AGLM_HW1

October 4, 2021

1 PHP2514: Applied Generalized Linear Models

1.1 Homework 1

Antonella Basso

1.1.1 Question 1:

The dataset "seeds.csv" contains data collected using a completely randomized experimental design. In particular, genetically similar seeds are randomly assigned to be raised in either a nutritionally enriched environment (treatment group) or standard conditions(control group). After a predetermined timepointall plants are harvested, dried and weighed.The dataset "seeds.csv" contains the weights (in grams)for 30 plants in each group.

a) Conduct a comprehensive Exploratory Data Analysis (EDA) to inspect, understand and describe the information collected in this dataset. Use appropriate summary statistics and plots to present your results from the EDA.

b) What do you think about the effectiveness of the two treatment groups based on the results of the EDA?

c) Use a suitable statistical procedure to compare the two groups. Express the question in terms of a hypothesis test. Explain the steps you followed to perform the statistical test and the results. What is your final conclusion?

[1]:
```
#importing "Seeds" data
seeds <- read.csv("/home/jovyan/AGLM/Seeds.csv")
seeds
```

A data.frame: 30 × 2

treatment <dbl>	control <dbl>
4.81	4.17
4.17	3.05
4.41	5.18
3.59	4.01
5.87	6.11
3.83	4.10
6.03	5.17
4.98	3.57
4.90	5.33
5.75	5.59
5.36	4.66
3.48	5.58
4.69	3.66
4.44	4.50
4.89	3.90
4.71	4.61
5.48	5.62
4.32	4.53
5.15	6.05
6.34	5.14
5.35	4.88
4.71	4.95
5.28	4.33
5.99	3.93
6.13	5.87
4.15	4.82
3.99	4.27
3.64	4.03
4.49	4.87
4.86	4.02

a) Exploratory Data Analysis (EDA) The variables in this dataset are as follows: - Outcome - (Y: weight) - Covariate - (X: treatment) -> 0: control, 1: treatment

The primary outcome of interest is a continuous random variable with ratio scale, while the predictor variable (covariate) is categorical with nominal scale and has two groups (binary); control and treatment (or no treatment and treatment).

This EDA consists of: - Descriptive Statistics - Boxplots - Histograms - Q-Q Plot

[2]:
```
#DATA WRANGLING

#installing tidyverse
#http://statseducation.com/Introduction-to-R/modules/tidy%20data/gather/
install.packages("tidyverse")
library(tidyverse)
```

```
#installing package to combine histograms
install.packages("gridExtra")
library(gridExtra)
```

```
Updating HTML index of packages in '.Library'

Making 'packages.html' …
 done

Warning message in system("timedatectl", intern = TRUE):
"running command 'timedatectl' had status 1"
  Attaching packages                        tidyverse
1.3.1

 ggplot2 3.3.5      purrr   0.3.4
 tibble  3.1.4      dplyr   1.0.7
 tidyr   1.1.3      stringr 1.4.0
 readr   1.4.0      forcats 0.5.1

  Conflicts
tidyverse_conflicts()
 dplyr::filter() masks stats::filter()
 dplyr::lag()    masks stats::lag()

Updating HTML index of packages in '.Library'

Making 'packages.html' …
 done


Attaching package: ‘gridExtra’


The following object is masked from ‘package:dplyr’:

    combine
```

[3]:
```
#arranging the data such that each line is a single observation (tidy format)
seeds_tidy <- seeds %>%
    gather('group', 'weight', 1:2)

head(seeds_tidy)
```

A data.frame: 6 × 2

	group <chr>	weight <dbl>
1	treatment	4.81
2	treatment	4.17
3	treatment	4.41
4	treatment	3.59
5	treatment	5.87
6	treatment	3.83

[4]:
```
#ceating a dummy variable for binary predictor (treatment)
#this adds a column that gives a 1 if the group is the treatment group and 0␣
 ↪otherwise (control)
#NOTE: 1*TRUE=1, 1*FALSE=0
seeds_tidy$treatment <- as.factor(1*(seeds_tidy$group=='treatment'))

#creating an index column for graphs/plots
seeds_tidy$indx <- row.names(seeds_tidy)

head(seeds_tidy)
```

A data.frame: 6 × 4

	group <chr>	weight <dbl>	treatment <fct>	indx <chr>
1	treatment	4.81	1	1
2	treatment	4.17	1	2
3	treatment	4.41	1	3
4	treatment	3.59	1	4
5	treatment	5.87	1	5
6	treatment	3.83	1	6

Descriptive Statistics:

- Summary of weight (minimum value, 1st quartile, median, mean, 3rd quartile, maximum value) for the whole data and for each treatment group
- Standard deviation (SD) and variance of weight data for each treatment group

[5]:
```
#summary of total weight and weight by group (control and treatment)
summary(seeds_tidy$weight)
by(seeds_tidy$weight, seeds_tidy$group, summary, na.rm=TRUE)

#SD and variance of weight for control and treatment groups
sd(seeds_tidy[seeds_tidy$group == "control",]$weight)
var(seeds_tidy[seeds_tidy$group == "control",]$weight)
sd(seeds_tidy[seeds_tidy$group == "treatment",]$weight)
var(seeds_tidy[seeds_tidy$group == "treatment",]$weight)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.050   4.165   4.760   4.771   5.335   6.340
```

```
seeds_tidy$group: control
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.050   4.048   4.635   4.683   5.178   6.110
------------------------------------------------------------
seeds_tidy$group: treatment
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.480   4.343   4.835   4.860   5.357   6.340
```

