Lan Vu HW5 Applied Regression

Run the code and library

```
library(readxl)
```

```
Warning: package 'readxl' was built under R version 4.5.1
```

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.5.1
```

```
Warning: package 'ggplot2' was built under R version 4.5.1
```

```
Warning: package 'tibble' was built under R version 4.5.1
```

```
Warning: package 'tidyr' was built under R version 4.5.1
```

```
Warning: package 'readr' was built under R version 4.5.1
```

```
Warning: package 'purrr' was built under R version 4.5.1
```

```
Warning: package 'dplyr' was built under R version 4.5.1
```

```
Warning: package 'stringr' was built under R version 4.5.1
```

```
Warning: package 'forcats' was built under R version 4.5.1
```

```
Warning: package 'lubridate' was built under R version 4.5.1
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(psych)
```

```
Warning: package 'psych' was built under R version 4.5.2
```

```
Attaching package: 'psych'

The following objects are masked from 'package:ggplot2':

    %+%, alpha
```

```
#df <- read_excel("~/Downloads/DCCT.xlsx")
df<- read_excel("C:/Users/lanvu/Downloads/DCCT.xlsx")
View(df)
```

Part 1

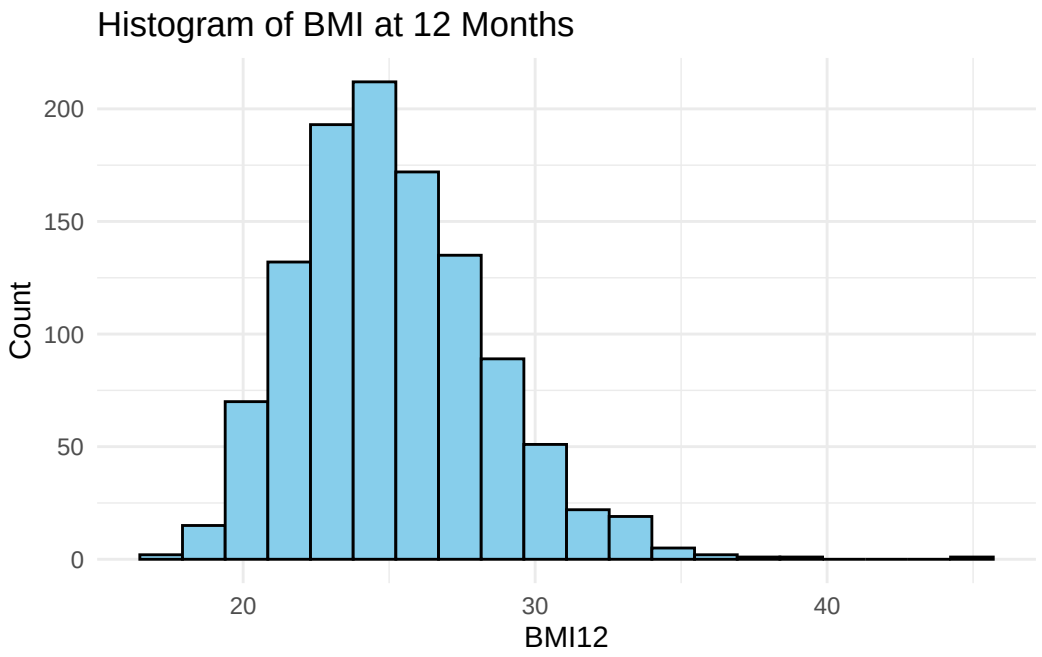
Question 1:Univariate Analysis

```
df %>%
  select(BMI12, BMI00, AGE) %>%
  psych::describe()
```

```
      vars    n  mean   sd median trimmed  mad   min   max range skew kurtosis
BMI12    1 1122 25.14 3.28  24.75   24.95 3.16 17.19 44.96 27.77 0.73     1.32
BMI00    2 1130 23.72 2.74  23.58   23.61 2.78 16.28 34.39 18.11 0.41     0.01
AGE      3 1130 34.86 5.79  35.00   34.83 7.41 21.00 48.00 27.00 0.03    -
0.80
        se
BMI12 0.10
BMI00 0.08
AGE   0.17
```

```r
#BMI 12
ggplot(df, aes(x = BMI12)) +
  geom_histogram(bins = 20, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of BMI at 12 Months", x = "BMI12", y = "Count")
```

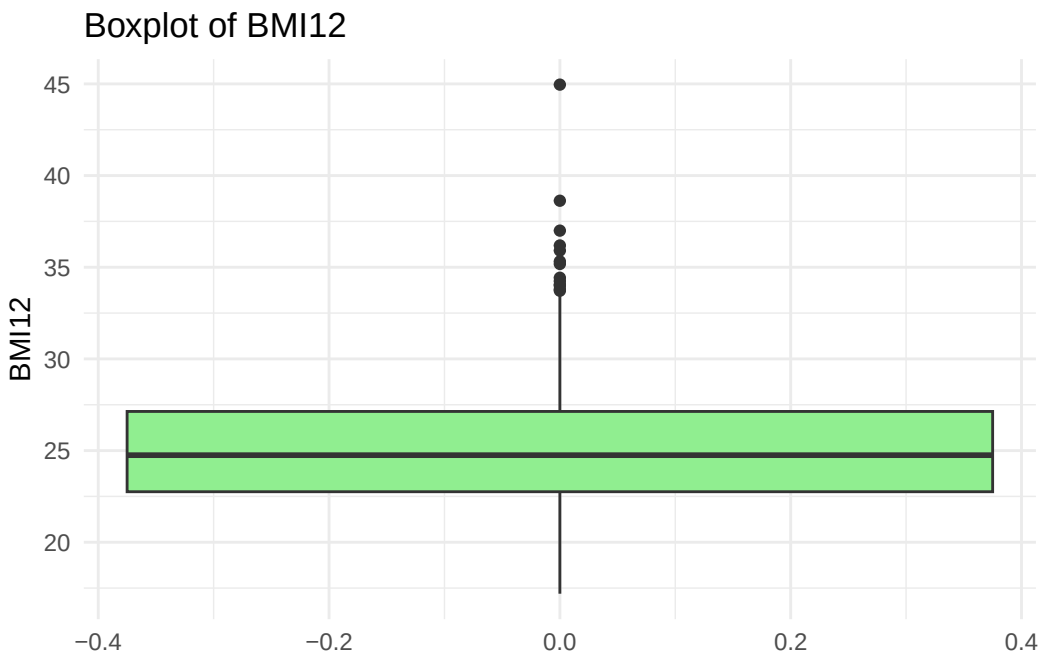
```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_bin()`).
```



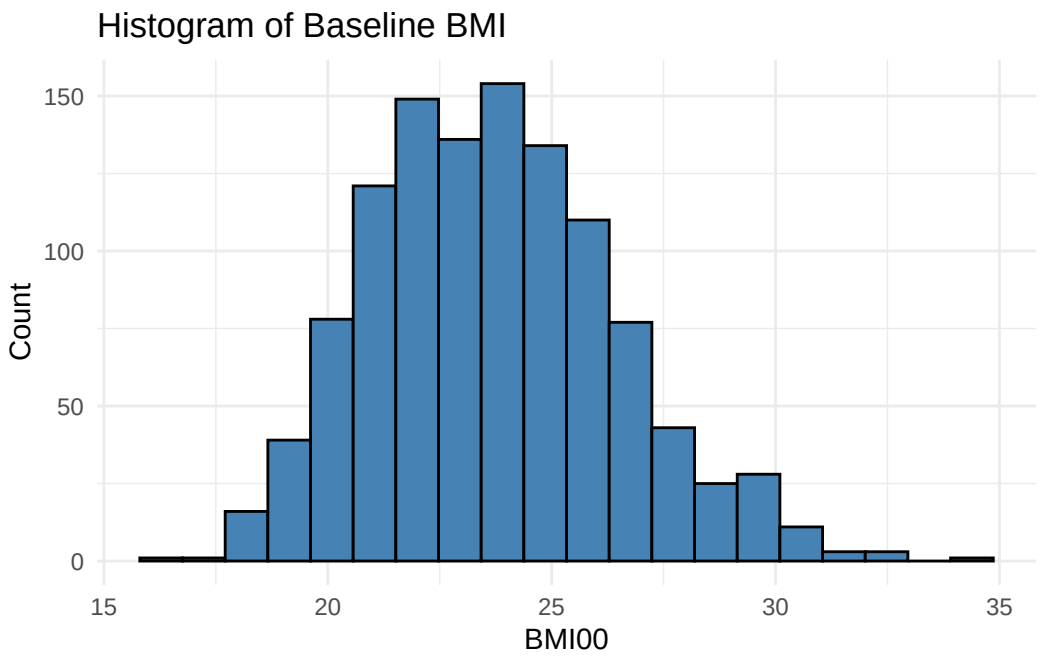

