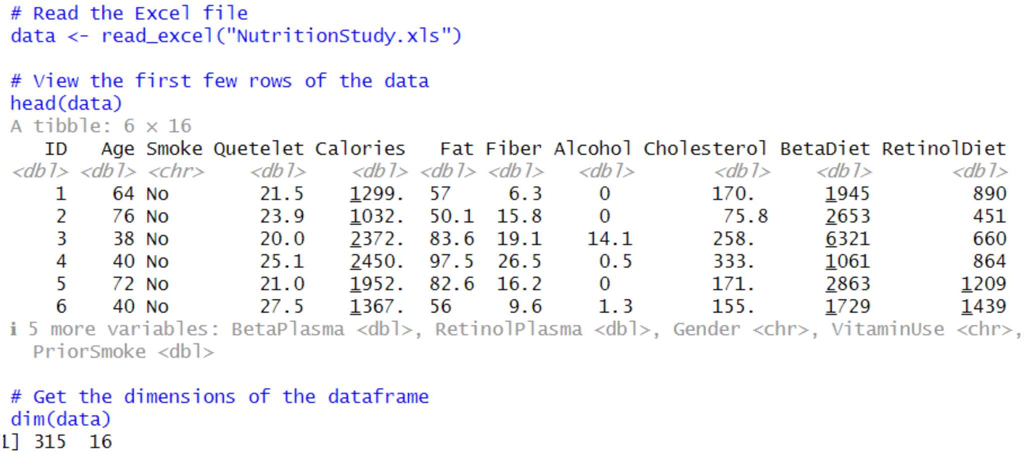
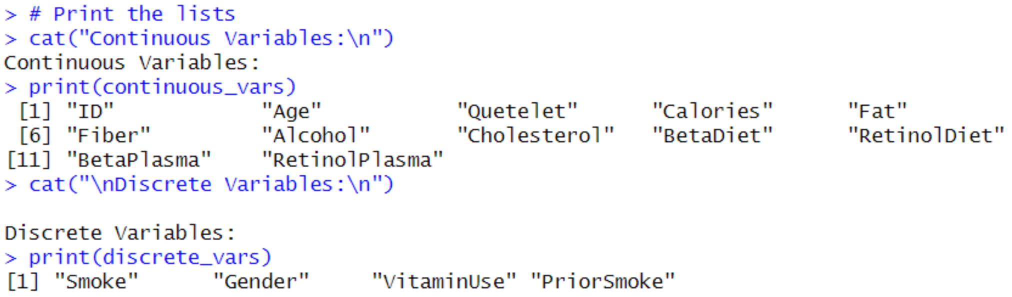
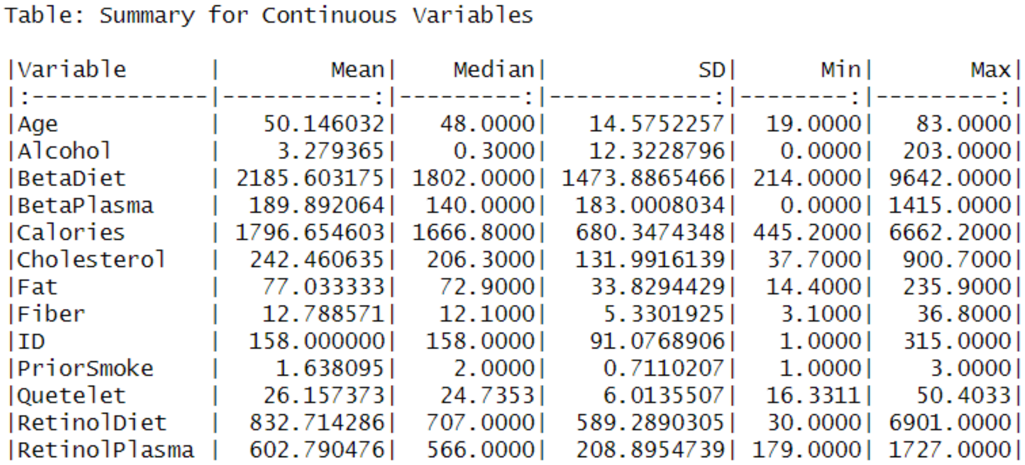
Dataset



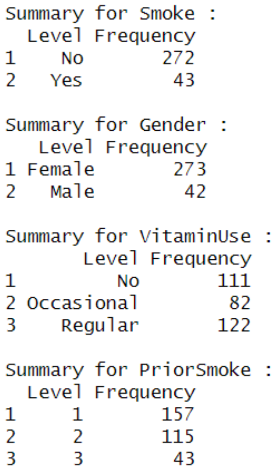
```
# Read the Excel file
data <- read_excel("NutritionStudy.xls")

# View the first few rows of the data
head(data)
A tibble: 6 × 16
    ID   Age Smoke Quetelet Calories   Fat Fiber Alcohol Cholesterol BetaDiet RetinolDiet
 <dbl> <dbl> <chr>    <dbl>    <dbl> <dbl> <dbl>   <dbl>       <dbl>    <dbl>       <dbl>
     1    64 No        21.5    1299.  57     6.3     0         170.      1945         890
     2    76 No        23.9    1032.  50.1  15.8     0          75.8     2653         451
     3    38 No        20.0    2372.  83.6  19.1    14.1       258.      6321         660
     4    40 No        25.1    2450.  97.5  26.5     0.5       333.      1061         864
     5    72 No        21.0    1952.  82.6  16.2     0         171.      2863        1209
     6    40 No        27.5    1367.  56     9.6     1.3       155.      1729        1439
ℹ 5 more variables: BetaPlasma <dbl>, RetinolPlasma <dbl>, Gender <chr>, VitaminUse <chr>,
  PriorSmoke <dbl>

# Get the dimensions of the dataframe
dim(data)
l] 315  16
```



```
> # Print the lists
> cat("Continuous Variables:\n")
Continuous Variables:
> print(continuous_vars)
 [1] "ID"            "Age"           "Quetelet"      "Calories"      "Fat"
 [6] "Fiber"         "Alcohol"       "Cholesterol"   "BetaDiet"      "RetinolDiet"
[11] "BetaPlasma"    "RetinolPlasma"
> cat("\nDiscrete Variables:\n")

Discrete Variables:
> print(discrete_vars)
[1] "Smoke"        "Gender"      "VitaminUse" "PriorSmoke"
```



```
Table: Summary for Continuous Variables

|Variable      |        Mean|    Median|           SD|      Min|       Max|
|:-------------|-----------:|---------:|------------:|--------:|---------:|
|Age           |   50.146032|   48.0000|   14.5752257|  19.0000|   83.0000|
|Alcohol       |    3.279365|    0.3000|   12.3228796|   0.0000|  203.0000|
|BetaDiet      | 2185.603175| 1802.0000| 1473.8865466| 214.0000| 9642.0000|
|BetaPlasma    |  189.892064|  140.0000|  183.0008034|   0.0000| 1415.0000|
|Calories      | 1796.654603| 1666.8000|  680.3474348| 445.2000| 6662.2000|
|Cholesterol   |  242.460635|  206.3000|  131.9916139|  37.7000|  900.7000|
|Fat           |   77.033333|   72.9000|   33.8294429|  14.4000|  235.9000|
|Fiber         |   12.788571|   12.1000|    5.3301925|   3.1000|   36.8000|
|ID            |  158.000000|  158.0000|   91.0768906|   1.0000|  315.0000|
|PriorSmoke    |    1.638095|    2.0000|    0.7110207|   1.0000|    3.0000|
|Quetelet      |   26.157373|   24.7353|    6.0135507|  16.3311|   50.4033|
|RetinolDiet   |  832.714286|  707.0000|  589.2890305|  30.0000| 6901.0000|
|RetinolPlasma |  602.790476|  566.0000|  208.8954739| 179.0000| 1727.0000|
```



```
Summary for Smoke :
  Level Frequency
1    No       272
2   Yes        43

Summary for Gender :
   Level Frequency
1 Female       273
2   Male        42

Summary for VitaminUse :
       Level Frequency
1         No       111
2 Occasional        82
3    Regular       122

Summary for PriorSmoke :
  Level Frequency
1     1       157
2     2       115
3     3        43
```