0.776967284330261

0.60367816091954

0.79003921043388

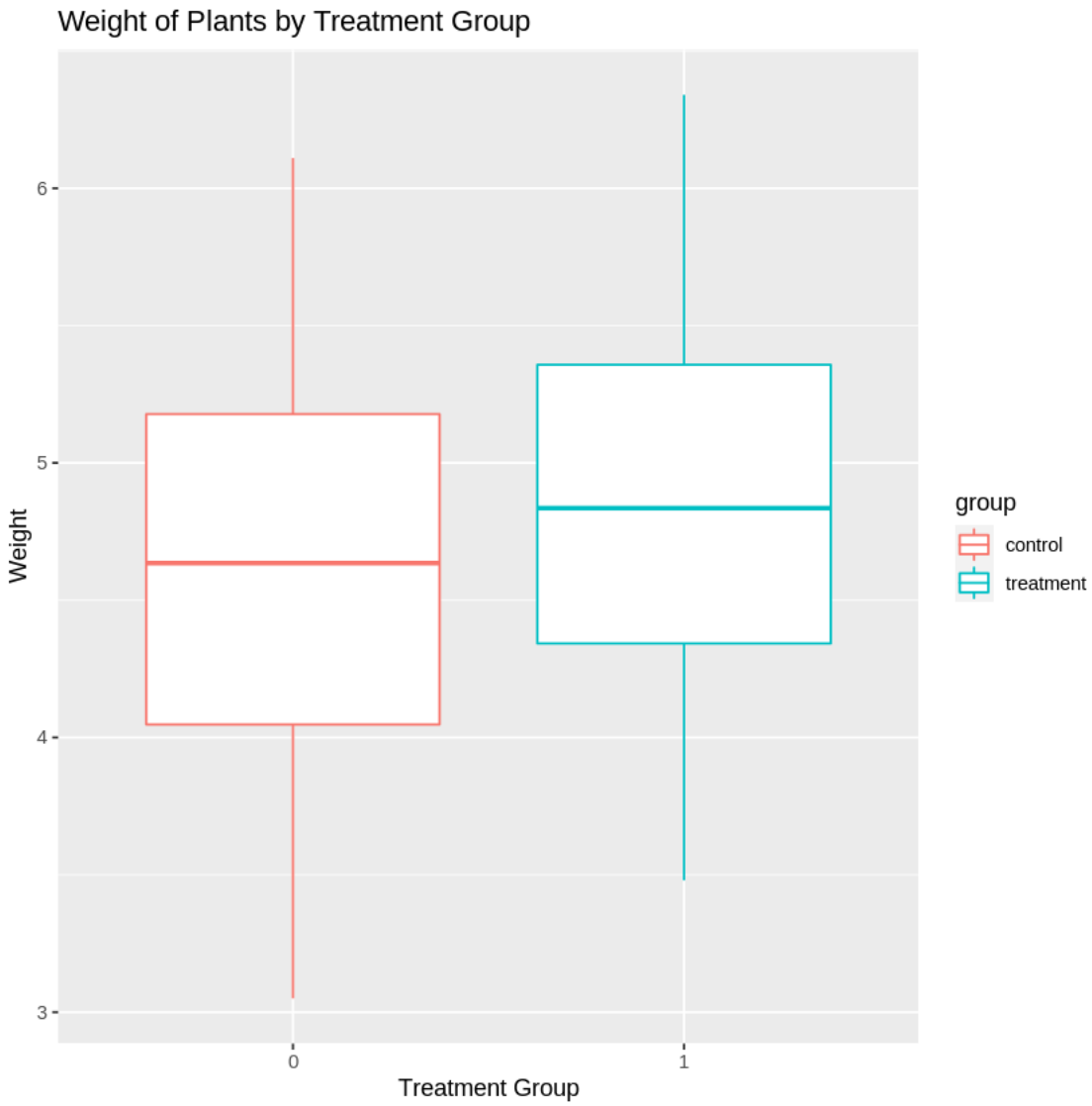
0.624161954022988

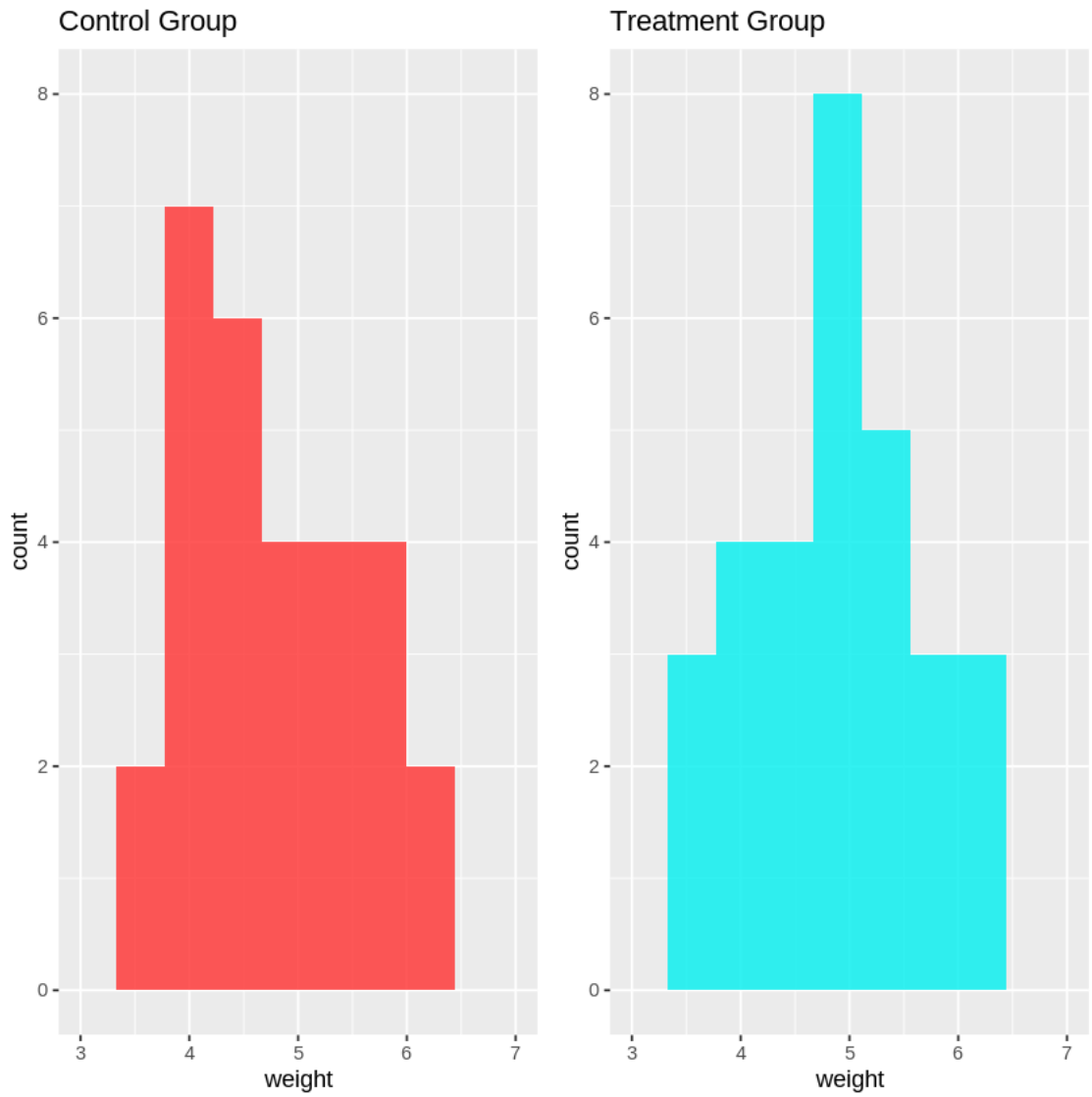
Graphs/Plots:

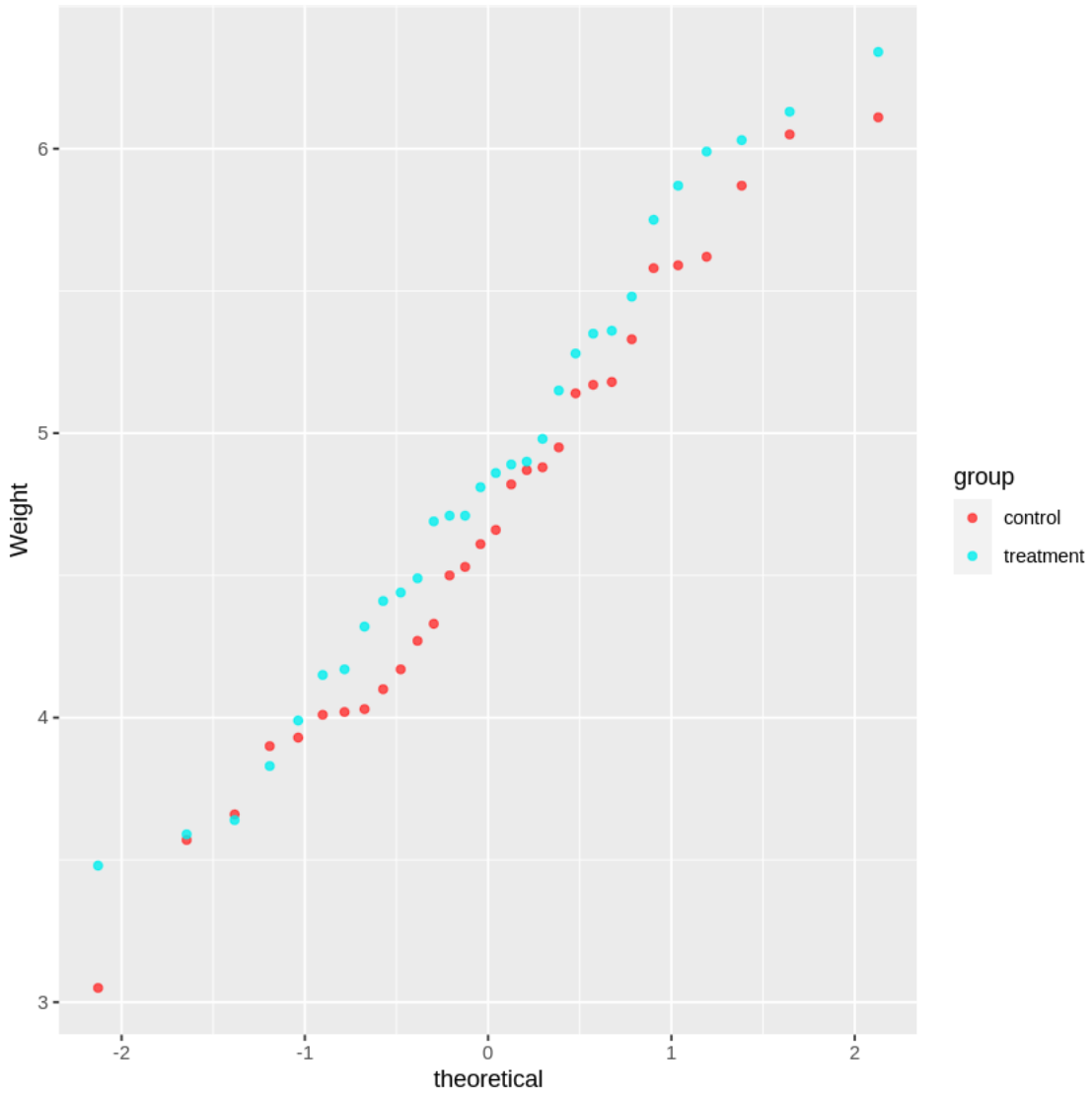
- Boxplots (show a side by side comparison of the mean and spread of plant weight for each treatment group)
- Histograms (show a side by side comparison of the distribution (appearing normal) of plant weight for each treatment group)
- Q-Q Plot (shows that the response variable (plant weight) is normally distributed and that the distribution for each treatment group is roughly the same (this is can be observed by noticing the parallel-like linear trend in the plot))

[6]:
```
#Boxplots
ggplot(seeds_tidy, aes(group=treatment, x=treatment, y=weight, color=group)) +␣
 ↪geom_boxplot() +
labs(x = "Treatment Group", y = "Weight", title = "Weight of Plants by␣
 ↪Treatment Group")

#Histograms
treatment = ggplot(seeds_tidy[seeds_tidy$group == "treatment", ]) +␣
 ↪geom_histogram(aes(x=weight), bins=10, fill="cyan2", alpha=0.8) + ylim(0,8)␣
 ↪+ xlim(3,7) + ggtitle("Treatment Group")
control = ggplot(seeds_tidy[seeds_tidy$group == "control", ]) +␣
 ↪geom_histogram(aes(x=weight), bins=10, fill="firebrick1", alpha=0.8) +␣
 ↪ylim(0,8) + xlim(3,7) + ggtitle("Control Group")
grid.arrange(control, treatment, ncol=2)

#Q-Q Plot
ggplot(seeds_tidy, aes(sample = weight)) + stat_qq(aes(color = group), alpha =␣
 ↪0.8) + scale_color_manual(values =c("firebrick1", "cyan2")) + labs(y =␣
 ↪"Weight")
```

```
Warning message:
"Removed 2 rows containing missing values (geom_bar)."
Warning message:
"Removed 2 rows containing missing values (geom_bar)."
```


Weight of Plants by Treatment Group
Weight
6
5
4
3
0
1
Treatment Group
group
control
treatment


Control Group
Treatment Group
count
weight
0
2
4
6
8
3
4
5
6
7



b) Interpretation of Results: Based on this EDA, it appears that seeds grown in a nutritionally enriched environment (treatment group) tend to produce plants that on average weigh more than those grown in standard conditions (control group). Given the relative position and spread of the data, it is possible (and expected) for some seeds in the control group to produce plants that weigh more than some of those in the treatment group. Although, generally speaking, it seems that plant weight is at least somewhat affected by the treatment. Despite this slight difference in means however, this EDA alone is not sufficient to make definitive claims about the relationship between the two variables (weight and treatment).

c) Hypothesis Test: To compare the two groups in this data (control and treatment) in a way that allows us to make a more sound conclusion about the effects of treatment on plant weight, it serves us best to conduct a two-sample independent t-test. Given that we have one categorical predictor variable with two independent groups we want to compare, this test allows us to not only

see whether plant weights differ between groups, but whether this difference is significant enough to reject the null hypothesis, which in this case, states that both group means are equal (and hence, the difference in group means equals 0). In turn, this will guide our conclusion regarding the strength and validity of our findings.