```r
ggplot(df, aes(y = BMI12)) +
  geom_boxplot(fill = "lightgreen") +
```

```
  theme_minimal() +
  labs(title = "Boxplot of BMI12", y = "BMI12")
```

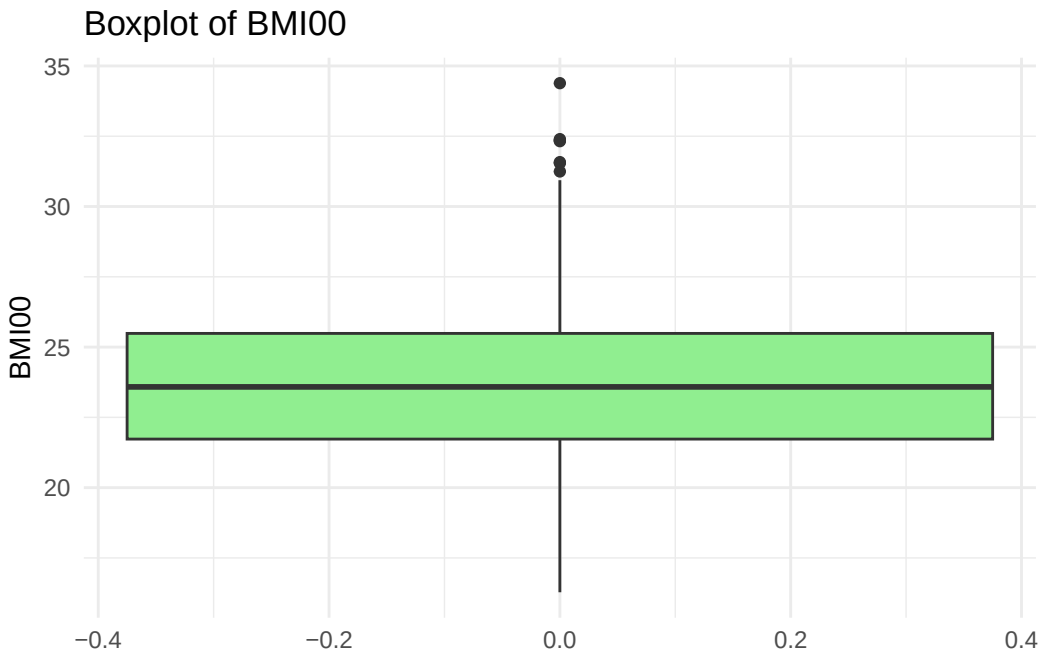
```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



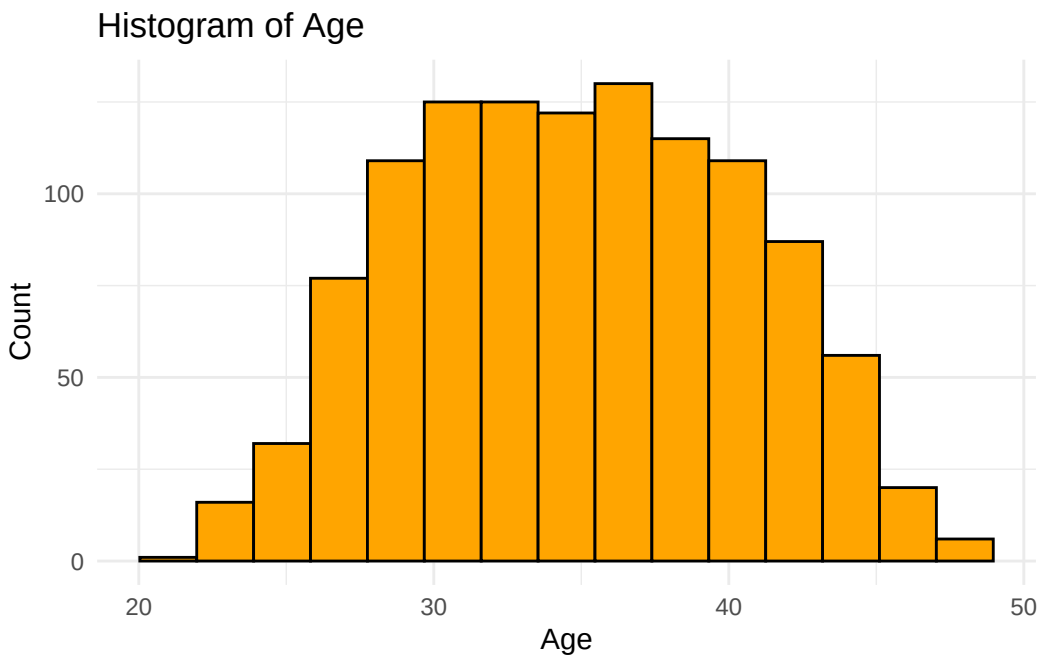

```
#BMI10

ggplot(df, aes(x = BMI00)) +
  geom_histogram(bins = 20, fill = "steelblue", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Baseline BMI", x = "BMI00", y = "Count")
```



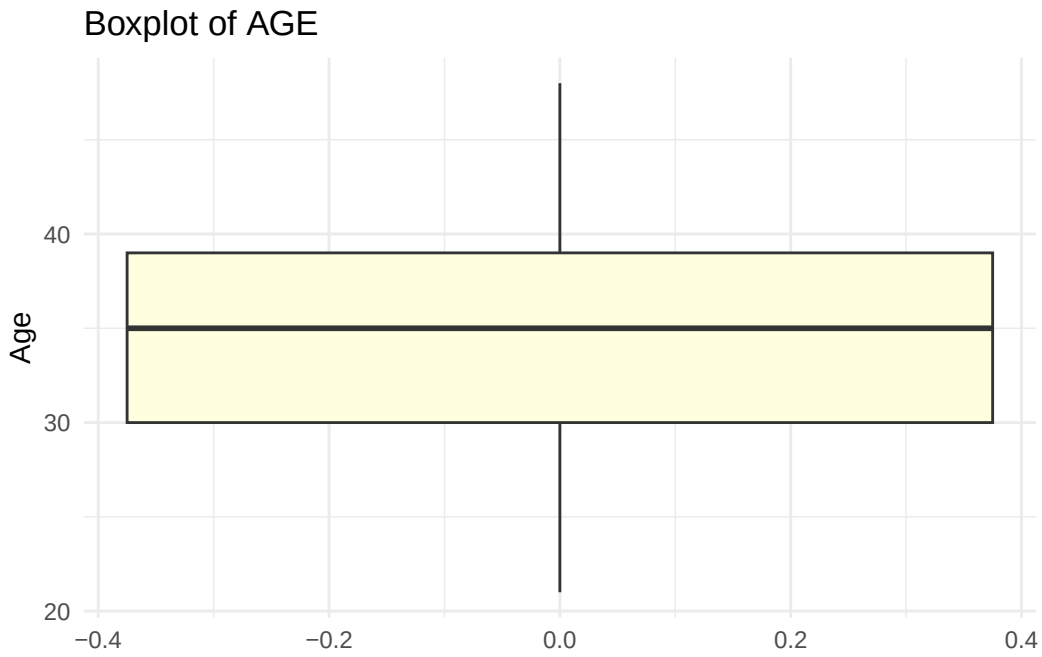

```
ggplot(df, aes(y = BMI00)) +
  geom_boxplot(fill = "lightgreen") +
  theme_minimal() +
  labs(title = "Boxplot of BMI00", y = "BMI00")
```




```
#Age
ggplot(df, aes(x = AGE)) +
  geom_histogram(bins = 15, fill = "orange", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Age", x = "Age", y = "Count")
```




```
ggplot(df, aes(y = AGE)) +
  geom_boxplot(fill = "lightyellow") +
  theme_minimal() +
  labs(title = "Boxplot of AGE", y = "Age")
```



AGE

For variable age, we can see that it has the mean of 34.86 and median of 34.83 its value range from 21.00 to 48.00.

From the histogram we can see that the graph is somewhat normal and uniform even though there is no peak,

BMI AT 12 MONTHS

We can see that there are some missing values in this variable as its seems to have lower number of observations compare to the other 2 variables. It have the mean ad median at 25.14 and 24.75 respectively. From the graph, we can see that its distribution had a little bit of a right-skewed tails not too much. Overall it seems to have a normal distribution with some outliners in the far right of the graph.

BASELINE BMI

From the table, we can see that people participate in this trial has a mean BMI of 23.72 and median of 23.58. Theirs BMI score range from 16.28 to 34.39. The distriubtion of baseline BMI is quite normally distributed.

From the boxplots, we can see the that the two whisker are somewhat symestric. However, we can see some outliers which can be worth investigate later in the analysis phrase

Distributions of GROUP and GENDER

```r
table(df$GROUP)
```

```
EXP STD
562 568
```

```r
prop.table(table(df$GROUP))
```

```
      EXP       STD
0.4973451 0.5026549
```

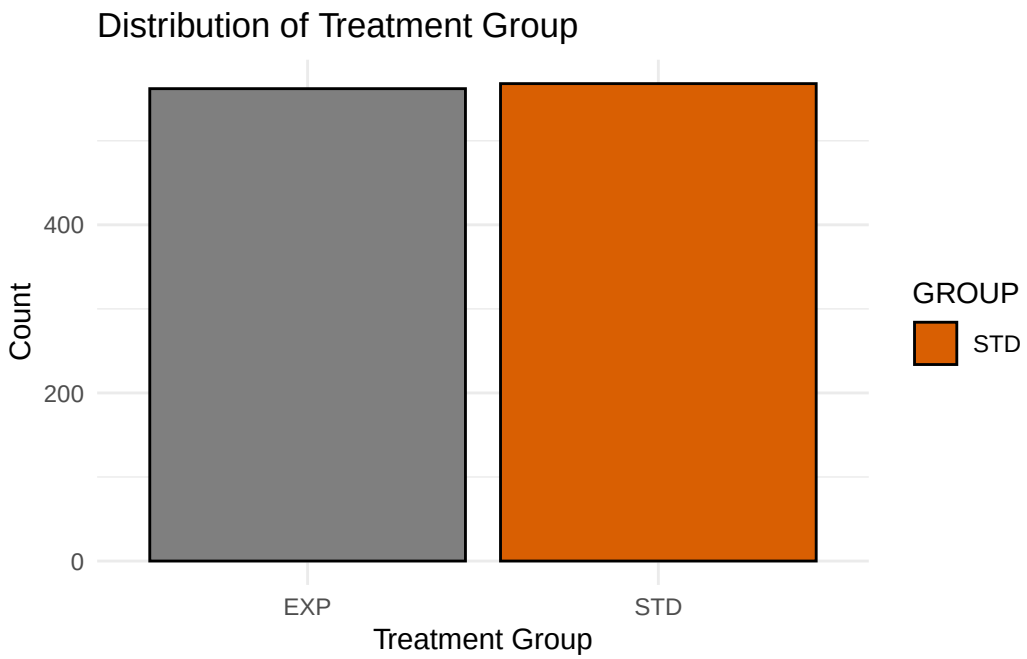
```r
table(df$GENDER)
```

```
  F   M
511 619
```

```r
prop.table(table(df$GENDER))
```