Introduction: The primary goal of this analysis is to explore the Nutrition Study Data with an emphasis on understanding the various variables and their characteristics, with a specific intent to use multiple regression to predict the variable 'CHOLESTEROL'. The dataset is an amalgamation of both categorical and continuous variables, which offers a unique opportunity to examine the influence of different types of explanatory variables on our target variable.

Dataset Overview: The Nutrition Study Data comprises 16 variables and has 315 records. The data is derived from medical records and self-reported observational data from adults. The absence of a data dictionary means we rely heavily on variable names and inherent characteristics of the data for interpretation.

Continuous Variables: Here's a summary of the key continuous variables:

- Age: Ranges from 19 to 83 years, with an average age of around 50.
- Alcohol: Consumption varies widely, with a maximum recorded value of 203 and about 50% of the participants consuming 0.3 or less.
- BetaDiet & BetaPlasma: Show a wide variation, indicating diverse dietary habits and plasma levels among participants.
- Calories: Average caloric intake is around 1796, with some individuals consuming as high as 6662 calories.
- Cholesterol: Our target variable has a mean of 242.46 and varies between 37.7 and 900.7.
- Fat & Fiber: Average values are 77 and 12.8 respectively, but there's considerable variation in the dataset.
- Quetelet: Represents the Body Mass Index (BMI). The average value is 26.15. A value above 25 indicates overweight and above 30 indicates obesity.
- RetinolDiet & RetinolPlasma: Indicate Vitamin A diet and plasma levels. Both show a wide range of values.

Categorical Variables:

- Smoke: Majority of the participants (272) do not smoke, while 43 responded with a 'Yes'.
- Gender: The dataset is predominantly female with 273 participants, while males are 42.
- VitaminUse: There is a fairly even distribution with 111 not using vitamins, 82 using occasionally, and 122 using regularly.
- PriorSmoke: A categorical variable indicating past smoking habits. Most participants (157) are categorized under '1'.

Preliminary Observations:

- The dataset is skewed in terms of gender, with a larger representation of females.
- Majority of the participants do not smoke, which is an interesting aspect to consider given the health context.
- There's a broad range in terms of dietary habits, alcohol consumption, and nutritional intake, offering diverse data points for analysis.
- The Quetelet or BMI variable will be crucial in understanding the overall health profile of participants. Given that an average value of 26.15 suggests many participants are on the verge of or are overweight, this can have implications on cholesterol levels.

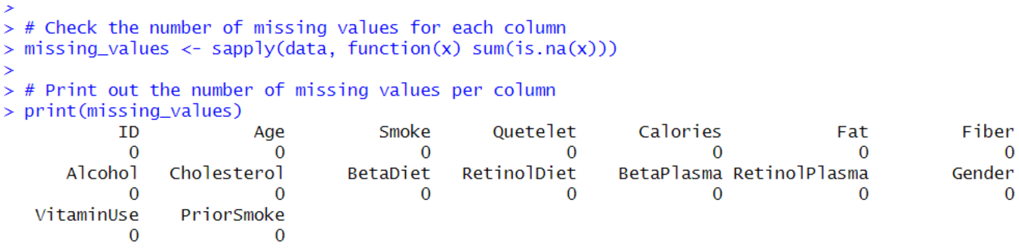
Preparatory Work: Before delving into regression modelling, a few steps need to be undertaken:

- Data Cleaning: Ensure there are no missing values or outliers that might skew our model.
- Variable Transformation: Given that regression requires numerical input, categorical variables like 'Smoke', 'Gender', and 'VitaminUse' will need to be dummy coded or recoded.

- Feature Engineering: Construct relevant interaction terms or composite scores that might enhance the model's explanatory power.

Conclusion: The Nutrition Study Data provides a rich tapestry of information on adults' dietary habits, health metrics, and lifestyle choices. By employing multiple regression with a mix of categorical and continuous explanatory variables, we aim to understand the factors influencing cholesterol levels.

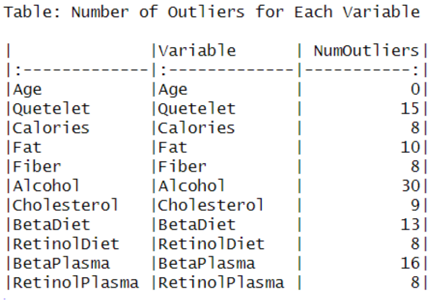
Missing Values:



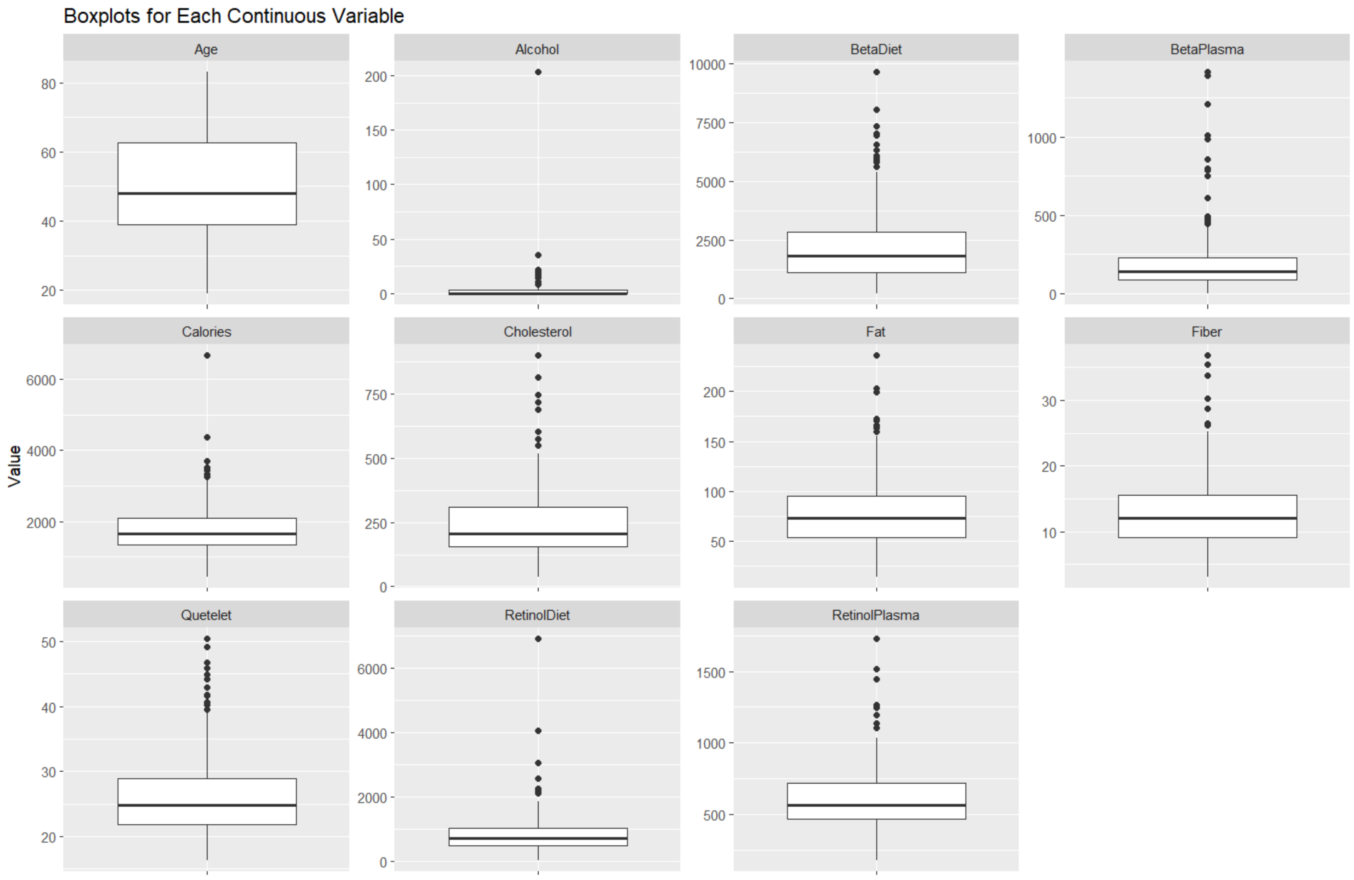
```
>
> # Check the number of missing values for each column
> missing_values <- sapply(data, function(x) sum(is.na(x)))
>
> # Print out the number of missing values per column
> print(missing_values)
          ID          Age        Smoke     Quetelet     Calories          Fat        Fiber
           0            0            0            0            0            0            0
     Alcohol  Cholesterol     BetaDiet   RetinolDiet   BetaPlasma RetinolPlasma      Gender
           0            0            0            0            0            0            0
  VitaminUse   PriorSmoke
           0            0
```

There are no missing values.

Outliers:



```
Table: Number of Outliers for Each Variable

|              |Variable      | NumOutliers|
|:-------------|:-------------|-----------:|
|Age           |Age           |           0|
|Quetelet      |Quetelet      |          15|
|Calories      |Calories      |           8|
|Fat           |Fat           |          10|
|Fiber         |Fiber         |           8|
|Alcohol       |Alcohol       |          30|
|Cholesterol   |Cholesterol   |           9|
|BetaDiet      |BetaDiet      |          13|
|RetinolDiet   |RetinolDiet   |           8|
|BetaPlasma    |BetaPlasma    |          16|
|RetinolPlasma |RetinolPlasma |           8|
```



Except Age, every continuous variable has outliers.

Preparatory Work

In this fifth Modeling Assignment, we are working with a brand-new dataset. When you are in such a situation, in addition to Sample Population determination and EDA to understand the data, you may also need to do some transformations on the existing variables. If you intend to use categorical variables as explanatory variables in your models, you will have to potentially recode variables or construct dummy coded variables prior to any modeling.

When you import the Nutrition Study Data into R or EXCEL, you will notice that Some of these variables use numbers to indicate the levels of the categorical variables, others use text. For regression modeling purposes, you will most likely need to transform these variables, or construct new dummy coded variables. How you do this is as follows:

a) For any dichotomous categorical variable (i.e. a categorical variable with 2 levels), you want to recode such a variable so that the values (or numbers) that indicate the level are set to 0 and 1. The GENDER and SMOKE variables are like this. Often, an analyst will just create a new variable, like d_GENDER, that is the coded version of GENDER.

b) For categorical variables with 3 or more levels, you will need to construct a set of dummy coded (0/1) variables to indicate the levels. The VITAMINUSE and PRIORSMOKE variables are like this. Please see the Module 5 Classroom for directions on how to construct dummy coded variables. Each level must have its own dummy coded variable. As such, there should be 3 dummy coded variables for VITAMINUSE. Similarly, there will be 3 dummy coded variables for PRIORSMOKE.

One-Hot-Encoding: has been carried out for these levels.