Results & Final Conclusion: The results of this t-test (below) show us that the mean plant weight for seeds in the control group is approximately 4.7 grams, while approximately 4.9 grams in the treatment group. At first glance, we see that the two are different, but we don't know whether this difference is significant enough to infer a true relationship between weight and treatment. Fortunately, the p-value of the t-statistic tells us that this difference is not in fact statistically significant. In other words, since our p-value is almost 0.4, this means that there is approximately a 40% chance that this difference is purely accidental. With a threshold of $p < 0.05$, we may not reject the null hypothesis, as the risk of rejecting it (40%) is too high. Still however, failing to reject the null hypothesis does not equate to accepting it. That is, just because we can't reject the possibility that the true population means are equal, does not mean that they in fact are. What we can say is that this data alone does not provide enough evidence to support any claims about the population, or the true effects of treatment on plant weight. Obtaining larger samples however, could be the first step that leads to deciding whether or not there actually exists a relationship between the environment a seed is grown in and the weight of the plant it produces.

[7]:
```
#two-sample independent t-test
t.test(seeds$control, seeds$treatment, alternative = "two.sided", var.equal =␣
 ↪FALSE)
```

```
        Welch Two Sample t-test

data:  seeds$control and seeds$treatment
t = -0.87161, df = 57.984, p-value = 0.387
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.5812966  0.2286299
sample estimates:
mean of x mean of y
 4.683333  4.859667
```

1.1.2 Question 2:

The dataset "AIDS_cases.csv" include the numbers of cases of AIDS in Australia by date of diagnosis for successive 3-month periods from 1984 to 1988.(Data from National Centre for HIV Epidemiology and Clinical Research,1994.)

a) Perform a comprehensive EDA to inspect, understand and describe the information collected in this dataset. Use appropriate summary statistics and plotsto present your results from the EDA.

b) What is your conclusion about AIDS incidence in Australia during this time period?

c) Suggest two probability distributions which could be good candidates for fitting the number of cases in this study. Which probability distribution would you prefer to fit the data in this particular example? Why?

[8]:
```
#importing "AIDS_cases" data
aids <- read.csv("/home/jovyan/AGLM/AIDS_cases.csv")
aids
```

A data.frame: 20 × 3

year <int>	quarter <int>	cases <int>
1984	1	1
1984	2	6
1984	3	16
1984	4	23
1985	1	27
1985	2	39
1985	3	31
1985	4	30
1986	1	43
1986	2	51
1986	3	63
1986	4	70
1987	1	88
1987	2	97
1987	3	91
1987	4	104
1988	1	110
1988	2	113
1988	3	149
1988	4	159

a) Exploratory Data Analysis (EDA) The variables in this dataset are as follows: - Outcome - (Y: cases) - Covariate - (X: time)

The primary outcome of interest is a discrete random variable (count data), while the predictor variable (covariate) is discrete with ordinal scale (discrete time intervals).

This EDA consists of: - Descriptive Statistics - Bar Graph - Line Plot - Regression Model

[10]:
```
#DATA WRANGLING

#adding another column that gives all quarters or 3-month periods (1-20) in␣
 ↪order (for bar graph)
aids$quarter_n <- as.numeric(row.names(aids))
head(aids)
```

A data.frame: 6 × 4

	year <int>	quarter <int>	cases <int>	quarter_n <dbl>
1	1984	1	1	1
2	1984	2	6	2
3	1984	3	16	3
4	1984	4	23	4
5	1985	1	27	5
6	1985	2	39	6

Descriptive Statistics:

- Summary of cases (minimum value, 1st quartile, median, mean, 3rd quartile, maximum value) for the whole data and for each year
- Standard deviation (SD) and variance of cases
- SD of cases for each year

[11]:
```
#summary of total cases and cases per year
summary(aids$cases)
by(aids$cases, aids$year, summary, na.rm=TRUE)

#SD and Variance of total cases and cases per year
sd(aids$cases)
var(aids$cases)

#SD of cases per year
by(aids$cases, aids$year, sd, na.rm=TRUE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   29.25   57.00   65.55   98.75  159.00
```

```
aids$year: 1984
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    4.75   11.00   11.50   17.75   23.00
------------------------------------------------------------
aids$year: 1985
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   29.25   30.50   31.75   33.00   39.00
------------------------------------------------------------
aids$year: 1986
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  43.00   49.00   57.00   56.75   64.75   70.00
------------------------------------------------------------
aids$year: 1987
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  88.00   90.25   94.00   95.00   98.75  104.00
------------------------------------------------------------
aids$year: 1988
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  110.0   112.2   131.0   132.8   151.5   159.0
```