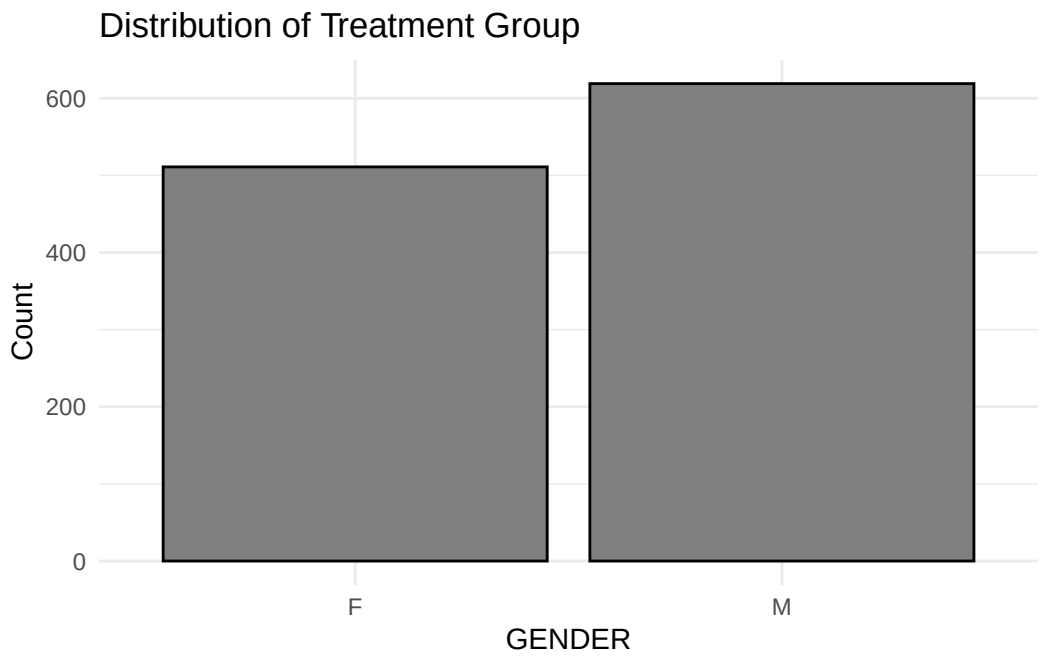
```
        F         M
0.4522124 0.5477876
```

```r
#FREQ TABLE

ggplot(df, aes(x = GROUP, fill = GROUP)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c("INT" = "#1b9e77", "STD" = "#d95f02")) +
  labs(title = "Distribution of Treatment Group", x = "Treatment Group", y = "Count") +
  theme_minimal()
```




```
ggplot(df, aes(x = GENDER, fill = GENDER)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c("INT" = "#1b9e77", "STD" = "#d95f02")) +
  labs(title = "Distribution of Treatment Group", x = "GENDER", y = "Count") +
  theme_minimal()
```

```
Warning: No shared levels found between `names(values)` of the manual scale and the
data's fill values.
No shared levels found between `names(values)` of the manual scale and the
data's fill values.
```



For Groups variable:
Based on the frequency table, we can tell that the disitribution within the groups is quite similar.

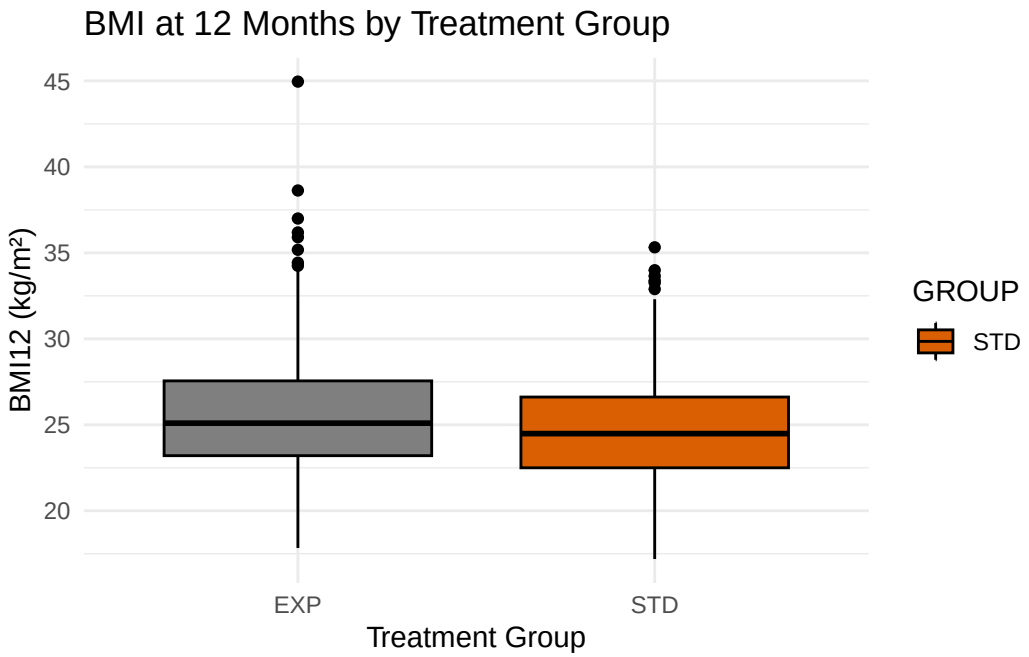
However, in Gender, we can tell that there are more male in this trial study compare to females participators.

Question 2: Bivariate Analysis

a. Examine BMI12 by GROUP

```
# Boxplot of BMI12 by treatment group
ggplot(df, aes(x = GROUP, y = BMI12, fill = GROUP)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("INT" = "#1b9e77", "STD" = "#d95f02")) +
  labs(title = "BMI at 12 Months by Treatment Group",
       x = "Treatment Group",
       y = "BMI12 (kg/m²)") +
  theme_minimal()
```

```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



The boxplot compares the distribution of BMI12 between the Intensive therapy (EXP) and Standard therapy (STD) groups. The median BMI12 is slightly higher in the EXP group than in the STD group, suggesting that participants receiving intensive therapy tend to have marginally higher BMI after one year.

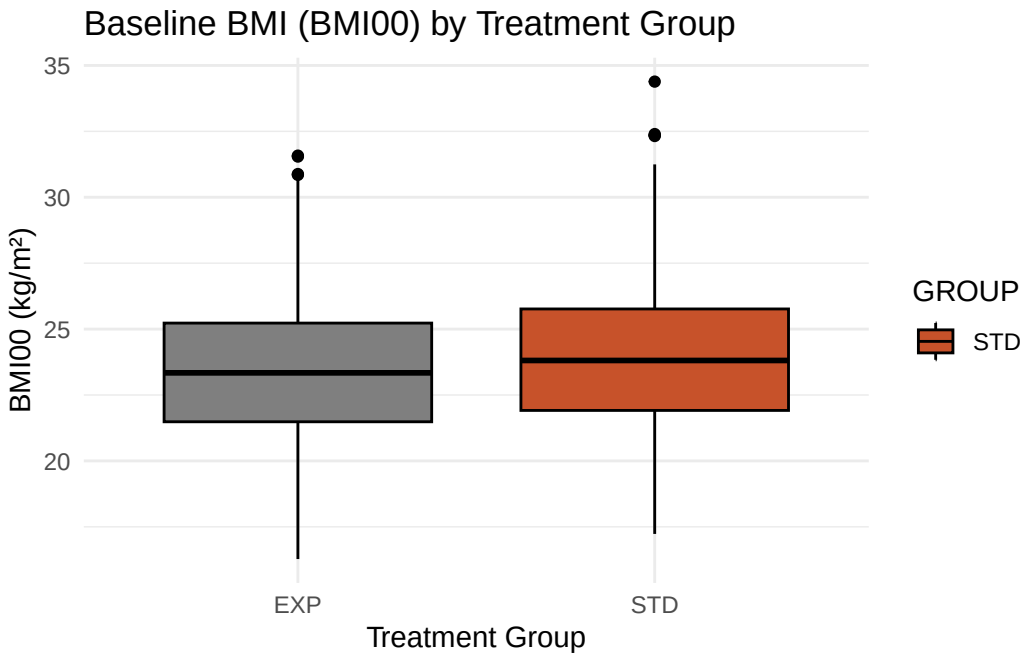
The interquartile ranges (IQRs) of the two groups overlap substantially, indicating that overall variability in BMI12 is similar between treatment cohorts. However, the EXP group exhibits more extreme high-end outliers, suggesting greater right-tail variability under intensive therapy.

In contrast, the STD group shows fewer and less extreme high BMI outliers.

Overall, while the central tendencies are fairly similar between groups, the EXP group displays greater dispersion and more extreme BMI values, suggesting that any treatment effect may be modest in magnitude and should be formally evaluated using regression modeling.

b. BMI100 by GROUP