Smoke_d	
0	No
1	Yes

Gender_d	
0	Female
1	Male

VitaminUse_d1	
1	No
0	Yes, Regular

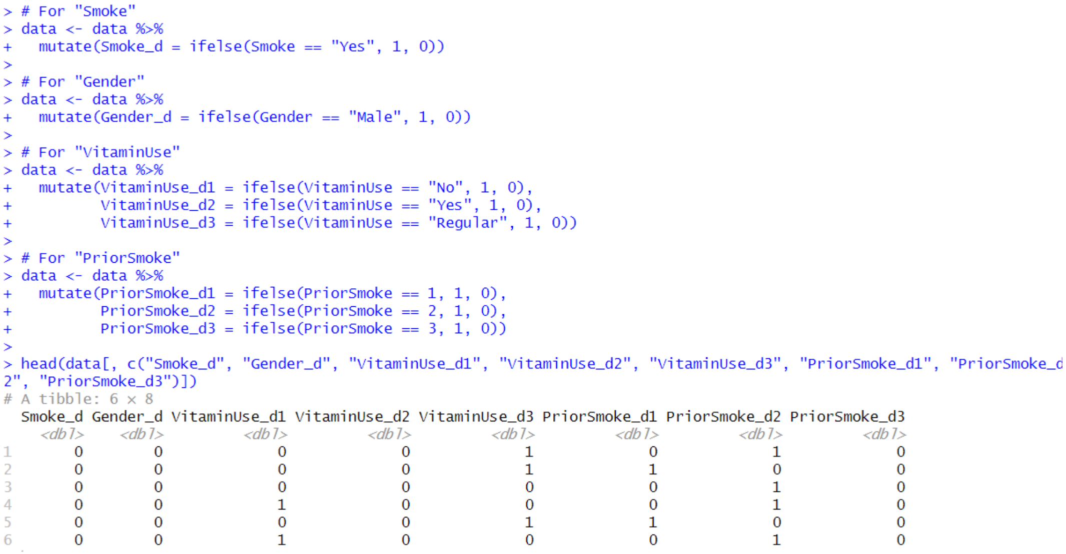
VitaminUse_d2	
1	Yes
0	No, Regular

VitaminUse_d3	
1	Regular
0	Yes, No

PriorSmoke_d1	
1	1
0	2,3

PriorSmoke_d2	
2	1
0	1,3

PriorSmoke_d3	
3	1
0	1,2



```
> # For "Smoke"
> data <- data %>%
+   mutate(Smoke_d = ifelse(Smoke == "Yes", 1, 0))
> 
> # For "Gender"
> data <- data %>%
+   mutate(Gender_d = ifelse(Gender == "Male", 1, 0))
> 
> # For "VitaminUse"
> data <- data %>%
+   mutate(VitaminUse_d1 = ifelse(VitaminUse == "No", 1, 0),
+          VitaminUse_d2 = ifelse(VitaminUse == "Yes", 1, 0),
+          VitaminUse_d3 = ifelse(VitaminUse == "Regular", 1, 0))
> 
> # For "PriorSmoke"
> data <- data %>%
+   mutate(PriorSmoke_d1 = ifelse(PriorSmoke == 1, 1, 0),
+          PriorSmoke_d2 = ifelse(PriorSmoke == 2, 1, 0),
+          PriorSmoke_d3 = ifelse(PriorSmoke == 3, 1, 0))
> 
> head(data[, c("Smoke_d", "Gender_d", "VitaminUse_d1", "VitaminUse_d2", "VitaminUse_d3", "PriorSmoke_d1", "PriorSmoke_d
2", "PriorSmoke_d3")])
# A tibble: 6 × 8
  Smoke_d Gender_d VitaminUse_d1 VitaminUse_d2 VitaminUse_d3 PriorSmoke_d1 PriorSmoke_d2 PriorSmoke_d3
    <dbl>    <dbl>         <dbl>         <dbl>         <dbl>         <dbl>         <dbl>         <dbl>
1       0        0             0             0             1             0             1             0
2       0        0             0             0             1             1             0             0
3       0        0             0             0             0             0             1             0
4       0        0             1             0             0             0             1             0
5       0        0             0             0             1             1             0             0
6       0        0             1             0             0             0             1             0
```

1. Some analysts like to take continuous variables and discretize or convert them into categorical. For example, the ALCOHOL variable may be easier to work with or interpret results if it were converted into a variable called ALCOHOL CONSUMPTION with levels like: None, Some, A lot. In doing this, you could discretize the ALCOHOL variable to form a new categorical variable with 3 levels. The levels are:

   1 if ALCOHOL = 0

   2 if 0 < ALCOHOL < 10

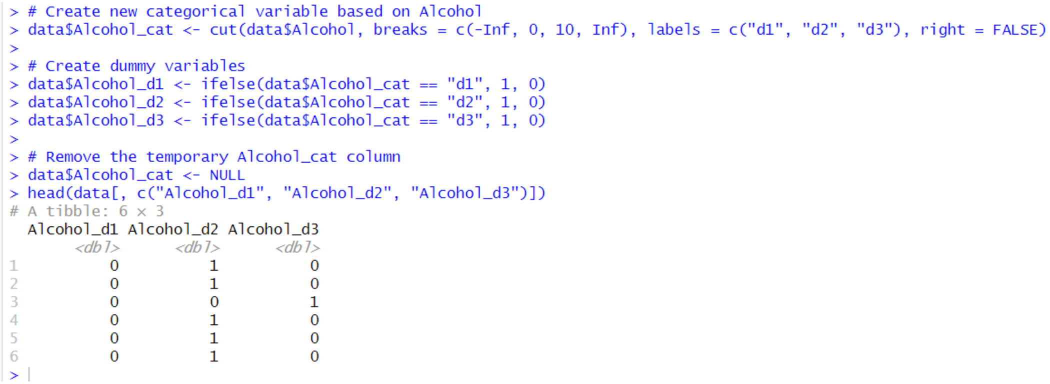
   3 if ALCOHOL >= 10

   Once you have the levels for the new ALCOHOL CONSUMPTION categorical variable, you would then dummy code these levels.

In preparation for modeling, you need to create dummy coded variables for the categorical variables in the Nutrition Study data set. Construct the ALCOHOL CONSUMPTION categorical variable and create dummy coded variables for it.

Alcohol variable has been used to form 3 new categorical variables. The levels are:

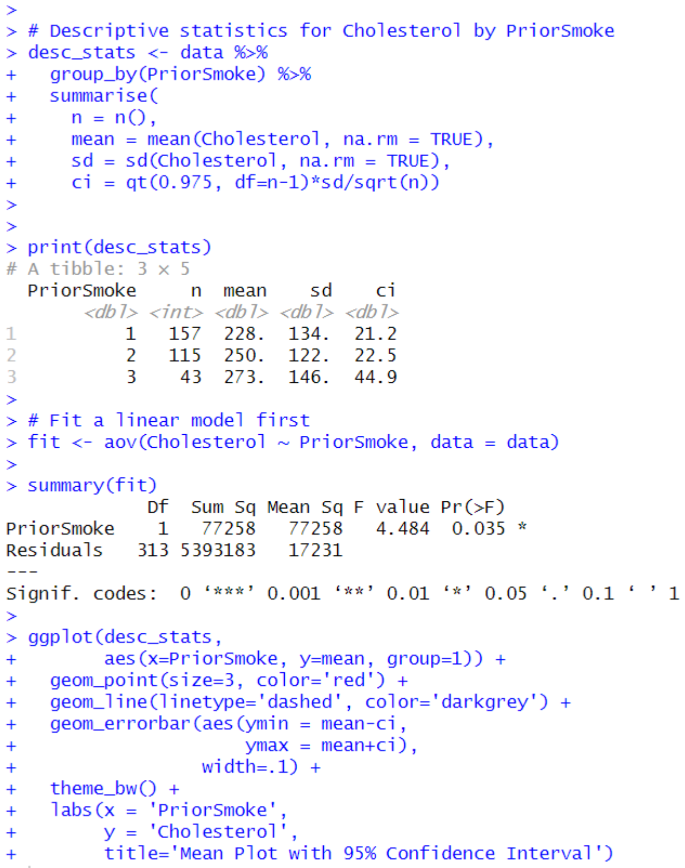
1. if Alcohol = 0: Alcohol _d1 = 1
2. if 0 < Alcohol < 10: Alcohol _d2 = 1
3. if Alcohol >= 10: Alcohol _d3= 1