46.4060737857721

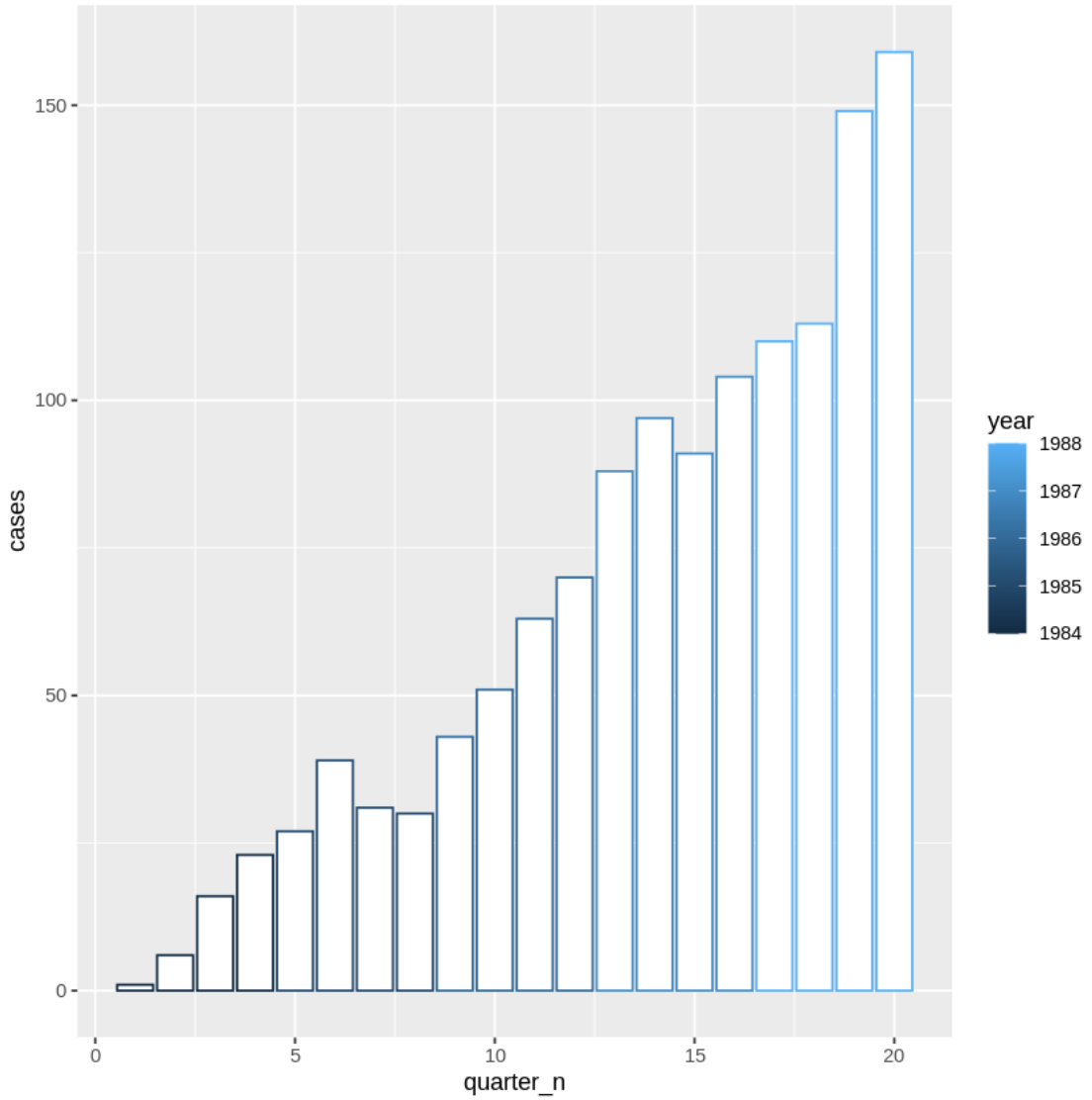
2153.52368421053

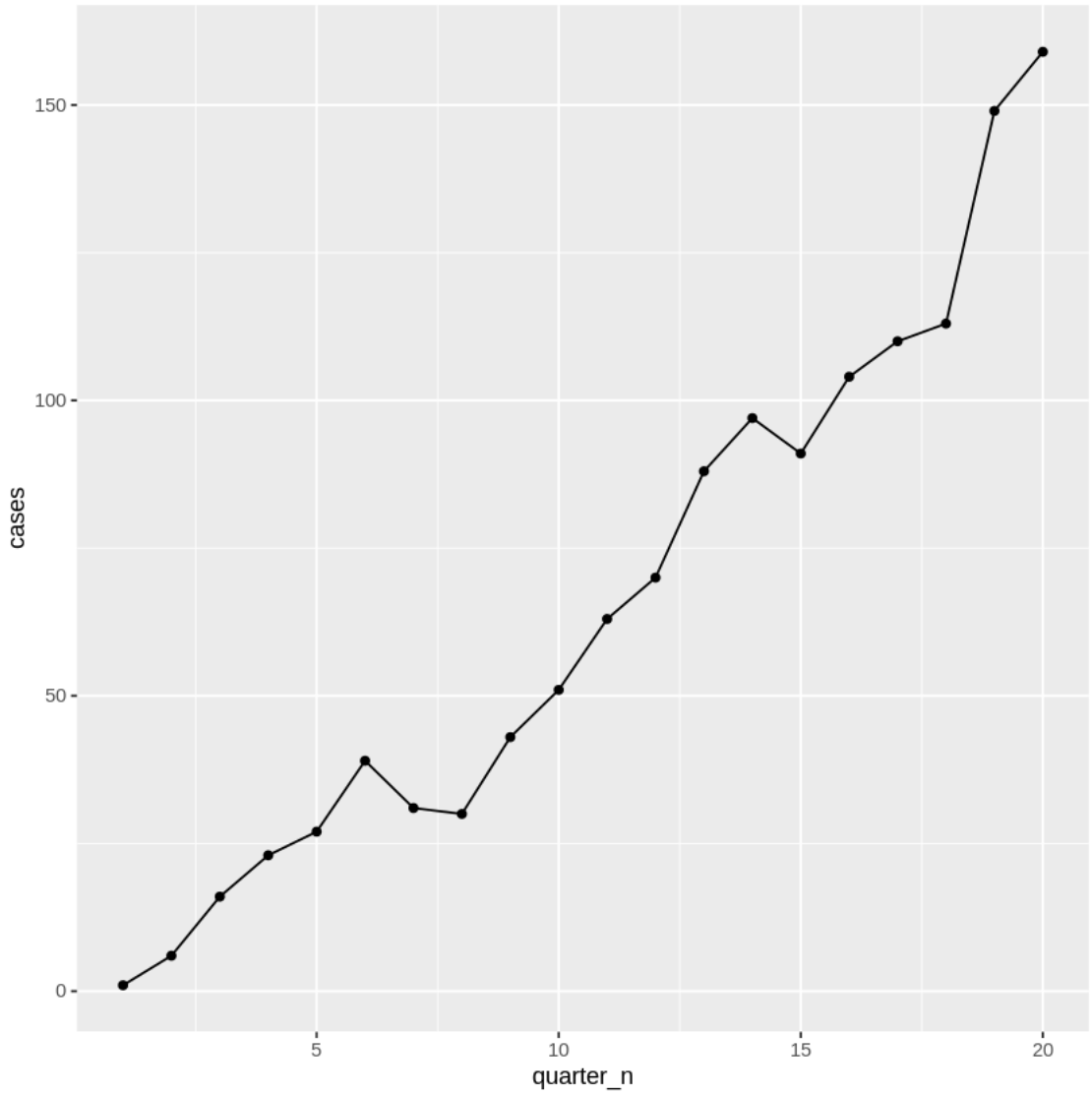
```
aids$year: 1984
[1] 9.882645
------------------------------------------------------------
aids$year: 1985
[1] 5.123475
------------------------------------------------------------
aids$year: 1986
[1] 12.06579
------------------------------------------------------------
aids$year: 1987
[1] 7.071068
------------------------------------------------------------
aids$year: 1988
[1] 24.90482
```

Graphs/Plots:

- Bar Graph (shows the number of AIDS cases for every 3-month period from 1984 to 1988 in chronological order)
- Line Plot (shows the increasing trend of AIDS cases from 1984 to 1988)
- Regression Model (shows the linear relationship between number of cases and time and models the general behavior of the data to allow for estimates and predictions)

[12]:
```
#Bar Graph
ggplot(aids, aes(x=quarter_n, y=cases, color=year)) + geom_bar(stat="identity",␣
↪fill="white")

#Line Plot
ggplot(aids, aes(x=quarter_n, y=cases, group=1)) + geom_line() + geom_point()
```

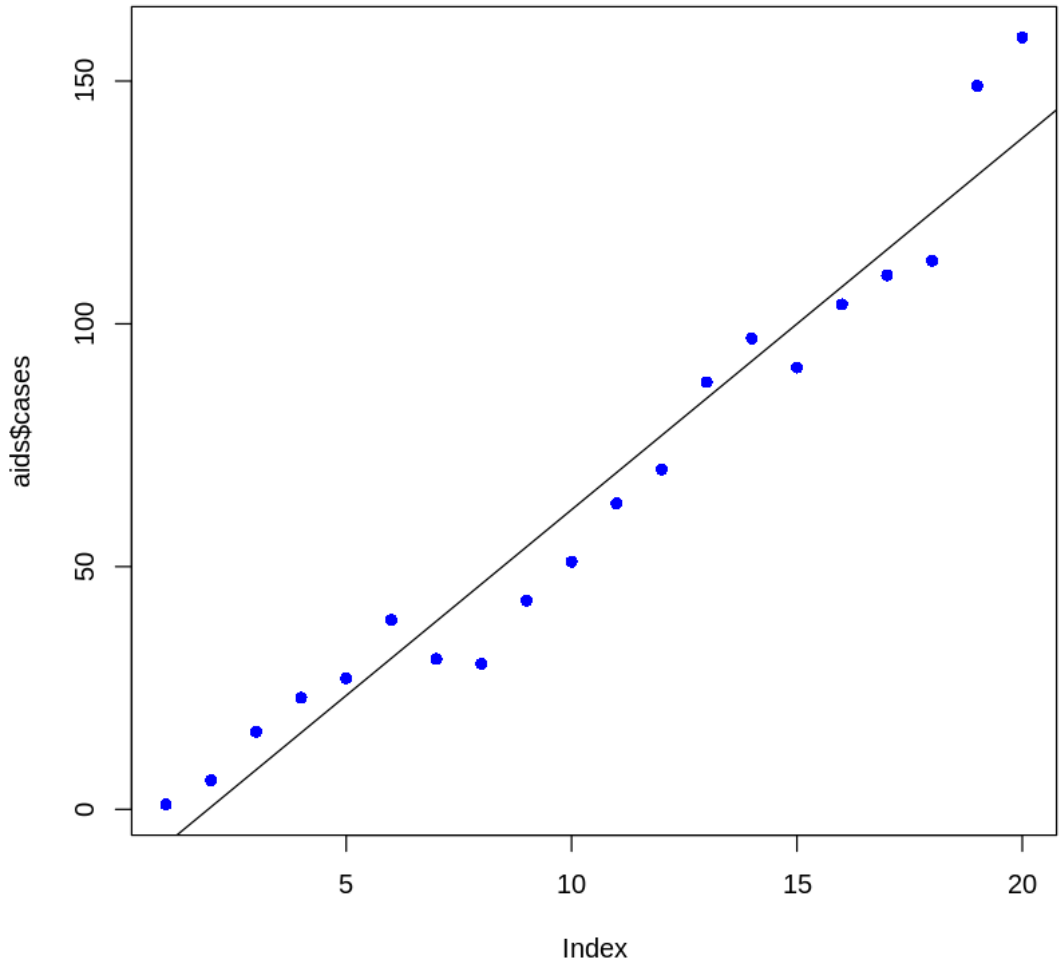

150
100
50
0
cases
0
5
10
15
20
quarter_n
year
1988
1987
1986
1985
1984



[13]:
```r
#Regression Model:
lm_aids = lm(cases ~ quarter_n, data = aids)
summary(lm_aids) #building a model (shows statistical signifigance)

plot(aids$cases, pch = 16, col = "blue") #plotting the data
abline(lm_aids) #adding the regression line from the model
```

```
Call:
lm(formula = cases ~ quarter_n, data = aids)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-16.4165  -8.0699  -0.1635   7.3603  20.7429

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.8105     4.8536  -3.051  0.00687 **
quarter_n     7.6534     0.4052  18.889 2.58e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.45 on 18 degrees of freedom
Multiple R-squared:  0.952,        Adjusted R-squared:  0.9493
F-statistic: 356.8 on 1 and 18 DF,  p-value: 2.577e-13
```