```
ggplot(df, aes(x = GROUP, y = BMI00, fill = GROUP)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("INT" = "#76a21e", "STD" = "#c7522a")) +
  labs(title = "Baseline BMI (BMI00) by Treatment Group",
       x = "Treatment Group",
       y = "BMI00 (kg/m²)") +
  theme_minimal()
```



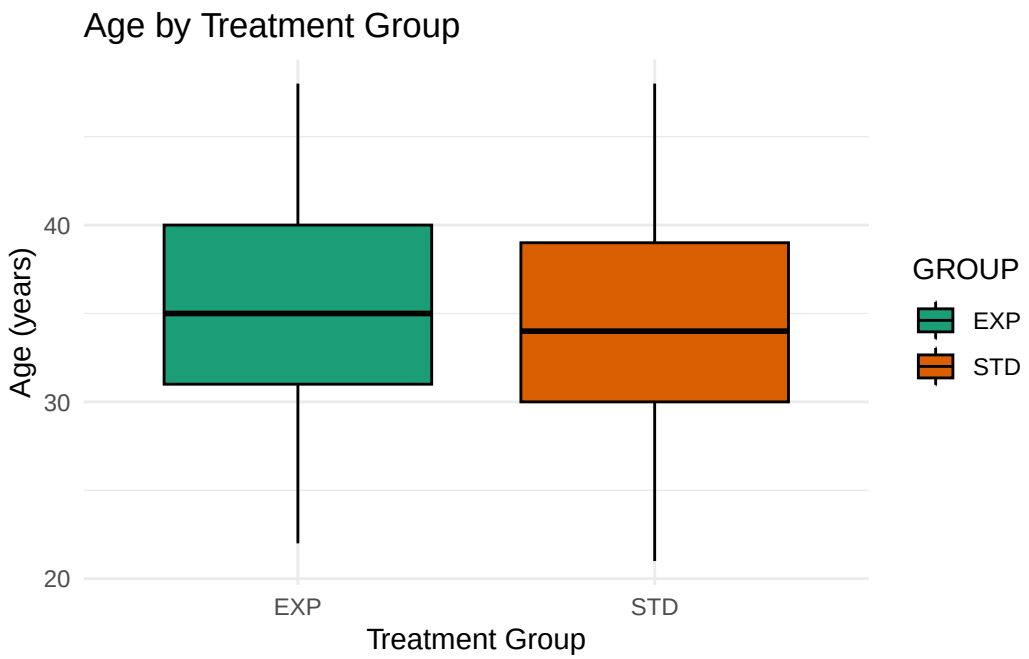
The boxplot of baseline BMI (BMI00) by treatment group shows that the EXP and STD groups are highly comparable at baseline. The median BMI00 values are nearly identical between the two cohorts, indicating strong balance in central tendency prior to treatment initiation.

The IQR overlap , and the overall spread of BMI values appears similar across groups. Both treatment groups exhibit a small number of upper-end outliers, but these are comparable in magnitude and frequency.

Overall, this distribution suggests that randomization was successful in producing well-balanced baseline BMI between treatment arms, minimizing concern for confounding due to pre-treatment BMI differences. This supports the validity of subsequent comparisons of BMI12 outcomes between groups.

c. Age by GROUP

```
ggplot(df, aes(x = GROUP, y = AGE, fill = GROUP)) +
  geom_boxplot(color = "black") +
  scale_fill_manual(values = c("EXP" = "#1b9e77", "STD" = "#d95f02")) +
  labs(
    title = "Age by Treatment Group",
    x = "Treatment Group",
    y = "Age (years)"
  ) +
  theme_minimal()
```



The boxplot shows that the age distributions for the intensive therapy and standard therapy groups are very similar. Both groups have comparable medians in the mid-30s, and the IQR largely overlap, indicating similar variability in age across treatment cohorts. The overall ranges of age also appear comparable, with no clear evidence of systematic age differences between groups.

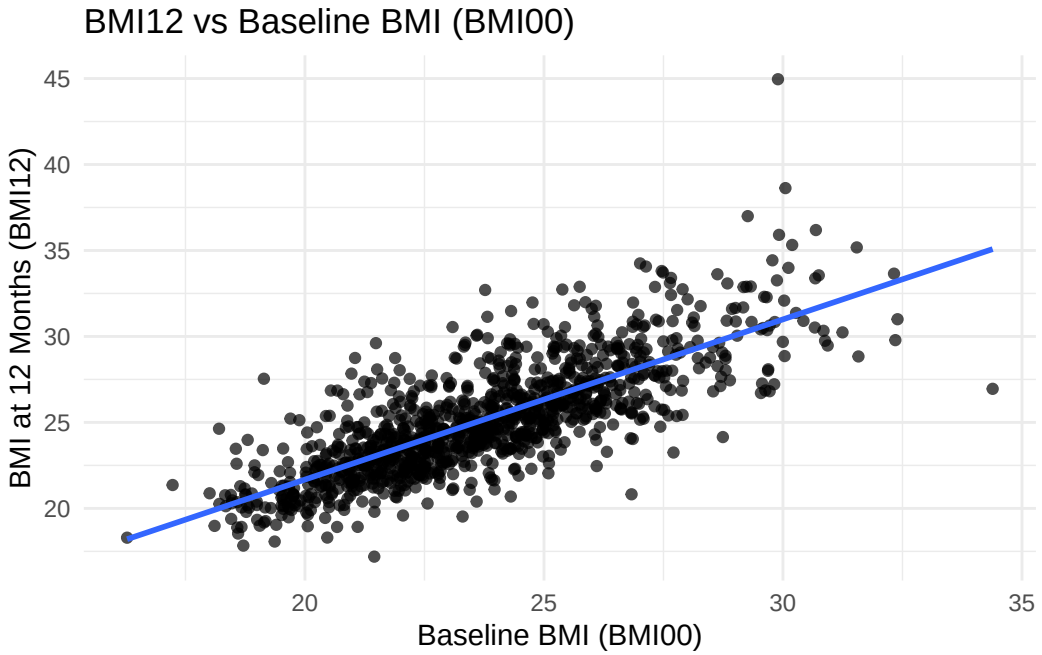
d. BMI12 versus BMI00 and BMI12 versus AGE

```
#BMI12 vs BMI00
ggplot(df, aes(x = BMI00, y = BMI12)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(
    title = "BMI12 vs Baseline BMI (BMI00)",
    x = "Baseline BMI (BMI00)",
    y = "BMI at 12 Months (BMI12)"
  )
```

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

```
Warning: Removed 8 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

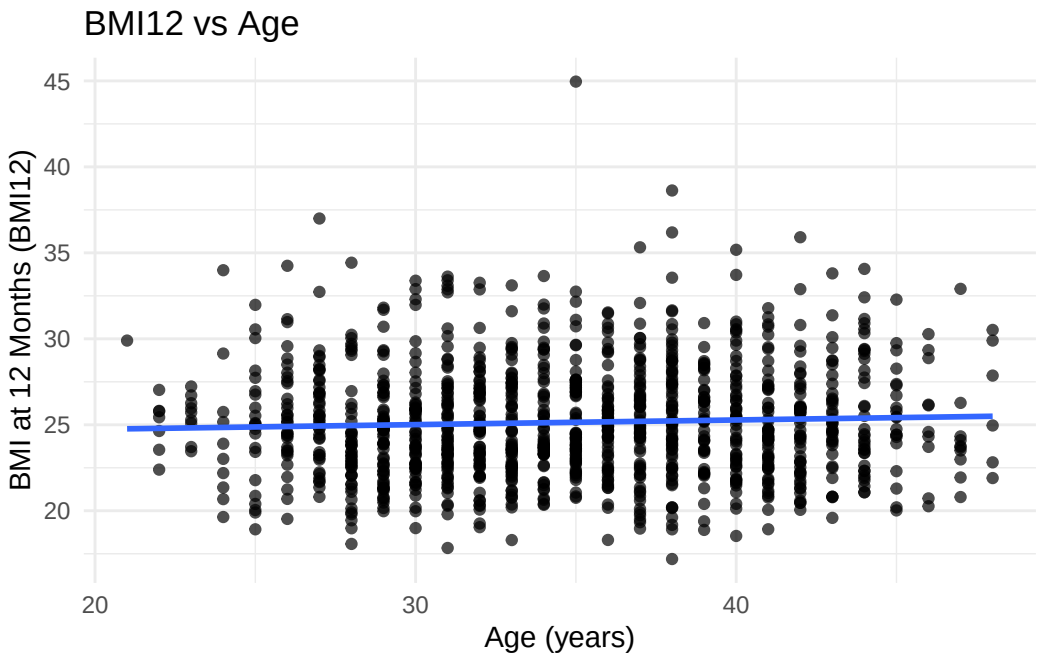
```
Warning: Removed 8 rows containing missing values or values outside the scale range
(`geom_point()`).
```




```r
#BMI12 vs AGE
ggplot(df, aes(x = AGE, y = BMI12)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal() +
  labs(
    title = "BMI12 vs Age",
    x = "Age (years)",
    y = "BMI at 12 Months (BMI12)"
  )
```

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

```
Warning: Removed 8 rows containing non-finite outside the scale range (`stat_smooth()`).
Removed 8 rows containing missing values or values outside the scale range
(`geom_point()`).
```



The scatterplot of BMI at 12 months (BMI12) versus baseline BMI (BMI00) shows a strong, positive, and approximately linear relationship. As baseline BMI increases, BMI12 increases in a nearly proportional manner.

The points are tightly clustered around the fitted regression line, indicating a strong association . This suggests that baseline BMI can be a dominant predictor of 12-month BMI and explains a substantial portion of the variation in BMI12.

There is no clear evidence of curvature or nonlinearity, and the spread of points remains relatively consistent across the range of BMI00 values, supporting the linearity and constant variance assumptions for regression.

