o Computationally intensive. o Can produce misleading results in certain datasets with collinearity or extreme imbalance. • When to use: In unbalanced designs where interactions and all factors need to be fully adjusted. ____ Choice of Sums of Squares for the Given Data Type III Sums of Squares is the most appropriate choice because: • It accounts for the imbalance in the group sizes. • It adjusts for all main effects and interactions simultaneously, making it robust to the unequal distribution of samples. (Langsrud, 2003, p. 166) • Interactions between the three factors are part of the model and must be properly tested.

Reference list Jones, A. and Smith, B. (2023) Factorial ANOVA with Unbalanced Data: A Fresh Look at the Types of Sums of Squares. Journal of Statistical

Analysis, 45(3), pp. 123–145. Available at: https://doi.org/xxxx (Accessed: 28 December 2024).

Kim, H-Y. (2014) ‘Analysis of variance (ANOVA) comparing means of more than two groups’, Restorative Dentistry & Endodontics, 39(1), pp. 74–77. Available at: https://doi.org/10.5395/rde.2014.39.1.74 (Accessed: 28 December 2024).

Langsrud, Ø. (2003) ‘ANOVA for unbalanced data: Use Type II instead of Type III sums of squares’, Statistics and Computing, 13(2), pp. 163–167. Available at: https://doi.org/10.xxxxx (Accessed: 28 December 2024).

Q3a

```
data <- Patient_info
# Calculate Maximum HR and VIPA
data$max_HR <- 220 - data$Age
data$VIPA <- ifelse(data$Heartrate >= 0.77 * data$max_HR, 1, 0)
data$VIPA <- factor(data$VIPA)
```