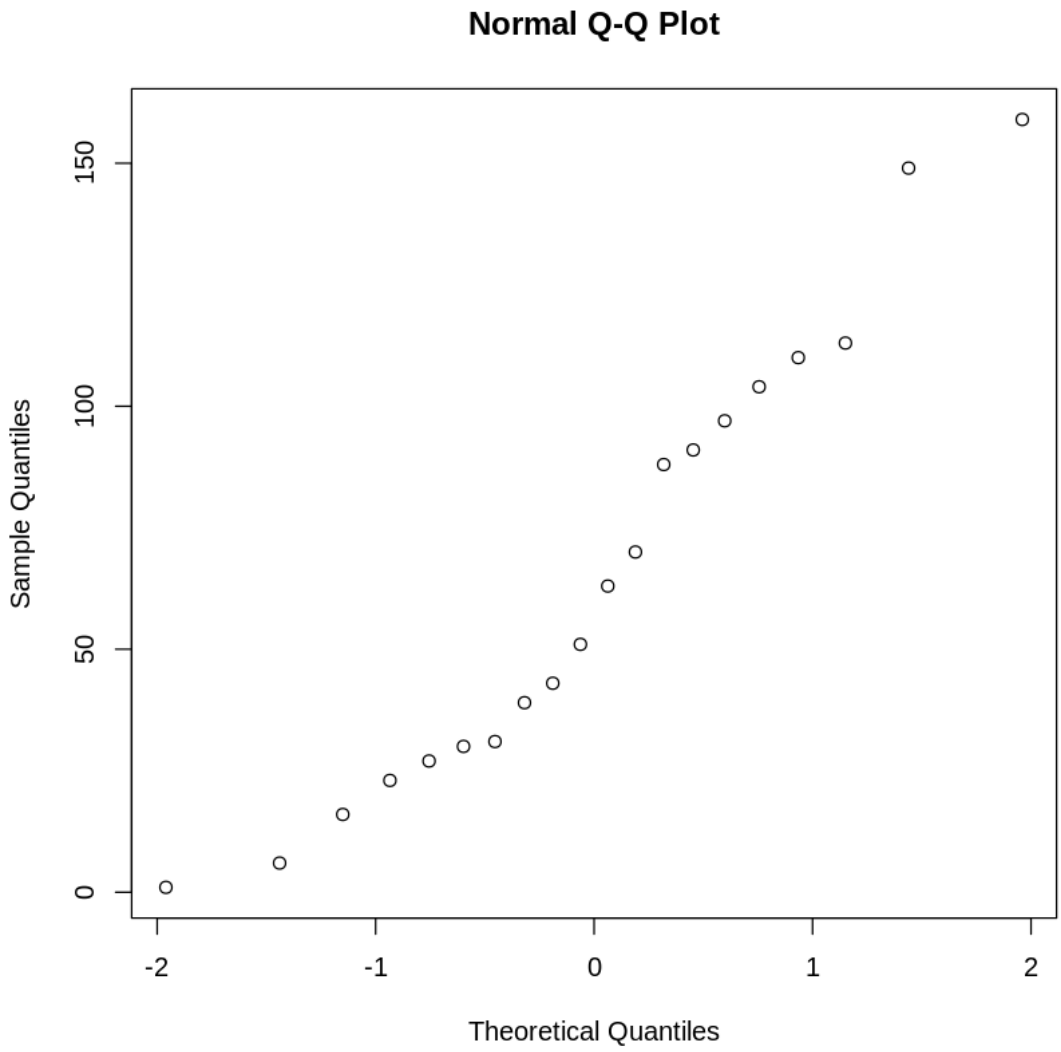


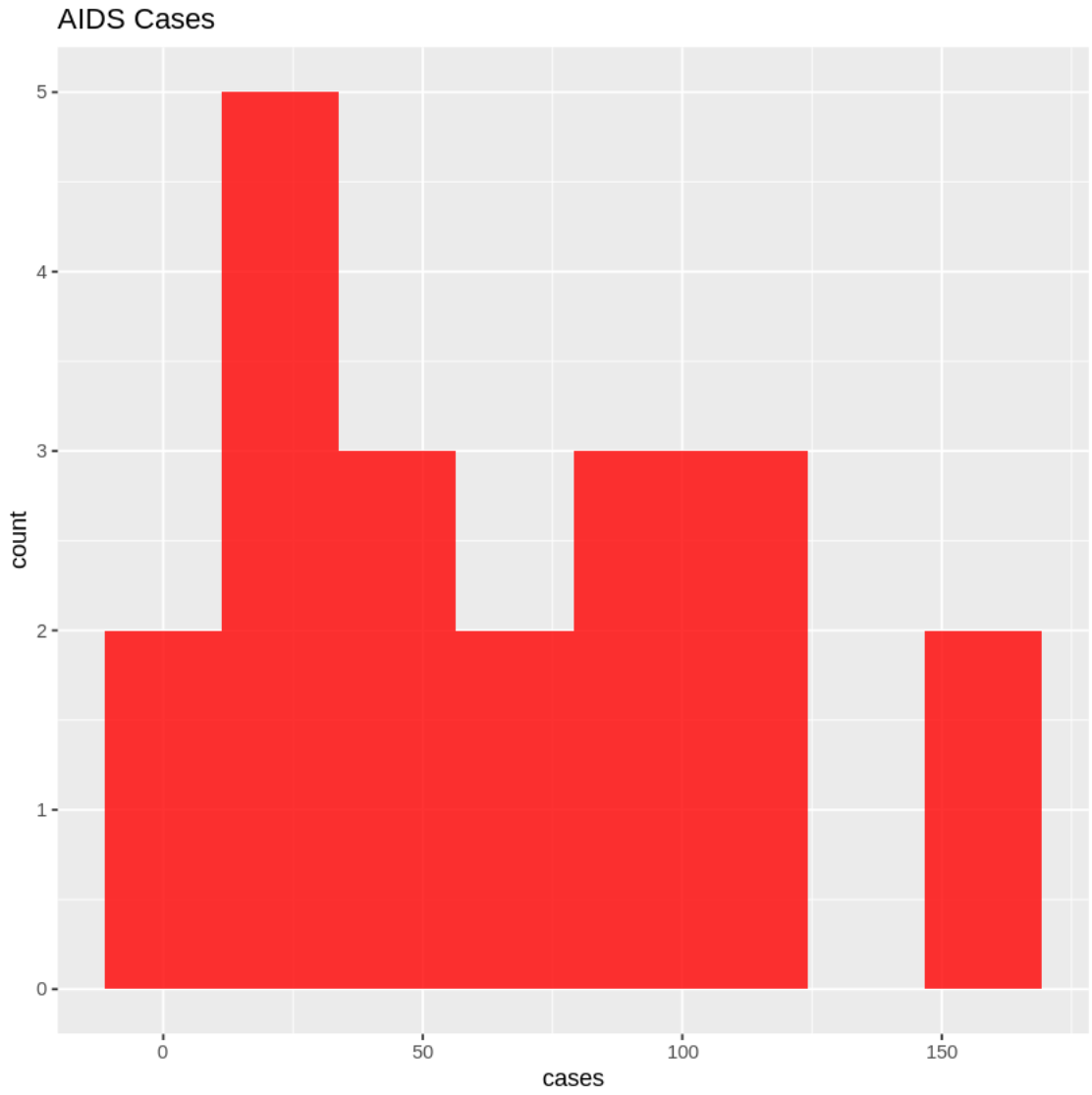
b) Interpretation of Results & Conclusion: Given the EDA, it is safe to conclude that AIDS cases in Australia increased at a relatively constant rate in this 5 year period, with the starting number of cases being 1, reaching a maximum of 159 cases in the last 3-month period. The results of the linear model above however, show that cases increased specifically at a rate of approximately 8 cases every 3 months from 1984 to 1988. That is, the model (y = -14.8105 + 7.6534x) predicted an increase of about 8 in the outcome variable (number of cases) for every 1 unit increase in the predictor variable (3 months). Moreover, looking at the adjusted R-squared value of approximately 0.95 and p-value of less than 0.001, we can say with confidence that this model is an adequate predictor of the number of AIDS cases for a given the yearly quarter in this time period. Thus, not only are we in a position to reject the null hypothesis that there is no linear increase in AIDS cases (respective to time), but we may claim that this model accounts for about 95% of the total error from the mean (TSS), meaning that only about 5% of error in the expected/predicted number of cases is unexplained. Therefore, we may claim that this model both describes the overall behavior of the data with high predictive power, and tells us that the number of AIDS cases in Australia increased at a relatively constant rate from 1984 to 1988 (on average 8 cases every 3 months).

c) Suggested Probability Distributions: Based on the type of data used in this study and the observations gethered from the Q-Q plot and histogram below, two possible probability distributions for fitting the number of cases are: - Normal Distribution - Poisson Distribution

The Q-Q plot below was used to observe whether the number of cases in this study are normally distributed. Given the general linearity of the graph, it appears that our outcome variable could be normally distributed. However, slight curves in the data indicate a potential skew. And, since the mean number of AIDS cases is greater than the median, it is safe to assume that this skew is positive (that is, the data is skewed to the right). This intuition is supported by the histogram below, which displays the frequency of cases in intervals of roughly 20. We see that the data is somewhat right skewed, although the size of our sample limits the validity of that assumption. Nonetheless, this potential right skew combined with the fact we are dealing with discrete count data, hints at the possibility that a Poisson distribution may also be a good fit. Still however, due to the results from the Q-Q plot and the spread of the data, it may be safer to assume that the number of cases follows a normal distribution with high sd/variance and a right skew.

[14]:
```
#Q-Q plot to see if number of cases is normally distributed
qqnorm(aids$cases)

#frequency histogram to observe the distrubution of the data
#since values (cases) range from about 1-160, we use 8 bins to capture the␣
 ↪frequency of cases in intervals of 20
ggplot(aids) + geom_histogram(aes(x=cases), bins=8, fill="red", alpha=0.8) +␣
 ↪ggtitle("AIDS Cases")
```


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0
50
100
150
-2
-1
0
1
2



1.1.3 Question 3:

(SEPARATE)

1.1.4 Question 4:

Studies have shown that the concentration of cholesterol in blood serum increases with age, but it is less clear whether cholesterol level is also associated with body weight. The dataset "Cholesterol.csv" includes values for serum cholesterol (in millimoles per liter), age (in years) and body mass index (BMIin kg/m^2)of n=30 women.

a) From BMI values create a new variable('BMIgrp')with the following three groups:

- "low" if BMI $<$ 21.5
- "medium" if 21.5 $\quad$ BMI $<$ 24.5

- "high" if BMI 24.5

What is the type of this new variable that you have created?

b) Perform a comprehensive exploratory analysis to describe your data.

c) Without fitting any statistical model and based on information in variables 'CHOL', 'Age' and 'BMIgrp' (the variable you created in part a)) use hypothesis testing to perform meaningful comparisons among the three BMI groups. Describe the statistical process you followed. Summarize and present the results of this analysis. What is/are your final conclusion(s) based on the data collected in this sample?

d) Suppose that the primary research question is to understand and describe the relationship among cholesterol levels, age and BMI. Based on knowledge from previous courses on Statistical Analyses, as well as your practical experience and intuition, apply suitable statistical methods to analyze your data and draw meaningful conclusions.