A few high-BMI observations are present at the upper end, but they follow the same overall linear trend and do not appear to unduly distort the relationship.

Overall, this visualization provides strong justification for including baseline BMI (BMI00) as a key covariate in all subsequent regression models of BMI12.

#BMI12 vs Age

The scatterplot of BMI at 12 months (BMI12) versus age shows a very weak positive association. The fitted regression line has only a slight upward slope, indicating that BMI12 increases marginally with age.

However, there is substantial vertical spread in BMI12 at every age level, demonstrating considerable variability in BMI outcomes across individuals of the same age. This wide dispersion indicates that age explains only a small portion of the variability in BMI12.

No evidence of curvature or nonlinearity is apparent, and the relationship appears approximately linear, though weak.

Overall, this plot suggests that age is not a strong standalone predictor of BMI at 12 months, but it may still serve as an important adjustment variable in multivariable models to control for potential confounding.

Question 3

```
model_1 <- lm(BMI12 ~ GROUP + GENDER + AGE, data = df)
summary(model_1)
```

```
Call:
lm(formula = BMI12 ~ GROUP + GENDER + AGE, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3659 -2.2920 -0.3856  1.8889 19.0882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.62600    0.60901  40.436  < 2e-16 ***
GROUPSTD    -0.91081    0.19391  -4.697 2.97e-06 ***
GENDERM      0.59503    0.19474   3.055   0.0023 **
AGE          0.01858    0.01675   1.109   0.2676
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.236 on 1118 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.02839,   Adjusted R-squared:  0.02578
F-statistic: 10.89 on 3 and 1118 DF,  p-value: 4.719e-07
```

```
df_model <- model.frame(model_1)

df_model$stud_resid <- rstudent(model_1)
df_model$fitted_vals <- fitted(model_1)
```

#ASSUMPTION CHECKS USING STUDENTIZED RESIDUALS

Linearity and Constant Variance

```
library(lmtest)
```

```
Warning: package 'lmtest' was built under R version 4.5.1
```

```
Loading required package: zoo
```

```
Warning: package 'zoo' was built under R version 4.5.1
```

```

Attaching package: 'zoo'
```

```
The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
resettest(model_1, power = 2:3, type = "fitted")
```

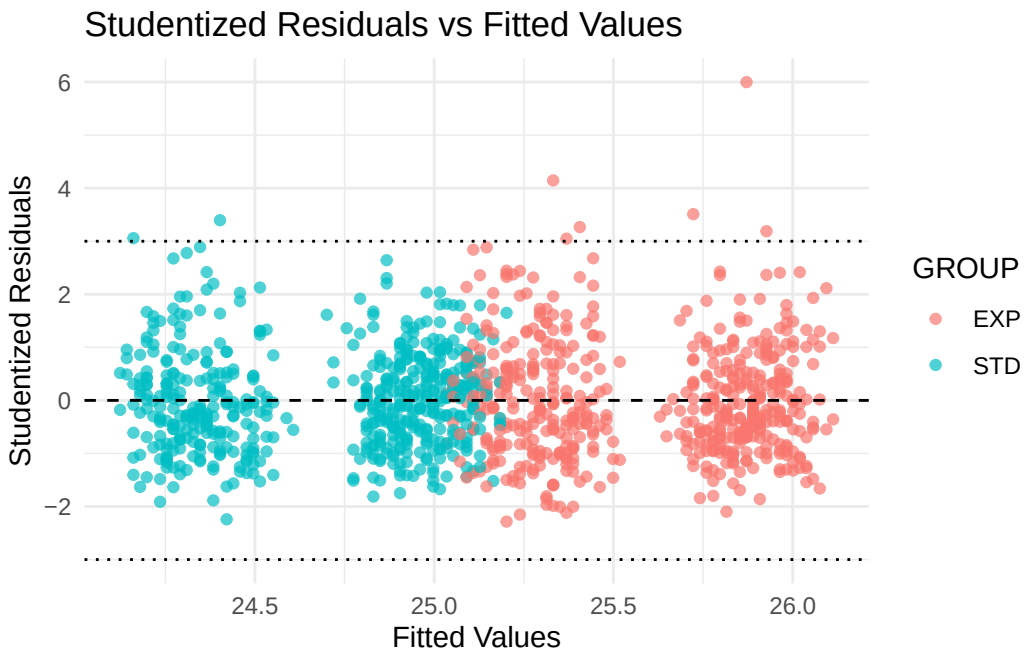
```

    RESET test

data:  model_1
RESET = 0.099544, df1 = 2, df2 = 1116, p-value = 0.9053
```

```
bptest(model_1)
```

```

    studentized Breusch-Pagan test

data:  model_1
BP = 16.845, df = 3, p-value = 0.0007606
```

```
ggplot(df_model, aes(x = fitted_vals, y = stud_resid,color=GROUP)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_hline(yintercept = c(-3, 3), linetype = "dotted") +
  theme_minimal() +
```

```
  labs(
    title = "Studentized Residuals vs Fitted Values",
    x = "Fitted Values",
    y = "Studentized Residuals"
  )
```



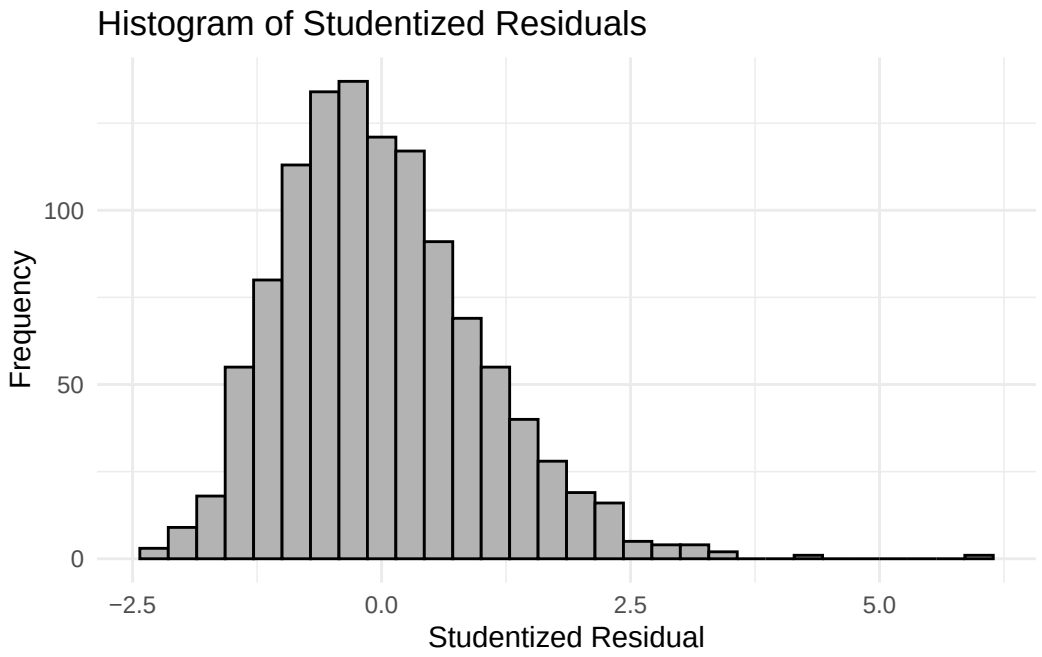
The apparent clustering of points in the residuals versus fitted values plot reflects the inclusion of treatment group as a categorical predictor. This pattern is expected and does not indicate nonlinearity. Importantly, within each cluster the residuals remain randomly scattered around zero with no visible curvature, and the Ramsey RESET test confirms correct functional form ($p = 0.9053$). Therefore, the linearity assumption is not violated.

Although the residual plot suggests only mild departures from constant variance, the significant BP test indicates that heteroscedasticity is present. Given the large sample size, this likely reflects a modest but statistically detectable variance structure rather than severe heteroscedasticity. As a result, standard errors based on the usual OLS assumptions may be slightly underestimated, and robust standard errors may be considered for inference.

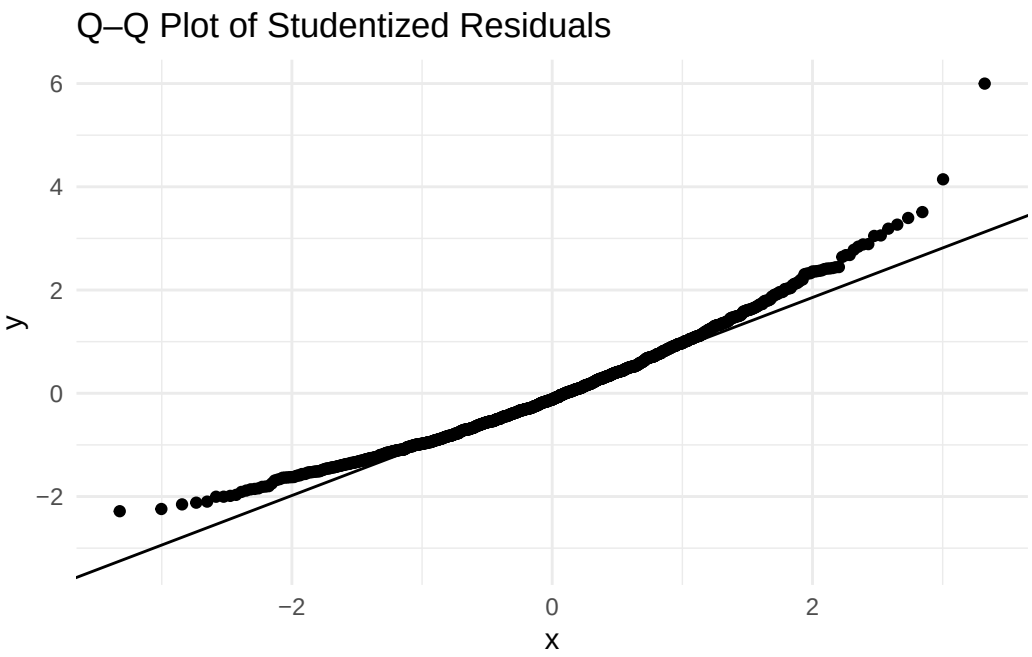
#Normality

```
ggplot(df_model, aes(x = stud_resid)) +
  geom_histogram(bins = 30, fill = "gray70", color = "black") +
  theme_minimal() +
  labs(
```

```
    title = "Histogram of Studentized Residuals",
    x = "Studentized Residual",
    y = "Frequency"
  )
```




```
ggplot(df_model, aes(sample = stud_resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_minimal() +
  labs(title = "Q-Q Plot of Studentized Residuals")
```




```
#Skewed
library(moments)
```

```
Warning: package 'moments' was built under R version 4.5.2
```

```
skewness(df_model$stud_resid)
```

```
[1] 0.7481051
```

```
kurtosis(df_model$stud_resid)
```

```
[1] 4.317217
```

```
shapiro.test(df_model$stud_resid)
```

```

    Shapiro-Wilk normality test

data:  df_model$stud_resid
W = 0.97183, p-value = 5.883e-14
```