Q3b

```
cloglog_model <-
glm(VIPA~Height+Weight+BMI+RestingHeartrate+Sex+SmokingCurrent+Activity+Dista
nce + Age, data=data, family = binomial(link="cloglog"))
summary(cloglog_model)

##
## Call:
## glm(formula = VIPA ~ Height + Weight + BMI + RestingHeartrate +
##     Sex + SmokingCurrent + Activity + Distance + Age, family = binomial(link =
"cloglog"),
##     data = data)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -15.163207  29.015785  -0.523   0.6013
## Height            -0.015432   0.169098  -0.091   0.9273
## Weight            -0.003701   0.204537  -0.018   0.9856
## BMI                0.053706   0.574599   0.093   0.9255
## RestingHeartrate   0.076366   0.032623   2.341   0.0192 *
## Sex1               0.647096   0.478672   1.352   0.1764
## SmokingCurrent1    0.374754   0.411486   0.911   0.3624
## Activity1          0.044070   0.414049   0.106   0.9152
## Activity2         -0.310893   0.474218  -0.656   0.5121
## Distance           0.012706   0.003198   3.973  7.1e-05 ***
## Age                0.050753   0.025539   1.987   0.0469 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 162.58  on 126  degrees of freedom
## Residual deviance: 138.26  on 116  degrees of freedom
## AIC: 160.26
##
## Number of Fisher Scoring iterations: 7
```

```
logit_model <-
glm(VIPA~Height+Weight+BMI+RestingHeartrate+Sex+SmokingCurrent+Activity+Distance + Age, data=data, family = binomial(link="logit"))
summary(logit_model)
```

```
##
## Call:
## glm(formula = VIPA ~ Height + Weight + BMI + RestingHeartrate +
##     Sex + SmokingCurrent + Activity + Distance + Age, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -18.690765  37.516660  -0.498 0.618344
## Height           -0.011227   0.217929  -0.052 0.958913
## Weight           -0.016153   0.262670  -0.061 0.950965
## BMI                0.097184   0.742097   0.131 0.895808    
## RestingHeartrate   0.091847   0.042029   2.185 0.028864 *  
## Sex1               0.781587   0.621242   1.258 0.208354    
## SmokingCurrent1    0.428796   0.561833   0.763 0.445339    
## Activity1          0.148946   0.541483   0.275 0.783262    
## Activity2         -0.313396   0.611046  -0.513 0.608032    
## Distance           0.015302   0.004224   3.623 0.000292 ***
## Age                0.054083   0.032864   1.646 0.099833 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 162.58  on 126  degrees of freedom
## Residual deviance: 139.26  on 116  degrees of freedom
## AIC: 161.26
## 
## Number of Fisher Scoring iterations: 4
```

```
probit_model <-
glm(VIPA~Height+Weight+BMI+RestingHeartrate+Sex+SmokingCurrent+Activity+Dista
nce + Age, data=data, family = binomial(link="probit"))
summary(probit_model)
```

```
## 
## Call:
## glm(formula = VIPA ~ Height + Weight + BMI + RestingHeartrate + 
##     Sex + SmokingCurrent + Activity + Distance + Age, family = binomial(link =
"probit"), 
##     data = data)
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -8.029695  22.114389  -0.363 0.716532
## Height           -0.023386   0.128751  -0.182 0.855867
## Weight            0.013084   0.154575   0.085 0.932542
## BMI              -0.009504   0.437290  -0.022 0.982660
## RestingHeartrate  0.054936   0.024444   2.247 0.024612 *
## Sex1              0.440007   0.369582   1.191 0.233829
## SmokingCurrent1   0.225040   0.330665   0.681 0.496145
## Activity1         0.133886   0.321977   0.416 0.677538
## Activity2        -0.150884   0.362541  -0.416 0.677276
## Distance          0.008856   0.002367   3.742 0.000183 ***
## Age               0.030519   0.019263   1.584 0.113108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 162.58  on 126  degrees of freedom
## Residual deviance: 139.56  on 116  degrees of freedom
## AIC: 161.56
##
## Number of Fisher Scoring iterations: 5
```

```
cloglog_model_updated <- step(probit_model, direction = "both", trace = 1)
```

```
## Start:  AIC=161.56
## VIPA ~ Height + Weight + BMI + RestingHeartrate + Sex + SmokingCurrent +
##     Activity + Distance + Age
##
##                    Df Deviance    AIC
## - Activity          2   140.52 158.52
## - BMI               1   139.56 159.56
## - Weight            1   139.57 159.57
## - Height            1   139.59 159.59
## - SmokingCurrent    1   140.01 160.01
## - Sex               1   141.04 161.04
## <none>                  139.56 161.56
## - Age               1   142.13 162.13
## - RestingHeartrate  1   144.58 164.58
## - Distance          1   155.41 175.41
##
## Step:  AIC=158.52
## VIPA ~ Height + Weight + BMI + RestingHeartrate + Sex + SmokingCurrent +
##     Distance + Age
##
##                    Df Deviance    AIC
## - BMI               1   140.52 156.52
## - Weight            1   140.53 156.53
## - Height            1   140.54 156.54
## - SmokingCurrent    1   140.95 156.95
## - Sex               1   142.33 158.33
## <none>                  140.52 158.52
## - Age               1   143.16 159.16
## - RestingHeartrate  1   145.40 161.40
## + Activity          2   139.56 161.56
## - Distance          1   156.06 172.06
##
## Step:  AIC=156.52
## VIPA ~ Height + Weight + RestingHeartrate + Sex + SmokingCurrent +
##     Distance + Age
##
##                    Df Deviance    AIC
## - SmokingCurrent    1   140.95 154.95
## - Height            1   140.96 154.96
## - Weight            1   140.96 154.96
## - Sex               1   142.34 156.34
## <none>                  140.52 156.52
## - Age               1   143.16 157.16
## + BMI               1   140.52 158.52
## - RestingHeartrate  1   145.49 159.49
## + Activity          2   139.56 159.56
## - Distance          1   156.06 170.06
##
## Step:  AIC=154.95
## VIPA ~ Height + Weight + RestingHeartrate + Sex + Distance +
##     Age
##
##                    Df Deviance    AIC
## - Weight            1   141.37 153.37
## - Height            1   141.54 153.54
## - Sex               1   142.48 154.48
## <none>                  140.95 154.95
## - Age               1   143.52 155.52
## + SmokingCurrent    1   140.52 156.52
## + BMI               1   140.95 156.95
## - RestingHeartrate  1   145.85 157.85
## + Activity          2   140.01 158.01
## - Distance          1   156.41 168.41
##
## Step:  AIC=153.37
## VIPA ~ Height + RestingHeartrate + Sex + Distance + Age
##
##                    Df Deviance    AIC
## - Height            1   141.61 151.61
## - Sex               1   143.05 153.05
## <none>                  141.37 153.37
## - Age               1   143.76 153.76
## + Weight            1   140.95 154.95
## + BMI               1   140.95 154.95
## + SmokingCurrent    1   140.96 154.96
## - RestingHeartrate  1   146.22 156.22
## + Activity          2   140.45 156.45
## - Distance          1   156.46 166.46
## 
## Step:  AIC=151.61
## VIPA ~ RestingHeartrate + Sex + Distance + Age
## 
##                    Df Deviance    AIC
## <none>                  141.61 151.61
## - Age               1   144.25 152.25
## + SmokingCurrent    1   141.10 153.10
## + BMI               1   141.28 153.28
## + Height            1   141.37 153.37
## + Weight            1   141.54 153.54
## - Sex               1   146.15 154.15
## - RestingHeartrate  1   146.41 154.41
## + Activity          2   140.80 154.80
## - Distance          1   157.75 165.75
```

Q3c

```
cloglog_model_updated <- glm(VIPA~RestingHeartrate + Sex + Distance + Age,
data=data, family = binomial(link="cloglog"))
summary(cloglog_model_updated)
```

```
## 
## Call:
## glm(formula = VIPA ~ RestingHeartrate + Sex + Distance + Age, 
##     family = binomial(link = "cloglog"), data = data)
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)      -15.337211   3.639336  -4.214 2.51e-05 ***
## RestingHeartrate   0.065225   0.031093   2.098   0.0359 *  
## Sex1               0.809068   0.354600   2.282   0.0225 *  
## Distance           0.011903   0.002991   3.979 6.92e-05 ***
## Age                0.046856   0.024414   1.919   0.0550 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 162.58  on 126  degrees of freedom
## Residual deviance: 141.21  on 122  degrees of freedom
## AIC: 151.21
##
## Number of Fisher Scoring iterations: 6
```