- Explain the statistical analysis plan, and steps you followed to implement the methods you used for the statistical analysis.
- Present your results and conclusions in an effective and coherent way (use tables and plots).
- Summarize in a paragraph your main findings.
- Summarize in a few sentences any important limitations of the statistical approach(es) you used for the analysis.

a) BMIgrp Let X3 = "BMIgrp" be a new predictor variable that places each observation in one of the following three categories: - "low" if BMI < 21.5 - "medium" if 21.5 BMI < 24.5 - "high" if BMI 24.5

Then, random variable X3 is categorical with ordinal scale.

[15]:
```
chol <- read.csv("/home/jovyan/AGLM/Cholesterol.csv") %>%  #importing␣
↪"Cholesterol" data
   mutate(BMIgrp = ifelse(bmi < 21.5, "low",
                          ifelse(bmi >= 21.5 & bmi < 24.5, "medium",
                                 ifelse(bmi >= 24.5, "high", NA))))  #adding␣
↪X3 as a column in the data frame
chol
```

A data.frame: 30 × 4

chol <dbl>	age <int>	bmi <dbl>	BMIgrp <chr>
5.94	52	20.7	low
4.71	46	21.3	low
5.86	51	25.4	high
6.52	44	22.7	medium
6.80	70	23.9	medium
5.23	33	24.3	medium
4.97	21	22.2	medium
8.78	63	26.2	high
5.13	56	23.3	medium
6.74	54	29.2	high
5.95	44	22.7	medium
5.83	71	21.9	medium
5.74	39	22.4	medium
4.92	58	20.2	low
6.69	58	24.4	medium
6.48	65	26.3	high
8.83	76	22.7	medium
5.10	47	21.5	medium
5.81	43	20.7	low
4.65	30	18.9	low
6.82	58	23.9	medium
6.28	78	24.3	medium
5.15	49	23.8	medium
2.92	36	19.6	low
9.27	67	24.3	medium
5.57	42	22.0	medium
4.92	29	22.5	medium
6.72	33	24.1	medium
5.57	42	22.7	medium
6.25	66	27.3	high

b) Exploratory Data Analysis (EDA) The variables in this dataset are as follows: - Outcome - (Y: cholesterol) - Covariates - (X1: age, X2: BMI, X3: BMI rank) - NOTE: Not considering X2 and X3 simultaneously (may use only 1 at a time to denote BMI)

The primary outcome of interest is a continuous random variable with ratio scale, while predictor variable (covariate) X1 is technically continuous, but discrete in this case, with ratio scale. Moreover, predictor variable (covariate) X2 is continuous with ratio scale, and predictor variable (covariate) X3 is categorical with ordinal scale.

This EDA consists of: - Descriptive Statistics - Scatter Plots - Correlation Coefficients

Descriptive Statistics:

- Summary of cholesterol (minimum value, 1st quartile, median, mean, 3rd quartile, maximum value) for the whole data and for each BMI group
- Standard deviation (SD) and variance of cholesterol

- SD of cholesterol for each BMI group

[16]:
```
#summary of cholesterol data and cholesterol given bmi group
summary(chol$chol)
by(chol$chol, chol$BMIgrp, summary, na.rm=TRUE) #notice mean cholesterol␣
 ↪increases with respect to BMI group

#SD and variance of cholesterol
sd(chol$chol)
var(chol$chol)

#SD of cholesterol for each group
by(chol$chol, chol$BMIgrp, sd, na.rm=TRUE)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.920   5.135   5.845   6.005   6.647   9.270
```

```
chol$BMIgrp: high
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.860   6.250   6.480   6.822   6.740   8.780
------------------------------------------------------------
chol$BMIgrp: low
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.920   4.665   4.815   4.825   5.588   5.940
------------------------------------------------------------
chol$BMIgrp: medium
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.920   5.190   5.830   6.163   6.705   9.270
```

1.30915542661461

1.71388793103448

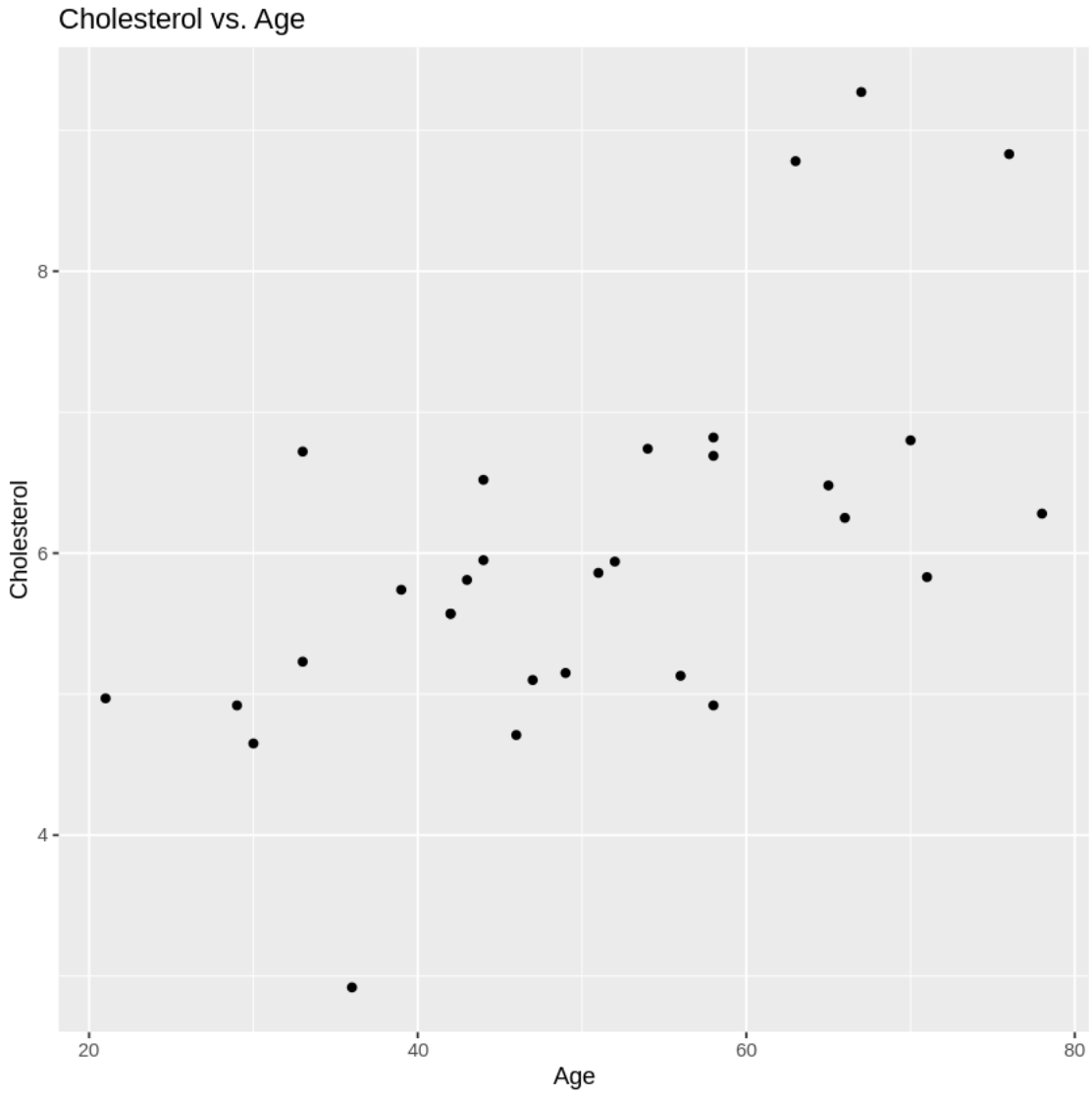
```
chol$BMIgrp: high
[1] 1.14128
------------------------------------------------------------
chol$BMIgrp: low
[1] 1.08585
------------------------------------------------------------
chol$BMIgrp: medium
[1] 1.211825
```

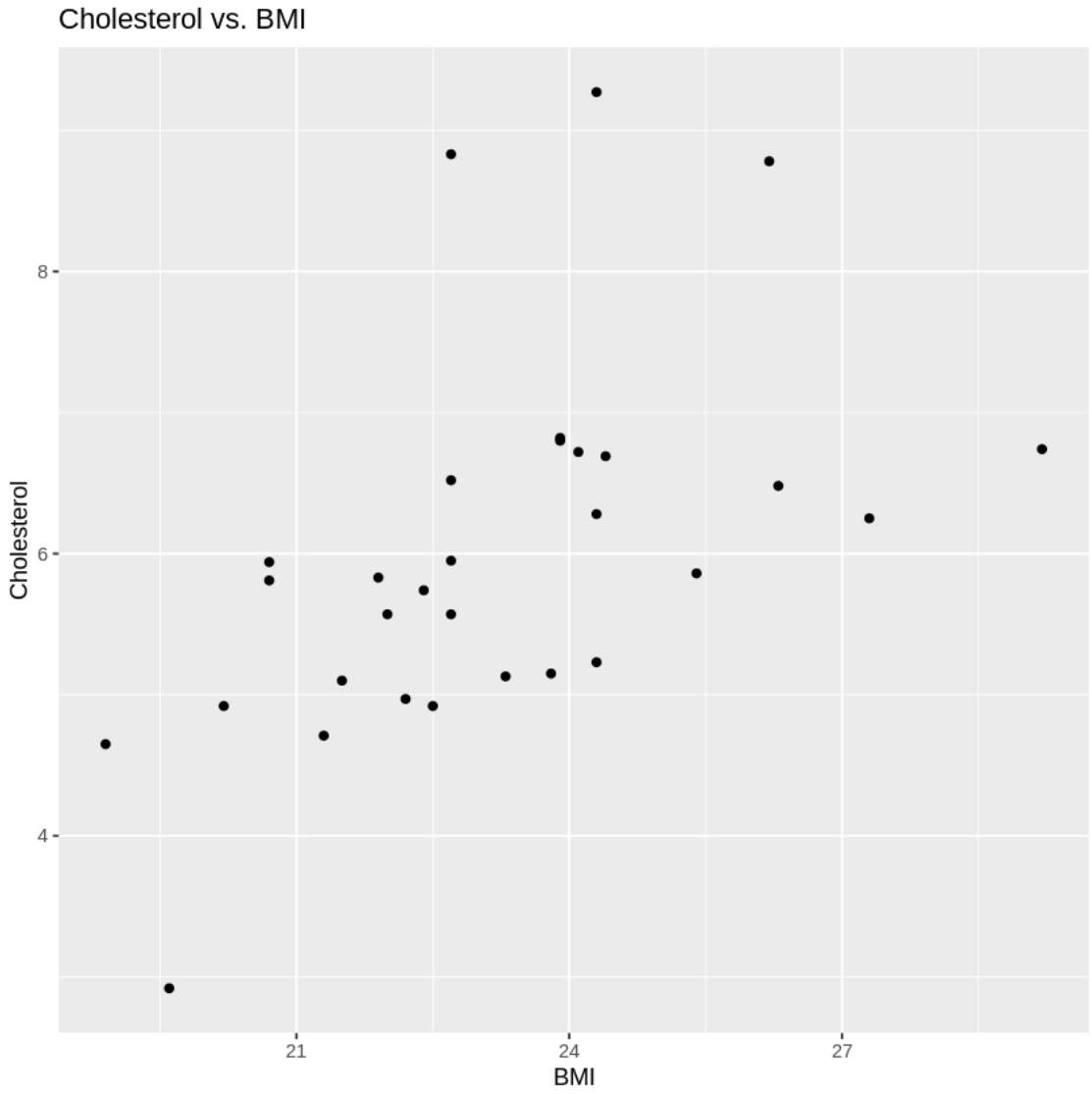
Graphs/Plots:

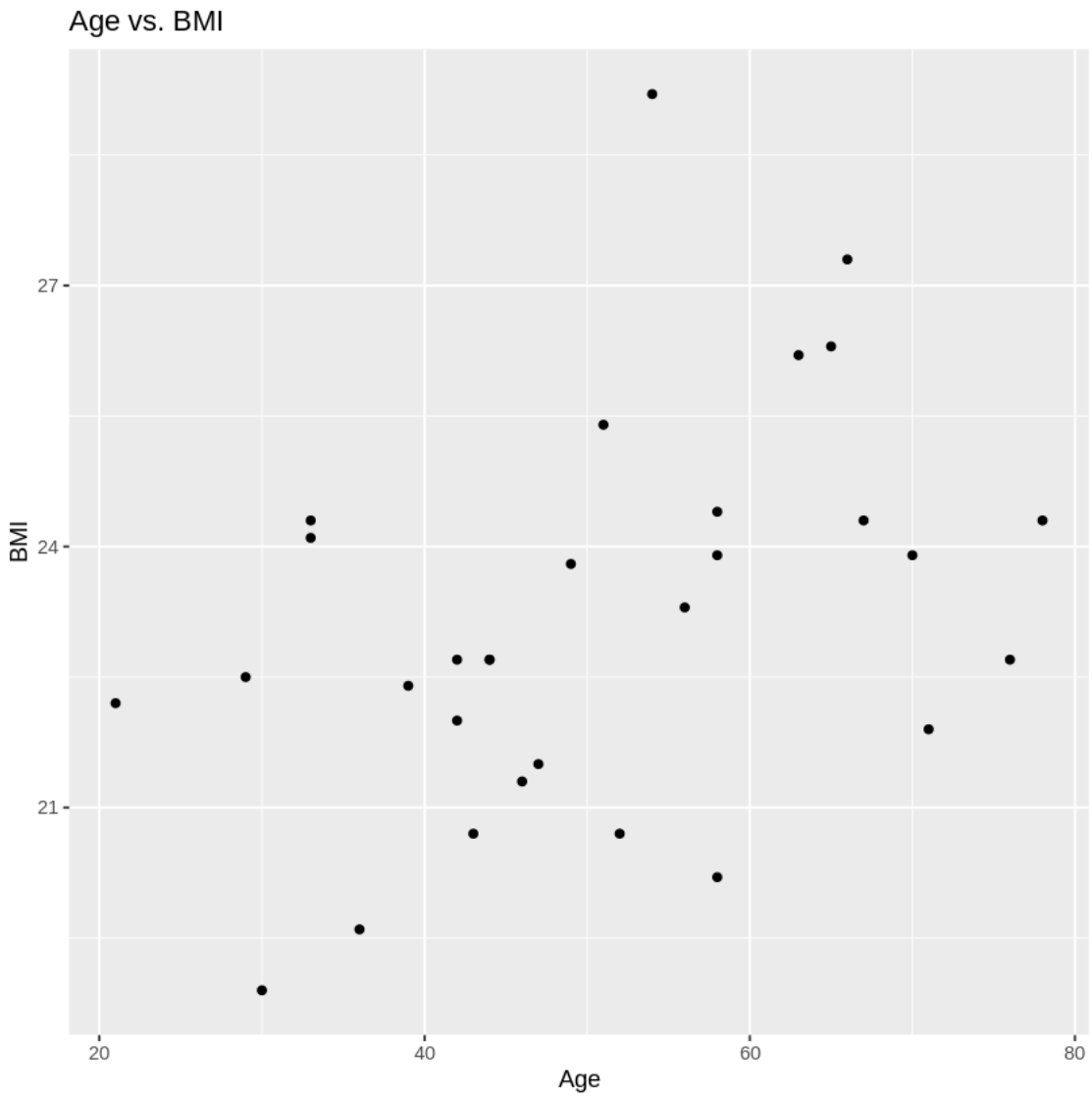
- Scatter Plots (show all the relationships between variables in the data)
- Correlation Coefficients (quantify the stregth of variable-variable relationships)

[17]:
```
#exploring relationships between variables

#Scatter Plot for Y and X1 (chol vs. age)
```

```r
ggplot(chol) + geom_point(aes(x = age, y = chol)) +
labs(x = "Age", y = "Cholesterol", title = "Cholesterol vs. Age")

#Scatter Plot for Y and X2 (chol vs. BMI)
ggplot(chol) + geom_point(aes(x = bmi, y = chol)) +
labs(x = "BMI", y = "Cholesterol", title = "Cholesterol vs. BMI")

#Scatter Plot for X1 and X2 (age vs. BMI)
#checking for collinearity
ggplot(chol) + geom_point(aes(x = age, y = bmi)) +
labs(x = "Age", y = "BMI", title = "Age vs. BMI")
```




Cholesterol vs. BMI
Cholesterol
8
6
4
21
24
27
BMI



[18]:
```
#correlation coefficients
cor(chol$chol, chol$age) #somewhat large effect size, so we can assume that the␣
 ↪positive linear relationship between cholesterol and age is practically␣
 ↪significant
cor(chol$chol, chol$bmi) #somewhat large effect size (though not as large as␣
 ↪the previous one), so we can assume that the positive linear relationship␣
 ↪between cholesterol and bmi is practically significant
cor(chol$age, chol$bmi) #not very large effect size, so we can't assume a␣
 ↪strong positive linear relationship between bmi and age, although there may␣
 ↪be some kind of relationship (is not practically significant)
```