The normality assumption was evaluated using a histogram, Q–Q plot, and the Shapiro–Wilk test applied to the studentized residuals. The histogram shows an approximately bell-shaped distribution centered near zero with a slight right-skew and a small number of large positive residuals.

The Q–Q plot indicates that the majority of points fall close to the theoretical reference line through the central portion of the distribution; however, noticeable deviations are present in the upper tail, where several large positive residuals depart substantially from normality.

Based on the Shapiro–Wilk test ($p < 0.0001$), the null hypothesis of normally distributed residuals is rejected, indicating a formal violation of the normality assumption. The histogram and Q–Q plot further reveal mild right-skewness and several large positive residuals.

#Multicolinarity

```
library(car)
```

```
Warning: package 'car' was built under R version 4.5.2
```

```
Loading required package: carData
```

```
Warning: package 'carData' was built under R version 4.5.1
```

```

Attaching package: 'car'
```

```
The following object is masked from 'package:psych':

    logit
```

```
The following object is masked from 'package:dplyr':

    recode
```

```
The following object is masked from 'package:purrr':

    some
```

```
vif(model_1)
```

```
   GROUP   GENDER      AGE
1.007369 1.005967 1.008740
```

Multicollinearity was assessed using variance inflation factors. All predictors show low VIF values well below the commonly used threshold of 5. This indicates no evidence of problematic multicollinearity among the predictors. Therefore, the estimated regression coefficients are stable and interpretable.

Independence of error

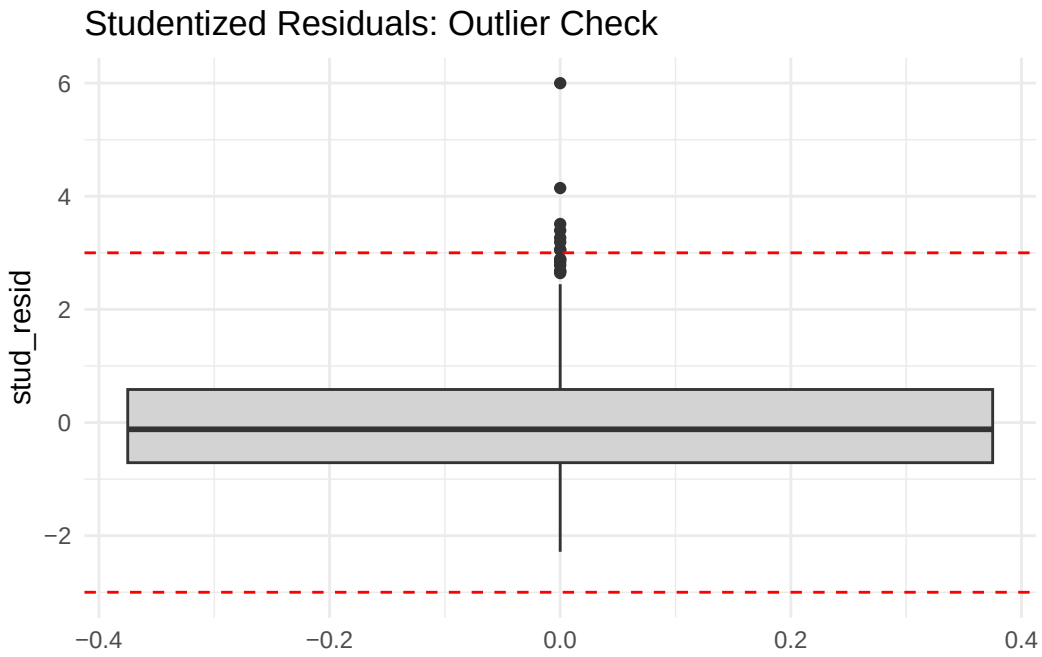
Independence of errors is assumed based on the randomized, subject-level structure of the clinical trial, with no repeated measurements per participant.

Question 4 Influence and outliers

```r
# Identify large residuals
which(abs(df_model$stud_resid) > 3)
```

```
[1]  64 191 243 281 414 416 535 908
```

```r
ggplot(df_model, aes(y = stud_resid)) +
  geom_boxplot(fill = "lightgray") +
  geom_hline(yintercept = c(-3, 3), linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(title = "Studentized Residuals: Outlier Check")
```



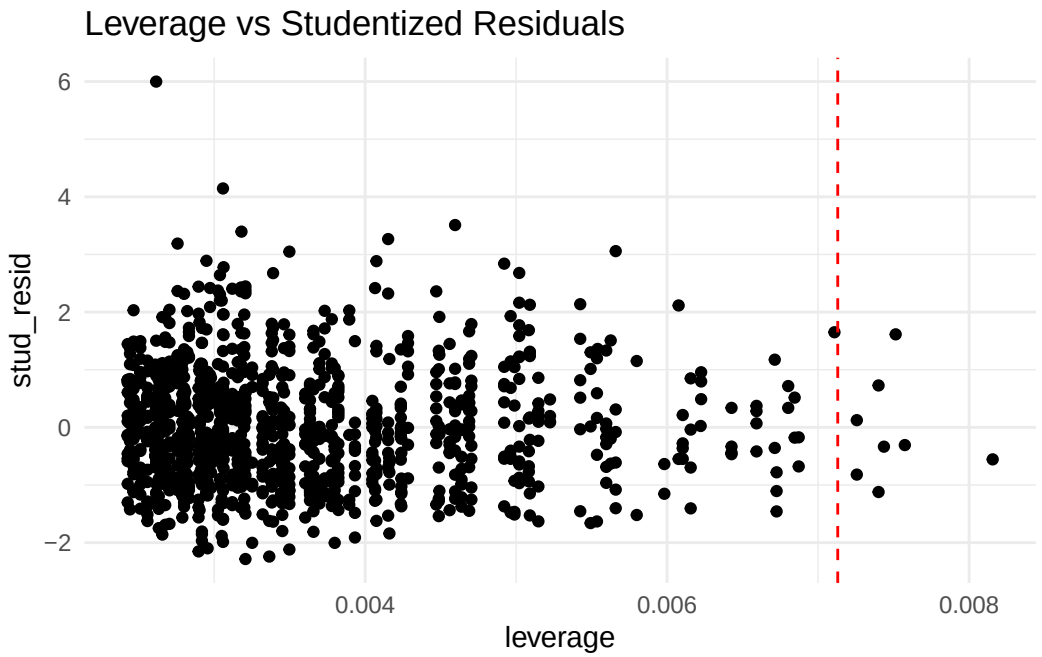

```
df_model$leverage <- hatvalues(model_1)

p <- length(coef(model_1)) - 1
n <- nrow(df_model)

lev_cutoff <- 2*(p+1)/n
lev_cutoff
```

```
[1] 0.007130125
```

```
which(df_model$leverage > lev_cutoff)
```

```
[1]  54 244 246 252 285 762 764 814
```

```
ggplot(df_model, aes(x = leverage, y = stud_resid)) +
  geom_point() +
  geom_vline(xintercept = lev_cutoff, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(title = "Leverage vs Studentized Residuals")
```