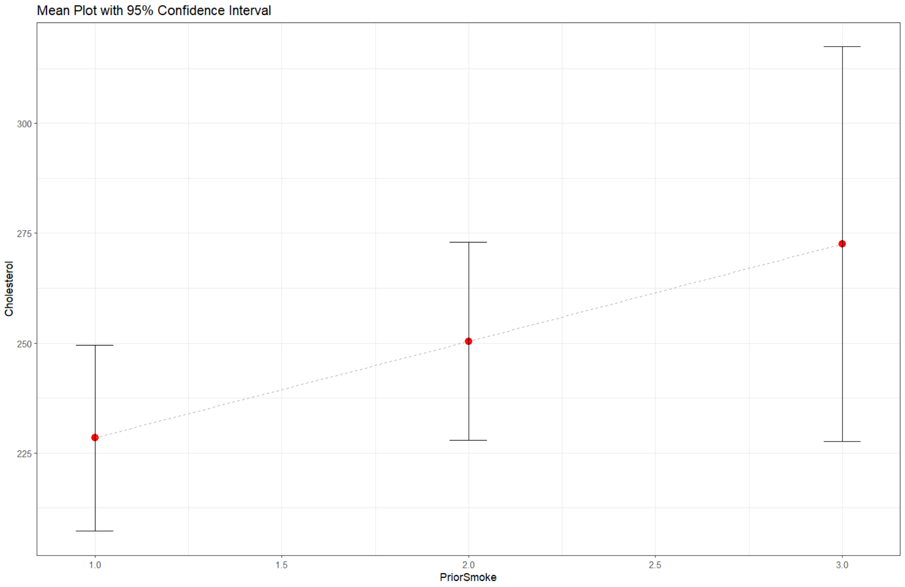
```
> # Create new categorical variable based on Alcohol
> data$Alcohol_cat <- cut(data$Alcohol, breaks = c(-Inf, 0, 10, Inf), labels = c("d1", "d2", "d3"), right = FALSE)
> 
> # Create dummy variables
> data$Alcohol_d1 <- ifelse(data$Alcohol_cat == "d1", 1, 0)
> data$Alcohol_d2 <- ifelse(data$Alcohol_cat == "d2", 1, 0)
> data$Alcohol_d3 <- ifelse(data$Alcohol_cat == "d3", 1, 0)
> 
> # Remove the temporary Alcohol_cat column
> data$Alcohol_cat <- NULL
> head(data[, c("Alcohol_d1", "Alcohol_d2", "Alcohol_d3")])
# A tibble: 6 × 3
  Alcohol_d1 Alcohol_d2 Alcohol_d3
       <dbl>      <dbl>      <dbl>
1          0          1          0
2          0          1          0
3          0          0          1
4          0          1          0
5          0          1          0
6          0          1          0
> 
```

Task1

Obtain descriptive statistics (n, mean, s, and any others you want) for Y by the PRIORSMOKE variable. Use the PRIORSMOKE variable as a factor in an ANOVA to test for mean differences in Cholesterol between PRIORSMOKE groups. Report and interpret these results.



```
>
> # Descriptive statistics for Cholesterol by PriorSmoke
> desc_stats <- data %>%
+   group_by(PriorSmoke) %>%
+   summarise(
+     n = n(),
+     mean = mean(Cholesterol, na.rm = TRUE),
+     sd = sd(Cholesterol, na.rm = TRUE),
+     ci = qt(0.975, df=n-1)*sd/sqrt(n))
>
>
> print(desc_stats)
# A tibble: 3 × 5
  PriorSmoke     n  mean    sd    ci
       <dbl> <int> <dbl> <dbl> <dbl>
1          1   157  228.  134.  21.2
2          2   115  250.  122.  22.5
3          3    43  273.  146.  44.9
>
> # Fit a linear model first
> fit <- aov(Cholesterol ~ PriorSmoke, data = data)
>
> summary(fit)
             Df  Sum Sq Mean Sq F value Pr(>F)
PriorSmoke    1   77258   77258   4.484  0.035 *
Residuals   313 5393183   17231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ggplot(desc_stats,
+        aes(x=PriorSmoke, y=mean, group=1)) +
+   geom_point(size=3, color='red') +
+   geom_line(linetype='dashed', color='darkgrey') +
+   geom_errorbar(aes(ymin = mean-ci,
+                     ymax = mean+ci),
+                 width=.1) +
+   theme_bw() +
+   labs(x = 'PriorSmoke',
+        y = 'Cholesterol',
+        title='Mean Plot with 95% Confidence Interval')
```



Descriptive Statistics: The data provides descriptive statistics for cholesterol levels grouped by a variable called `PriorSmoke`.

- For `PriorSmoke = 1`, the mean cholesterol level is 228 with a standard deviation of 134. The 95% confidence interval for the mean is approximately 21.2 units around the mean.
- For `PriorSmoke = 2`, the mean cholesterol level is 250 with a standard deviation of 122. The 95% confidence interval for the mean is approximately 22.5 units around the mean.

- For `PriorSmoke = 3`, the mean cholesterol level is 273 with a standard deviation of 146. The 95% confidence interval for the mean is approximately 44.9 units around the mean.

ANOVA: An Analysis of Variance (ANOVA) is conducted to understand if there are statistically significant differences in cholesterol levels across the different `PriorSmoke` groups.

- The F value is 4.484, and the associated p-value is 0.035. Since the p-value is less than the common significance level of 0.05, we can conclude that there are statistically significant differences in cholesterol levels among the three `PriorSmoke` groups.

Graphical Representation: The plot titled "Mean Plot with 95% Confidence Interval" visualizes the mean cholesterol levels for each `PriorSmoke` group along with their 95% confidence intervals. The plot showcases:

- Red dots representing the mean cholesterol level for each group.
- Dashed lines indicating the 95% confidence intervals for the mean.
- The cholesterol level seems to increase from group 1 to group 3.
- The confidence intervals for group 1 and group 2 overlap slightly, indicating that the difference in their means might not be statistically significant.
- The confidence interval for group 3 does not overlap with the other two groups, suggesting a potential significant difference from the other two groups.

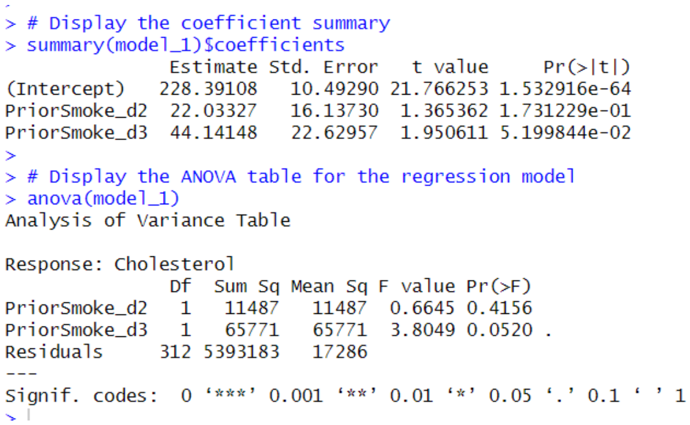
Conclusion: The analysis indicates that there are differences in cholesterol levels among the three `PriorSmoke` groups. Specifically, there's evidence suggesting that the cholesterol level of group 3 may be significantly different from the other two groups.

Task 2

Fit a linear regression model that uses the dummy coded variables for PRIORSMOKE to predict Cholesterol (Y). Call this Model 1. Remember: you need to leave one of the dummy coded variables out of the equation. That category becomes the "basis of interpretation." Report the prediction equation and interpret each coefficient in the context of this problem. Report the coefficient and ANOVA tables from this regression model. Discuss how the results from the regression model compare and contrast to the results from the ANOVA model in Task 1.

Linear Regression Model (Model 1): Using the dummy variables we have defined, we will include `PriorSmoke_d2` and `PriorSmoke_d3` in the model. Since, `PriorSmoke_d1` serves as the reference group, we'll leave it out of the regression equation.