0.602903709671754

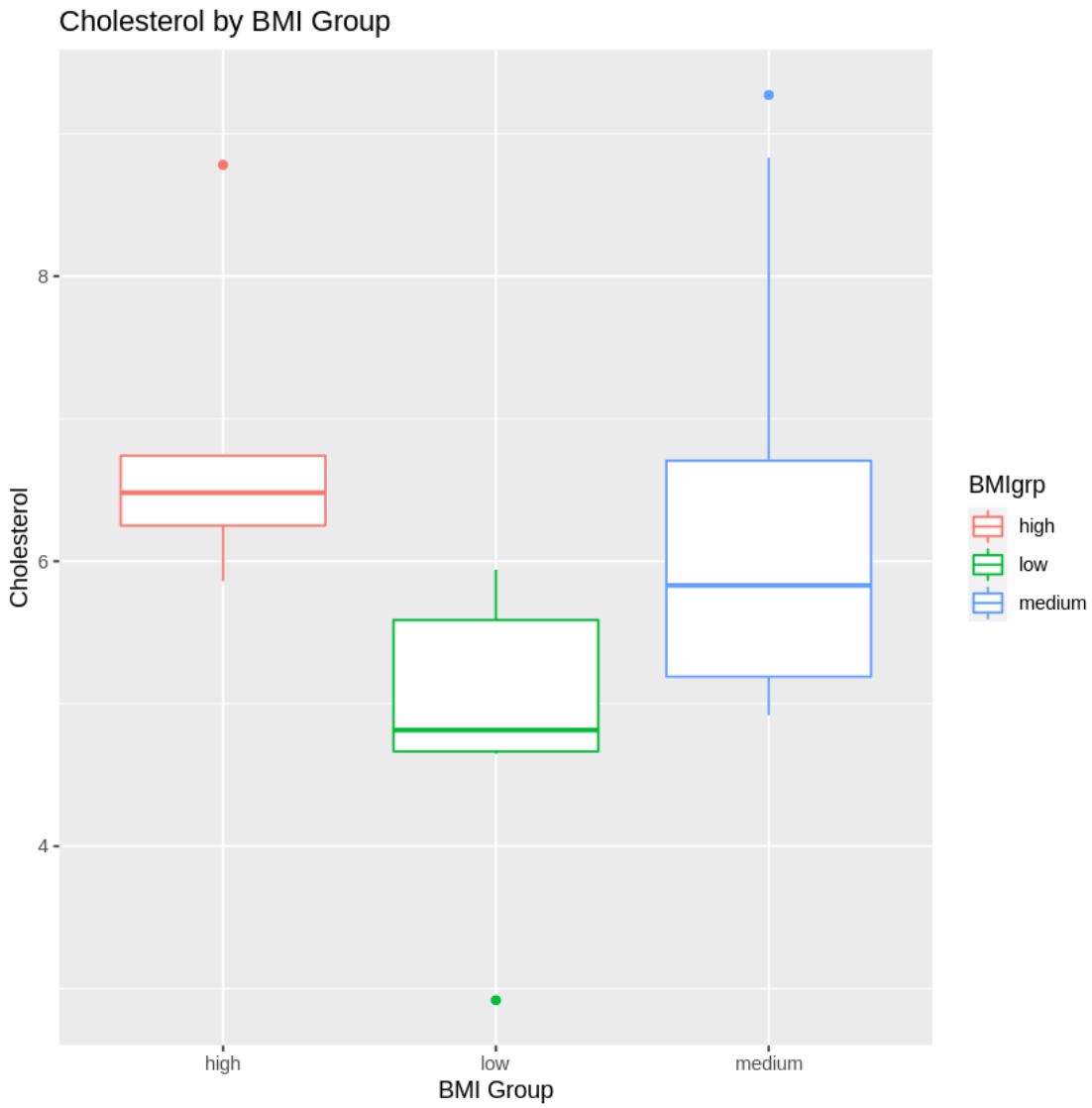
0.535287673092347

0.403315360753672

c) Hypothesis Test: To get a visual for how cholesterol levels differ with respect to BMI group, the boxplots below were constructed. However, given that we want to numerically compare cholesterol levels (a continuous variable) among three BMI groups (categorical predictor variable with 3 groups), we ought to use a one-way ANOVA. Using this hypothesis test, allows us to determine whether or not mean cholesterol levels differ between groups and whether this difference, if any, is significant enough to reject the null hypothesis, which under ANOVA, states that all group means are equal (and hence, the difference in group means equals 0). The more tangible results from this test will then help guide our intuitions about the effect that BMI may have on cholesterol.

Results & Final Conclusion: The p-value of approximately 0.023 corresponding to the calculated F-value in the ANOVA (below) tells us there exists some difference in cholesterol means between groups, and with a threshold of p-value < 0.05, that this difference is statistically significant. That is, the probability that this difference occurred under the null hypothesis is only about 2.3%, which indicates that it was not likely a result of pure chance. With that, we are in a position to reject the null hypothesis, stating that all group means are equal. This is verified by the boxplots below. However, ANOVA helps us to the extent that it affirms there is a significant difference in group means. It does not tell us whether all group means are different from each other. To see if any two means are the same or not significantly different we would need to perform three separate t-tests to compare the means between all three BMI groups. That is, one t-test for each pair: low and medium; low and high; and medium and high. The results from all such t-tests (below) prove that all group means are different, but also show that they are not all statistically significant. From the p-values obtained, we see that there is a statistically significant difference in cholesterol means between low and medium BMI groups, as well as low and high BMI groups, yet not for medium and high BMI groups. That is, the difference in cholesterol means between the medium (6.16) and high (6.82) BMI groups is much smaller than the difference in means between the low (4.83) and medium/high BMI groups, meaning that the former difference is not statistically significant. With a p-value of approximately 0.3, and hence a 30% chance that this difference is purely coincidental, we may not reject the possibility that the true average cholesterol levels for medium and high BMI groups are the same (null hypothesis). Thus, we can infer from this sample data that cholesterol levels tend to be significantly lower for those belonging to the low BMI group, but whether or not there is any significant difference in cholesterol levels among medium and high BMI groups is something that would require more data and further exploration to confirm. In the boxplots below, the cholesterol levels for the medium BMI group appear to me more spread out, which could have impacted the outcome of the t-tests. Nonetheless, more data would provide us with more clarity regarding this issue. Despite this uncertainty, the results and observations from these tests, together, provide sufficient grounds to conclude that there is in fact a relationship between cholesterol and BMI.

[19]:
```
#Boxplots
ggplot(chol, aes(group=BMIgrp, x=BMIgrp, y=chol, color=BMIgrp)) +␣
 ↪geom_boxplot() +
labs(x = "BMI Group", y = "Cholesterol", title = "Cholesterol by BMI Group")
```



[20]:
```r
#Hypothesis Test: ANOVA (one-way)
chol_anova <- aov(chol ~ BMIgrp, data = chol)
summary(chol_anova)

#t-tests for group-group comparisons
low_df <- chol[which(chol$BMIgrp == "low"), ]
med_df <- chol[which(chol$BMIgrp == "medium"), ]
high_df <- chol[which(chol$BMIgrp == "high"), ]

t.test(low_df$chol, med_df$chol) #low vs. medium
t.test(low_df$chol, high_df$chol) #low vs. high
t.test(med_df$chol, high_df$chol) #medium vs. high
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
BMIgrp       2  12.16   6.082   4.374 0.0226 *
Residuals   27  37.54   1.390
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
        Welch Two Sample t-test

data:  low_df$chol and med_df$chol
t = -2.5563, df = 9.3067, p-value = 0.0301
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.5154111 -0.1598521
sample estimates:
mean of x mean of y
 4.825000  6.162632
```

```
        Welch Two Sample t-test

data:  low_df$chol and high_df$chol
t = -2.954, df = 8.4598, p-value = 0.01723
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.5413009 -0.4526991
sample estimates:
mean of x mean of y
    4.825     6.822
```

```
        Welch Two Sample t-test

data:  med_df$chol and high_df$chol
t = -1.1345, df = 6.5966, p-value = 0.2961
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.0509105  0.7321737
sample estimates:
mean of x mean of y
 6.162632  6.822000
```

d) Linear Regression Models: To understand the relationship between cholesterol levels, age, and BMI, as well as to see whether age and BMI are effective predictors of cholesterol levels, it seemed appropriate to resort to linear regression. Specifically, three linear models were created: cholesterol given age and BMI (full model); cholesterol given age; and cholesterol given BMI. This

provided a more clear picture of the predictive capabilities of age and BMI on cholesterol levels, together and separately.

Results & Final Conclusion: From the linear models constructed (below), we notice that the full model produced the lowest p-value (despite all being below the threshold of 0.05) and the highest adjusted R-squared value (which is expected as this model has the most variables). Moreover, although the p-values from the other two models indicate that age is a better predictor of cholesterol level than BMI, the combination of the two proves better than age alone. That is, according to the full model, a 1 unit increase in both age and BMI corresponds to an increase in cholesterol levels, meaning that despite not having the same degree of impact on cholesterol levels, age and BMI both contribute to explaining at least some of its variation. Given the adjusted R-squared of the full model, it seems that about 43% of the variation in cholesterol levels can be explained by age and BMI. Although this value is not high enough to give this model strong predictive power, it is significant enough to show that there is a strong relationship between outcome and predictor variables. Moreover, with a p-value < 0.05 we may conclude that at least one of the predictor variables is significantly related to the outcome (but, given the results of the other models and p-values associated with their coefficients in the full model, we know that they both are). What can be taken away from this statistical analysis is that although age and BMI are not the only predictors of cholesterol levels (there may be many others or better ones), they certainly have some sort of predictive power over cholesterol levels.

*Note: The sum of adjusted R-squared values for age and BMI is larger than the adjusted R-squared value of the full model because the individual models don't account for the shared R-squared values of both variables (that is, part of the R-squared value corresponding to age is due to BMI and vise versa, and the combined model takes this into account so as to not double count).

Limitations: Some limitations of the linear regression approach include assumptions about the distribution of the outcome variable and the type of relationship between dependent and independent variables, as well as a large focus on means and possible outliers. Given that linear regression assumes a normally distributed outcome variable and a linear relationship between outcome and predictor variables, these findings could be flawed if the data does not meet these assumptions. Moreover, given this method's emphasis on mean values and the possibility of outliers, it may be the case that our results were impacted by extreme values or inconsistencies within the data. In the boxplots above, we see that a few outliers were detected, meaning that it is possible our results and observations would have been different had they been excluded from the data and proceeding statistical analysis. However, this is not guaranteed.

[21]:
```r
#Regression Models:
chol_lm <- lm(chol ~ age + bmi, data = chol) #full model
summary(chol_lm)

chol_lm1 <- lm(chol ~ age, data = chol) #age model
summary(chol_lm1)

chol_lm2 <- lm(chol ~ bmi, data = chol) #BMI model
summary(chol_lm2)

#error rate for full model (%17) - estimated error of prediction
```

```r
sigma(chol_lm)/mean(chol$chol)
```

```
Call:
lm(formula = chol ~ age + bmi, data = chol)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7619 -0.7353 -0.0205  0.3772  2.3717

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.73983    1.89641  -0.390  0.69951
age          0.04097    0.01363   3.006  0.00567 **
bmi          0.20137    0.08876   2.269  0.03149 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.992 on 27 degrees of freedom
Multiple R-squared:  0.4654,        Adjusted R-squared:  0.4258
F-statistic: 11.75 on 2 and 27 DF,  p-value: 0.000213



Call:
lm(formula = chol ~ age, data = chol)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29944 -0.67361  0.02992  0.40873  2.39393

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.29561    0.70480   4.676 6.72e-05 ***
age          0.05344    0.01336   3.999 0.000422 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.063 on 28 degrees of freedom
Multiple R-squared:  0.3635,        Adjusted R-squared:  0.3408
F-statistic: 15.99 on 1 and 28 DF,  p-value: 0.0004216



Call:
lm(formula = chol ~ bmi, data = chol)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-1.97890 -0.80623 -0.07073  0.53611  2.97330

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.15683    2.14558  -0.539   0.5940
bmi          0.30897    0.09214   3.353   0.0023 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.125 on 28 degrees of freedom
Multiple R-squared:  0.2865,        Adjusted R-squared:  0.2611
F-statistic: 11.24 on 1 and 28 DF,  p-value: 0.002303
```

0.165198462656938