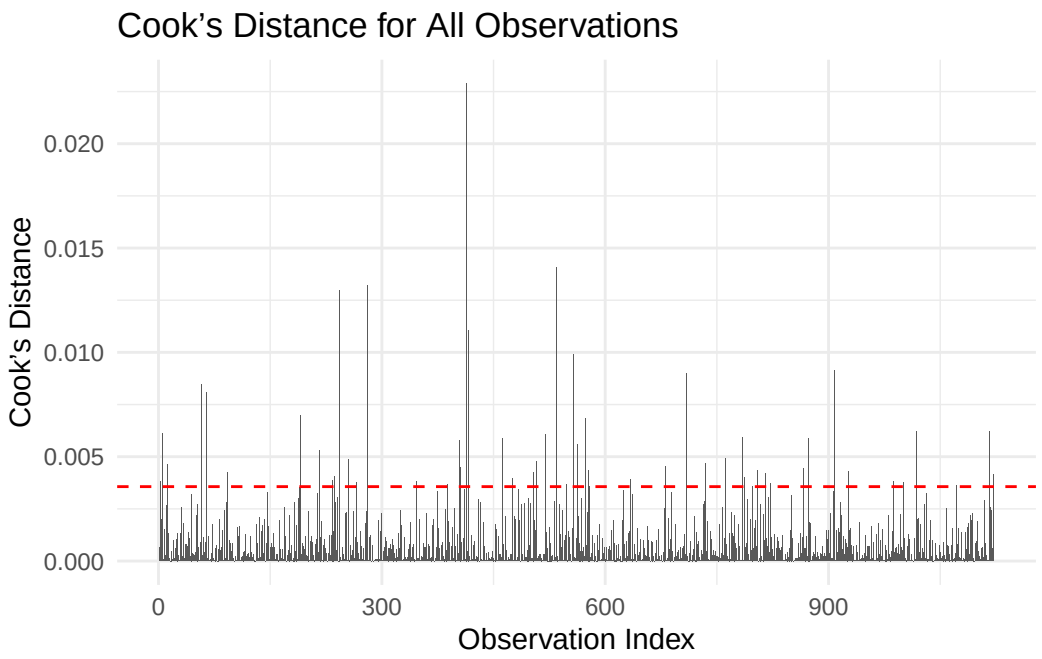

```
df_model$cooks <- cooks.distance(model_1)
cook_cutoff <- 4/n
cook_cutoff
```

```
[1] 0.003565062
```

```
which(df_model$cooks > cook_cutoff)
```

```
 [1]    3    6   12   58   64   92  191  216  233  237  243  255  266  281  347
[16]  388  404  406  414  416  462  476  504  508  520  535  548  558  563  574
[31]  577  579  634  681  709  735  762  785  787  798  805  815  822  866  873
[46]  908  927  987 1000 1018 1072 1116 1122
```

```
ggplot(df_model, aes(x = seq_along(cooks), y = cooks)) +
  geom_bar(stat = "identity") +
  geom_hline(yintercept = cook_cutoff, linetype = "dashed", color = "red") +
  theme_minimal() +
  labs(title = "Cook's Distance for All Observations",
       x = "Observation Index",
       y = "Cook's Distance")
```



Several observations were flagged as potential outliers based on large studentized residuals (> 3), indicating unusually high or low BMI12 values relative to model predictions. A small number of observations also showed high leverage, suggesting uncommon combinations of predictor values. Cook's distance identified a subset of observations as moderately influential.

However, all Cook's distance values remained well below 1, indicating no single observation show high influence on the model. Together, although a few points are statistically unusual, they do not appear to meaningfully distort the regression results, and therefore, all observations were retained in the final analysis.

Part 2

Question 1

Yes, the initial regression model can be interpreted, but with some caution. The linearity assumption was satisfied, and independence of observations is supported by the randomized study design. No serious multicollinearity issues were detected, and no individual observation was found to have undue influence on the model.

However, formal tests indicated mild violations of the constant variance and normality assumptions. Because the sample size is large, these tests are very sensitive and likely detected small departures rather than major problems. Visual diagnostics suggest that these violations are not severe.

Therefore, the initial model is appropriate for a preliminary interpretation of the treatment effect, but results should be viewed cautiously and improved in later models.

Question 2

```
summary(model_1)
```

```
Call:
lm(formula = BMI12 ~ GROUP + GENDER + AGE, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3659 -2.2920 -0.3856  1.8889 19.0882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.62600    0.60901  40.436  < 2e-16 ***
GROUPSTD    -0.91081    0.19391  -4.697 2.97e-06 ***
GENDERM      0.59503    0.19474   3.055   0.0023 **
AGE          0.01858    0.01675   1.109   0.2676
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.236 on 1118 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.02839,   Adjusted R-squared:  0.02578
F-statistic: 10.89 on 3 and 1118 DF,  p-value: 4.719e-07
```

The overall multiple linear regression model predicting BMI at 12 months (BMI12) from treatment group (GROUP), gender, and age was evaluated using an omnibus F-test. The results indicate that the model is statistically significant,

$F(3,1118)=10.89, p=4.72\times10-7$

This result indicates that, taken together, GROUP, GENDER, and AGE explain a statistically significant portion of the variability in BMI at 12 months. The model explains approximately 2.84% of the variability in BMI12, indicating a small overall effect size.

With respect to individual predictors, treatment group is statistically significant. Participants in the standard therapy (STD) group have, on average, a BMI that is 0.91 units lower than those

in the intensive therapy group, after adjusting for gender and age. Gender is also statistically significant, with males having, on average, a BMI that is 0.60 units higher than females. Age is not a statistically significant predictor of BMI at 12 months ($= 0.0186$, $p = 0.268$).

Overall, while the model is statistically significant, the proportion of variance explained is small, indicating that additional predictors such as baseline BMI are needed to better explain BMI outcomes at 12 months.

Part 3

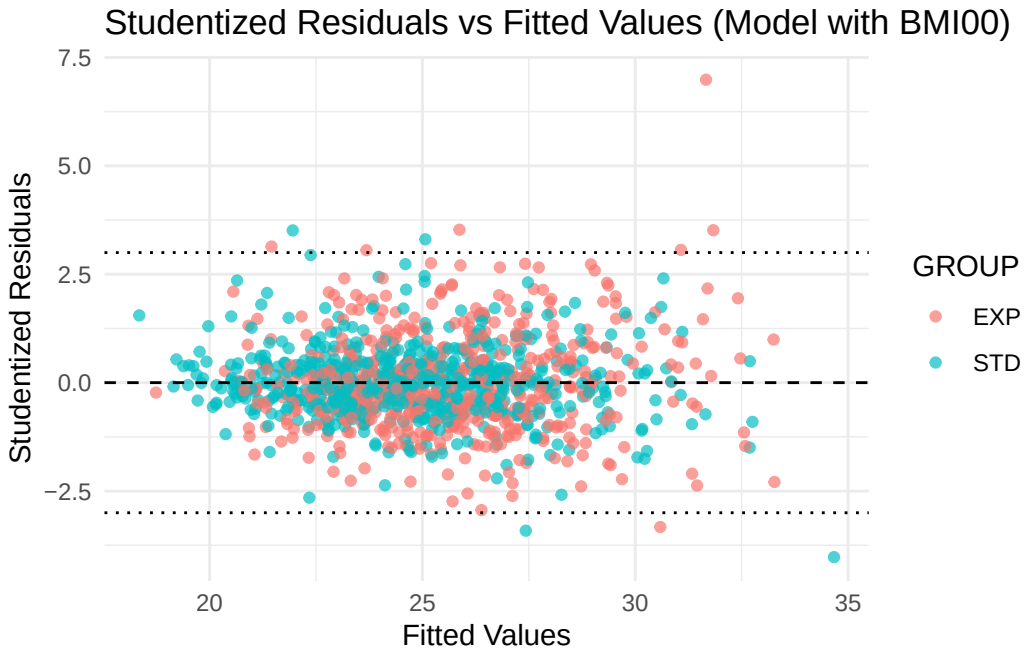
Question 1

```r
model_2 <- lm(BMI12 ~ GROUP + GENDER + AGE + BMI00, data = df)
summary(model_2)
```

```
Call:
lm(formula = BMI12 ~ GROUP + GENDER + AGE + BMI00, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7167 -1.1457 -0.1288  0.9257 13.2966

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.307428   0.605540   5.462  5.8e-08 ***
GROUPSTD    -1.307897   0.117235 -11.156  < 2e-16 ***
GENDERM     -0.027362   0.118231  -0.231    0.817
AGE         -0.001557   0.010106  -0.154    0.878
BMI00        0.951224   0.021487  44.269  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.95 on 1117 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.6473,    Adjusted R-squared:  0.646
F-statistic: 512.4 on 4 and 1117 DF,  p-value: < 2.2e-16
```

```r
df_model2 <- model.frame(model_2)

df_model2$stud_resid  <- rstudent(model_2)
df_model2$fitted_vals <- fitted(model_2)
```

```r
ggplot(df_model2, aes(x = fitted_vals, y = stud_resid, color = GROUP)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_hline(yintercept = c(-3, 3), linetype = "dotted") +
  theme_minimal() +
  labs(
    title = "Studentized Residuals vs Fitted Values (Model with BMI00)",
    x = "Fitted Values",
    y = "Studentized Residuals"
  )
```



Adding baseline BMI improved the model. It greatly improved explanatory power, with the R^2 increasing from approximately 0.03 in the initial model to 0.65 in the revised model. The residual standard error decreased from about 3.24 to 1.95, demonstrating a reduction in unexplained variability and improved prediction accuracy.

After adding baseline BMI, the treatment effect became stronger and more precisely estimated, with participants in the standard therapy group having, on average, a BMI that is approximately

1.31 units lower than those in the intensive therapy group at 12 months, after adjustment. Overall, the revised model provides a far more accurate, stable, and clinically meaningful assessment of treatment differences in BMI at 12 months.

Question 2

The revised regression model was highly statistically significant, $F(4,1117)=512.4$,$p<0.0000000000000022$, and explained approximately 64.7% of the variability in BMI12. This demonstrates that the predictors jointly provide a strong explanation of BMI at 12 months.

Question 3

In the revised multiple linear regression model predicting BMI at 12 months (BMI12), treatment group (GROUP) and baseline BMI (BMI00) were statistically significant predictors, while gender and age were not significant after adjustment.

For treatment group,after adjusting for gender, age, and baseline BMI, participants in the standard therapy group have, on average, a BMI that is 1.31 kg/m^2 lower at 12 months compared to participants in the intensive therapy group. This difference is highly statistically significant, providing strong evidence that treatment cohort is associated with BMI outcomes after one year.

For Baseline BMI, holding treatment group, gender, and age constant, a one-unit increase in baseline BMI is associated with an average increase of 0.95 kg/m^2 in BMI at 12 months. This is strong and statistically significant association that confirms that baseline BMI is the dominant predictor of BMI at one year.