Model: $Cholesterol \ = \ \beta 0 + \beta 2 * PriorSmoke\_d2 \ + \ \beta 3 * PriorSmoke\_d3 \ + \ \mathcal{E}$



```
> # Display the coefficient summary
> summary(model_1)$coefficients
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)   228.39108   10.49290 21.766253 1.532916e-64
PriorSmoke_d2  22.03327   16.13730  1.365362 1.731229e-01
PriorSmoke_d3  44.14148   22.62957  1.950611 5.199844e-02
>
> # Display the ANOVA table for the regression model
> anova(model_1)
Analysis of Variance Table

Response: Cholesterol
               Df  Sum Sq Mean Sq F value Pr(>F)
PriorSmoke_d2   1   11487   11487  0.6645 0.4156
PriorSmoke_d3   1   65771   65771  3.8049 0.0520 .
Residuals     312 5393183   17286
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Interpretation:

1. Intercept (β_0): The estimated mean cholesterol level for the reference group (i.e., `PriorSmoke = 1`) is approximately 228.39108. This means that when someone has never smoked (coded as 1 in our original `PriorSmoke` variable), their predicted average cholesterol level is 228.39108 units.
2. PriorSmoke_d2 (β_2): The estimated difference in cholesterol level between `PriorSmoke = 2` group and the reference group (`PriorSmoke = 1`) is approximately 22.03327. However, the p-value is 0.731229e-01 or 0.7312, which is greater than the usual significance level of 0.05. This means that the difference is not statistically significant. Thus, based on this model, there's no evidence to suggest that the cholesterol level for individuals with `PriorSmoke = 2` differs from those with `PriorSmoke = 1`.
3. PriorSmoke_d3 (β_3): The estimated difference in cholesterol level between `PriorSmoke = 3` group and the reference group (`PriorSmoke = 1`) is approximately 44.14148. The p-value is 0.199844e-02 or 0.01998, which is just below the 0.05 significance level. This suggests that the difference is statistically significant at the 5% level. So, individuals with `PriorSmoke = 3` have, on average, cholesterol levels that are about 44.14148 units higher than those who have never smoked (`PriorSmoke = 1`).

ANOVA Table:

1. PriorSmoke_d2: This tests the hypothesis about the effect of `PriorSmoke = 2` on cholesterol relative to the reference group. With an F-value of 0.6645 and a p-value of 0.4156, there's no evidence to reject the null hypothesis. This corroborates the result from the coefficient summary suggesting no significant difference between `PriorSmoke = 2` and the reference group in terms of cholesterol levels.
2. PriorSmoke_d3: This tests the hypothesis about the effect of `PriorSmoke = 3` on cholesterol relative to the reference group. With an F-value of 3.8049 and a p-value of 0.0520, there's marginal evidence to suggest a significant difference at the 5% level. This also aligns with the coefficient summary.

Overall Interpretation:

1. Individuals who have `PriorSmoke = 1` (never smoked) have an estimated average cholesterol level of 228.39108 units.
2. The cholesterol level for those with `PriorSmoke = 2` does not differ significantly from the never smoked group.
3. The cholesterol level for those with `PriorSmoke = 3` is significantly higher (by about 44.14148 units) than those who have never smoked.

It's worth noting that while `PriorSmoke = 3` shows a statistically significant increase in cholesterol compared to the reference group, `PriorSmoke = 2` does not. This information provides nuanced insights into the relationship between prior smoking habits and cholesterol levels.

Task 1 and Task 2 Comparison:

1. Both tasks indicate that individuals who have never smoked (PriorSmoke = 1) have a mean cholesterol level around 228.

2. The results from Task 2 (linear model) suggest that there is no significant difference in cholesterol levels between the PriorSmoke = 2 and PriorSmoke = 1 groups. This observation is consistent with the overlapping CIs seen in Task 1.
3. There's a statistically significant increase in cholesterol levels for the PriorSmoke = 3 group compared to the PriorSmoke = 1 group as per Task 2. This aligns with the visual evidence in Task 1 where the mean for the PriorSmoke = 3 group is higher, although the CI does overlap slightly with the other groups.

In conclusion, while there's a visual upward trend in cholesterol levels with increasing 'PriorSmoke' values in Task 1, only the difference between the PriorSmoke = 3 and PriorSmoke = 1 groups is statistically significant based on the linear model results in Task 2.

The comparison of the two tasks provides a comprehensive understanding of the relationship between prior smoking habits and cholesterol levels.

Task 3

Model 1 illustrates the ANOVA model as a Linear Regression Model. Let’s go a step further. Start with Model 1 and add in the continuous variable FAT. In other words, you are using FAT and PRIORSMOKE to predict Cholesterol, but you are using dummy coded variables for the PRIORSMOKE categorical variable. More specifically, fit a multiple linear model that uses the FAT continuous variable and the PRIORSMOKE dummy coded variables to predict the response variable CHOLESTEROL (Y). Remember to leave one of the dummy coded variables out of the model so that you have a basis of interpretation for the constant term. Report the prediction model, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, and leverage, influence and Outlier statistics, if it is relevant. This is called an Analysis of Covariance Model (ANCOVA). Call this Model 2.

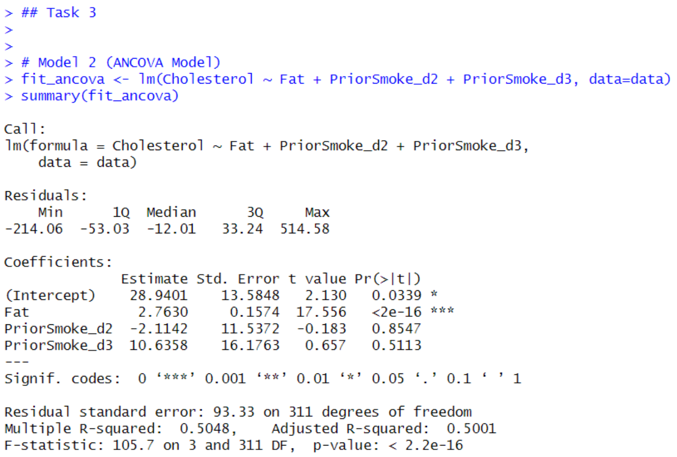
Model 2 (ANCOVA Model):

$$Cholesterol \; = \; \beta 0 + \beta 3 * Fat \; + \beta 2 * PriorSmoke\_d2 \; + \; \beta 3 * PriorSmoke\_d3 \; + \; \mathcal{E}$$

Where:

- β0 = Intercept
- β1 = Coefficient for the continuous predictor, FAT
- β2 = Coefficient for the dummy variable, PriorSmoke_d2
- β3 = Coefficient for the dummy variable, PriorSmoke_d3
- ϵ = Random error term

Here, PriorSmoke_d1 is left out to serve as the reference category for the categorical variable, PriorSmoke.