Part 4

Question 1

```
model_3 <- lm(BMI12 ~ GROUP + GENDER + AGE + BMI00 + BMI00*GROUP, data = df)
summary(model_3)
```

```
Call:
lm(formula = BMI12 ~ GROUP + GENDER + AGE + BMI00 + BMI00 * GROUP,
    data = df)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-7.0596 -1.1297 -0.1382  0.9383 12.8429

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.647827   0.808762   2.037  0.04184 *
GROUPSTD        1.803050   1.016510   1.774  0.07637 .
GENDERM        -0.009662   0.117924  -0.082  0.93471
AGE            -0.001202   0.010068  -0.119  0.90501
BMI00           1.020906   0.031142  32.783  < 2e-16 ***
GROUPSTD:BMI00 -0.131258   0.042605  -3.081  0.00211 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.943 on 1116 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.6502,    Adjusted R-squared:  0.6487
F-statistic: 414.9 on 5 and 1116 DF,  p-value: < 2.2e-16
```

The multiple linear regression model including the interaction between baseline BMI (BMI00) and treatment group, along with gender and age, was highly statistically significant,

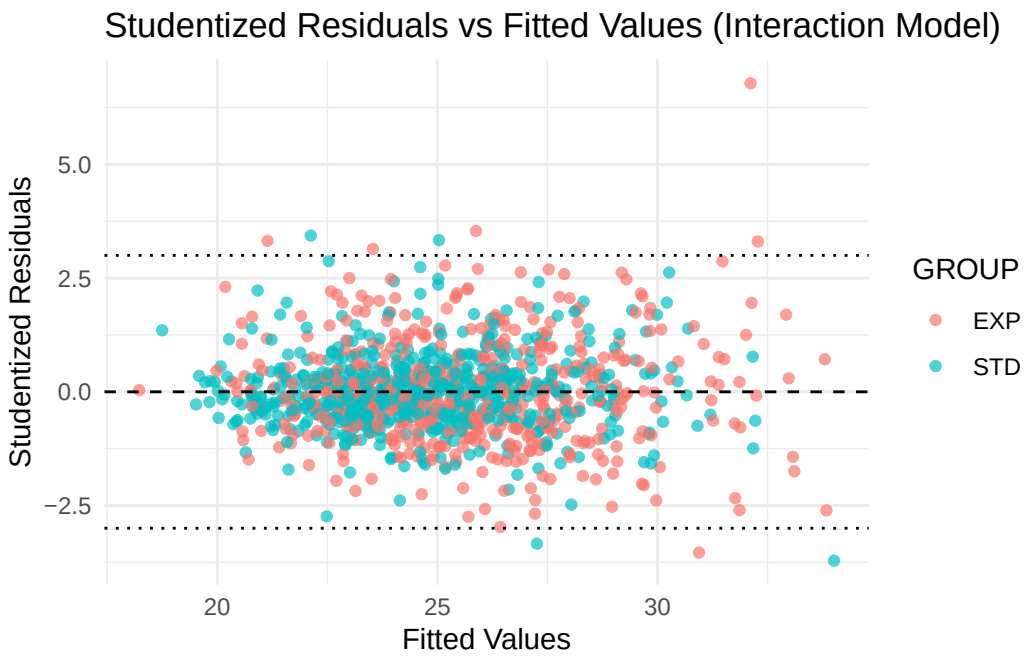
$F(5,1116)=414.9, p<0.0000001$

$R2=0.6487$, indicating that the model explains approximately 65% of the variability in BMI at 12 months.

Question 2

```r
library(car)
library(lmtest)

df_model3 <- model.frame(model_3)

df_model3$stud_resid  <- rstudent(model_3)
df_model3$fitted_vals <- fitted(model_3)

# residual scatterplot (color coded by cohort)
ggplot(df_model3, aes(x = fitted_vals, y = stud_resid, color = GROUP)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 0, linetype = "dashed") +
```

```
  geom_hline(yintercept = c(-3, 3), linetype = "dotted") +
  theme_minimal() +
  labs(
    title = "Studentized Residuals vs Fitted Values (Interaction Model)",
    x = "Fitted Values",
    y = "Studentized Residuals"
  )
```




```
bptest(model_3)
```

```
    studentized Breusch-Pagan test

data:  model_3
BP = 79.345, df = 5, p-value = 1.151e-15
```

```
shapiro.test(df_model3$stud_resid)
```

```
    Shapiro-Wilk normality test
```

```
data:  df_model3$stud_resid
W = 0.97072, p-value = 2.829e-14
```

```r
vif(model_3)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

```
      GROUP      GENDER         AGE       BMI00 GROUP:BMI00
  76.759571    1.022823    1.010919    2.165978   79.390645
```

High variance inflation factors were observed for GROUP and the GROUP × BMI00 interaction due to structural collinearity introduced by the interaction term. This is a known and expected consequence of including interactions and does not represent a data-related multicollinearity issue. Predictor-level VIFs or centering of BMI00 may be used to mitigate this effect.

```r
df$BMI00_c <- df$BMI00 - mean(df$BMI00, na.rm = TRUE)
model_3c <- lm(BMI12 ~ GROUP + GENDER + AGE + BMI00_c + BMI00_c:GROUP, data = df)
vif(model_3c)
```

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

```
        GROUP        GENDER           AGE       BMI00_c GROUP:BMI00_c
     1.013356      1.022823      1.010919      2.165978      2.158276
```

Baseline BMI was mean-centered to reduce structural multicollinearity introduced by the interaction term. After centering, variance inflation factors were all near 1, confirming that the data do not show multicollinearity.

After mean-centering baseline BMI to reduce structural collinearity in the interaction model, all variance inflation factors were below 2.2, indicating no evidence of multicollinearity among the predictors.

Question 3

```
#Partial f-test
anova(model_2, model_3c)
```

```
Analysis of Variance Table

Model 1: BMI12 ~ GROUP + GENDER + AGE + BMI00
Model 2: BMI12 ~ GROUP + GENDER + AGE + BMI00_c + BMI00_c:GROUP
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1   1117 4249.4
2   1116 4213.6  1    35.836 9.4915 0.002115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null Hypothesis:

H0: $\text{Beta}_{\text{BMI00xGROUP}} = 0$

Alternate Hypothesis:

Ha: $\text{Beta}_{\text{BMI00xGROUP}} =/= 0$

The partial F-test yield the result F(1,1116)=9.49, p=0.0021.

Because the p-value is less than 0.05, we reject the null hypothesis. This provides strong statistical evidence that the BMI00 × GROUP interaction significantly improves model fit. Therefore, the relationship between baseline BMI and BMI at 12 months differs significantly between the intensive therapy and standard therapy groups, and the interaction term should be retained in the final model.

Question 4

```
summary(model_3c)
```

```
Call:
lm(formula = BMI12 ~ GROUP + GENDER + AGE + BMI00_c + BMI00_c:GROUP,
    data = df)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-7.0596 -1.1297 -0.1382  0.9383 12.8429

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      25.864988   0.366805  70.514  < 2e-16 ***
GROUPSTD         -1.310553   0.116796 -11.221  < 2e-16 ***
GENDERM          -0.009662   0.117924  -0.082  0.93471
AGE              -0.001202   0.010068  -0.119  0.90501
BMI00_c           1.020906   0.031142  32.783  < 2e-16 ***
GROUPSTD:BMI00_c -0.131258   0.042605  -3.081  0.00211 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.943 on 1116 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.6502,    Adjusted R-squared:  0.6487
F-statistic: 414.9 on 5 and 1116 DF,  p-value: < 2.2e-16
```

The model was found to be highly statistically significant with F(5,1116) and p-value $<$ 0.0000001.

This result indicates that, taken together, treatment group, gender, age, centered baseline BMI, and the interaction between baseline BMI and treatment group explain a statistically significant portion of the variability in BMI at 12 months. The model explains approximately 65.0% of the total variability in BMI12 with R2 = 0.65.

Overall, these results confirm that the final interaction model provides a substantially improved and clinically meaningful explanation of BMI outcomes at 12 months compared to earlier models, and strongly support the inclusion of the BMI00 $\times$ GROUP interaction in the final analysis.

Question 5

The final regression model includes treatment group, gender, age, mean-centered baseline BMI, and the BMI00 $\times$ GROUP interaction. The statistically significant coefficients are GROUP, BMI00, and the interaction term.

For Baseline BMI

Holding treatment group, gender, and age constant, a 1-unit increase in baseline BMI is associated with an average increase of 1.02 kg/m^2 in BMI at 12 months among participants in the intensive therapy group.

Treatment Group

At the mean baseline BMI, participants in the standard therapy (STD) group have, on average, a BMI at 12 months that is 1.31 kg/m^2 lower than participants in the intensive therapy (INT) group, after adjusting for gender and age.

Interaction factors

The effect of baseline BMI on BMI at 12 months is 0.13 kg/m^2 weaker in the standard therapy group than in the intensive therapy group.

Question 6

The interaction between baseline BMI and treatment group indicates that patients' starting body weight influences their 12-month BMI differently depending on which therapy they receive. Specifically, patients who begin the study with a higher baseline BMI tend to retain more of that excess weight over time if they are treated with intensive therapy, compared with those receiving standard therapy.

In contrast, among patients in the standard therapy group, increases in baseline BMI have a weaker carry-over effect on BMI at 12 months. This suggests that standard therapy may help dampen the long-term impact of starting at a higher body weight, while intensive therapy appears to allow baseline BMI to persist more strongly into long-term outcomes.

Clinically, this means that patients with higher baseline BMI may respond differently to intensive versus standard therapy, and treatment decisions may benefit from considering a patient's starting BMI when choosing between therapy approaches.