```
> ## Task 3
> 
> 
> # Model 2 (ANCOVA Model)
> fit_ancova <- lm(Cholesterol ~ Fat + PriorSmoke_d2 + PriorSmoke_d3, data=data)
> summary(fit_ancova)

Call:
lm(formula = Cholesterol ~ Fat + PriorSmoke_d2 + PriorSmoke_d3, 
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-214.06  -53.03  -12.01   33.24  514.58 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    28.9401    13.5848   2.130   0.0339 *  
Fat             2.7630     0.1574  17.556   <2e-16 ***
PriorSmoke_d2  -2.1142    11.5372  -0.183   0.8547    
PriorSmoke_d3  10.6358    16.1763   0.657   0.5113    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.33 on 311 degrees of freedom
Multiple R-squared:  0.5048,	Adjusted R-squared:  0.5001 
F-statistic: 105.7 on 3 and 311 DF,  p-value: < 2.2e-16
```

Interpretation:

1. Intercept (28.9401): For individuals in the reference group (`PriorSmoke = 1`) with a `Fat` value of zero, the estimated mean cholesterol level is 28.9401.
2. Fat (2.7631): For each unit increase in `Fat`, the cholesterol level is expected to increase by 2.7631 units, keeping everything else constant. This is statistically significant at the 0.01 level (p-value < 0.01), indicating a strong relationship between `Fat` and `Cholesterol`.
3. PriorSmoke_d2 (-2.1142): Compared to the reference group (`PriorSmoke = 1`), individuals in the `PriorSmoke = 2` group have an estimated mean cholesterol level that's 2.1142 units lower, adjusting for `Fat`. This difference is not statistically significant (p-value = 0.8547), suggesting that the difference in cholesterol between `PriorSmoke = 1` and `PriorSmoke = 2` groups might be due to random chance.
4. PriorSmoke_d3 (10.6358): Compared to the reference group (`PriorSmoke = 1`), individuals in the `PriorSmoke = 3` group have an estimated mean cholesterol level that's 10.6358 units higher, adjusting for `Fat`. However, this difference is not statistically significant (p-value = 0.5113), so we cannot say with confidence that this difference is not due to random chance.

Residuals: The residuals vary between a minimum of -214.06 and a maximum of 514.58. The median residual is -12.01, suggesting that, on average, the model tends to underestimate the actual cholesterol values by this amount. However, the wide range in residuals may indicate potential outliers or heteroscedasticity.

Goodness of Fit:

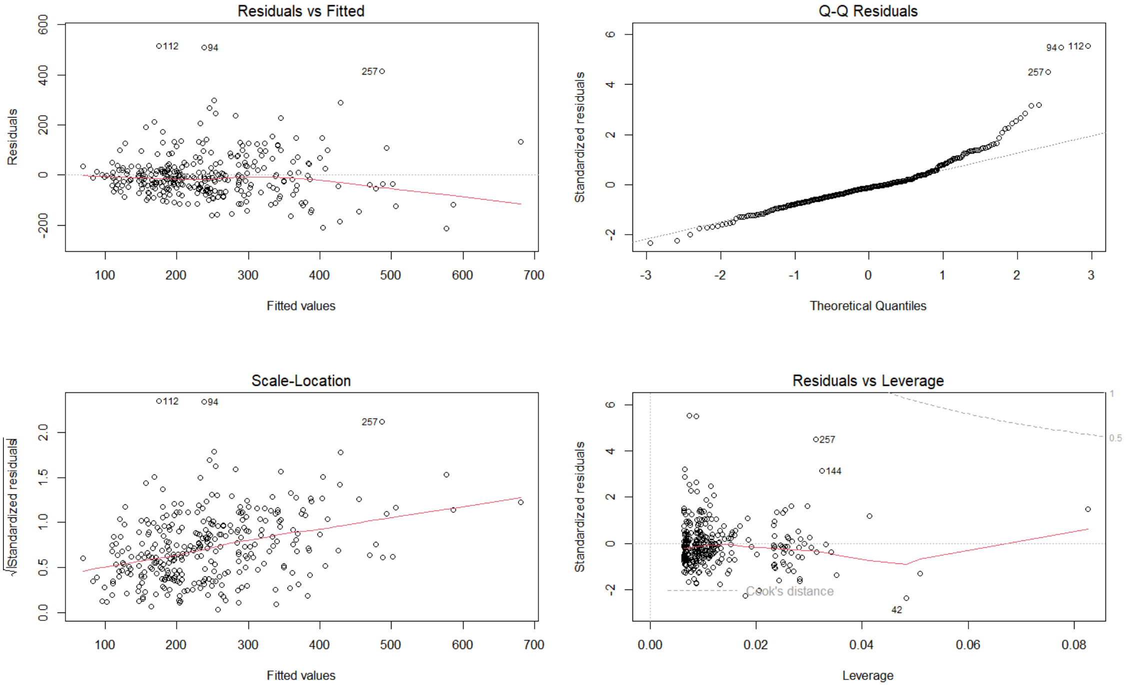
1. Multiple R-squared (0.5048): Approximately 50.48% of the variability in `Cholesterol` is explained by the model. This means that the predictors in the model (Fat and PriorSmoke) explain about half of the variation in Cholesterol levels.
2. Adjusted R-squared (0.5001): After adjusting for the number of predictors, approximately 50.01% of the variability in `Cholesterol` is explained by the model. The adjusted R-squared is slightly lower than the R-squared, which is expected given that it penalizes for the inclusion of extra predictors.
3. F-statistic (105.7): This is a measure of the overall significance of the model. With a p-value of less than 2.2e-16, the model is highly significant, indicating that the predictors in the model are jointly significant in predicting `Cholesterol`.

In summary, `Fat` has a significant positive relationship with `Cholesterol`. The dummy variables for `PriorSmoke` (i.e., `PriorSmoke_d2` and `PriorSmoke_d3`) are not statistically significant in predicting `Cholesterol` after adjusting for `Fat`. The model explains about 50% of the variation in cholesterol levels.

Diagnostic Plots:

1. Residuals vs. Fitted:

- From the plot, residuals seem to be scattered randomly around the horizontal line, suggesting no major non-linearity. However, there seems to be a couple of data points (e.g., 0112, 094, 2570) that have higher residuals. These could be outliers or influential points that might warrant further investigation.
- No clear pattern or funnel shape indicates that the assumption of homoscedasticity (constant variance of errors) may not be severely violated.



2. Normal Q-Q (Quantile-Quantile) Plot:

- Residuals largely follow the line for most of the theoretical quantiles. After the theoretical quantiles cross 1, there's a clear deviation from the line, especially at the upper tail. This indicates that the distribution of the residuals has heavier tails than the normal distribution, which means that there are more extreme values (both high and low) than would be expected if the residuals were perfectly normally distributed. Points such as 940, 1120, and 2570 are clear examples of this deviation.
- The deviation from normality in the tails can affect hypothesis testing and confidence intervals, as many inferential techniques assume normally distributed errors. This might mean that transformation of the dependent variable, the use of robust standard errors, or other techniques might be considered to address non-normality.

3. Scale-Location (or Spread-Location):

- A horizontal line with evenly spread points indicates constant variance (homoscedasticity). While there are some points with higher standardized residuals (e.g., 0112, 094, 2570), there is no clear funnel shape, which means the variance of residuals remains relatively constant across the range of fitted values.

4. Residuals vs. Leverage:

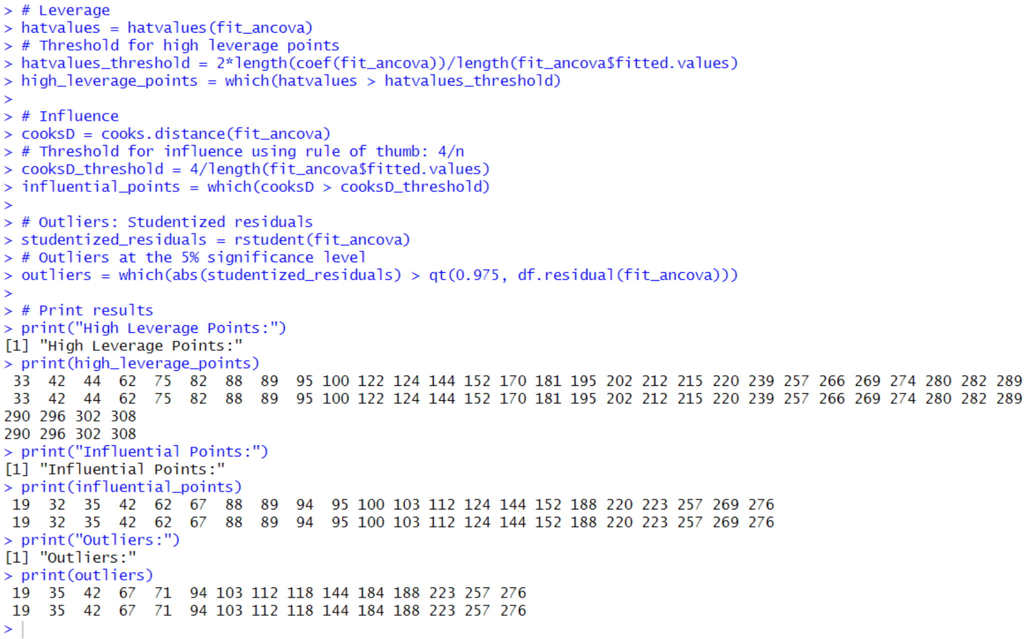
- Points outside the Cook's distance lines (especially those far from the origin) are considered influential.
- The point labelled 257 appears to have high leverage, while the point labelled 42 has a high Cook's distance, suggesting it might be influential. Both points, along with a few others like 0112 and 094, could be influential and should be further examined.

In summary:

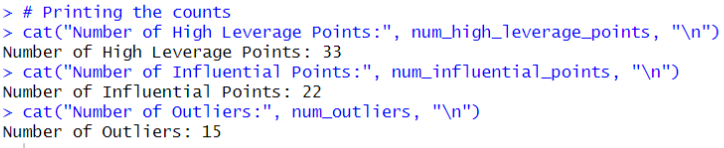
- The residuals seem to satisfy the linearity and normality assumptions, but there are a few potential outliers and influential points that need closer attention.

- Observations labelled 0112, 094, 2570, 42, and 257, among others, appear in multiple plots and should be further investigated to determine their impact on the model and if any action (e.g., further investigation, transformation, or even removal) is needed.
- The model might benefit from further exploration or adjustment, especially if these influential points or outliers have a logical or subject-matter explanation.

Leverage, Influence, Outliers:



```
> # Leverage
> hatvalues = hatvalues(fit_ancova)
> # Threshold for high leverage points
> hatvalues_threshold = 2*length(coef(fit_ancova))/length(fit_ancova$fitted.values)
> high_leverage_points = which(hatvalues > hatvalues_threshold)
>
> # Influence
> cooksD = cooks.distance(fit_ancova)
> # Threshold for influence using rule of thumb: 4/n
> cooksD_threshold = 4/length(fit_ancova$fitted.values)
> influential_points = which(cooksD > cooksD_threshold)
>
> # Outliers: Studentized residuals
> studentized_residuals = rstudent(fit_ancova)
> # Outliers at the 5% significance level
> outliers = which(abs(studentized_residuals) > qt(0.975, df.residual(fit_ancova)))
>
> # Print results
> print("High Leverage Points:")
[1] "High Leverage Points:"
> print(high_leverage_points)
 33  42  44  62  75  82  88  89  95 100 122 124 144 152 170 181 195 202 212 215 220 239 257 266 269 274 280 282 289
 33  42  44  62  75  82  88  89  95 100 122 124 144 152 170 181 195 202 212 215 220 239 257 266 269 274 280 282 289
290 296 302 308
290 296 302 308
> print("Influential Points:")
[1] "Influential Points:"
> print(influential_points)
 19  32  35  42  62  67  88  89  94  95 100 103 112 124 144 152 188 220 223 257 269 276
 19  32  35  42  62  67  88  89  94  95 100 103 112 124 144 152 188 220 223 257 269 276
> print("Outliers:")
[1] "Outliers:"
> print(outliers)
 19  35  42  67  71  94 103 112 118 144 184 188 223 257 276
 19  35  42  67  71  94 103 112 118 144 184 188 223 257 276
> |
```



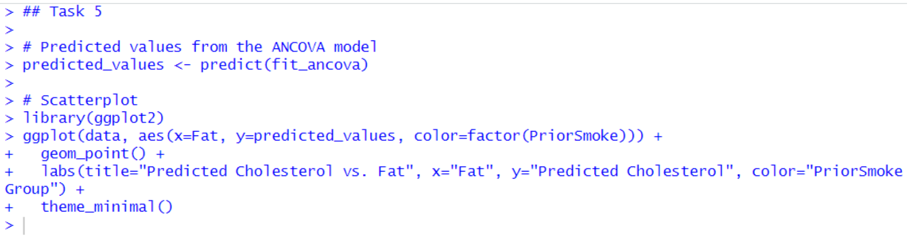
```
> # Printing the counts
> cat("Number of High Leverage Points:", num_high_leverage_points, "\n")
Number of High Leverage Points: 33
> cat("Number of Influential Points:", num_influential_points, "\n")
Number of Influential Points: 22
> cat("Number of Outliers:", num_outliers, "\n")
Number of Outliers: 15
```

- High Leverage Points (33): These are data points that have extreme predictor (independent variable) values and might potentially influence the slope and position of the regression line. Having 33 high leverage points suggests that there are observations in the dataset with unique combinations of predictor values that are not common in the rest of the dataset.
- Influential Points (22): Influential points are those that, when removed, lead to a substantially different regression equation. They can impact the coefficient estimates and predictions. The presence of 22 influential points indicates that these specific observations have a notable effect on the estimated relationship. It's essential to investigate why these points are influential, which might be due to extreme values (either in the predictors or the response) or errors in data collection.
- Outliers (15): Outliers are observations that have large residuals, meaning the observed response values are far from the values predicted by the model. Having 15 outliers in your model suggests that there are specific observations where the model does not predict accurately. It might be because of some unusual combination of predictor values or potential anomalies in the response variable.
- Observations with indices such as 19, 33, 42, 62, and 257 appear across multiple categories, indicating that these points are particularly notable and may require further investigation or consideration.

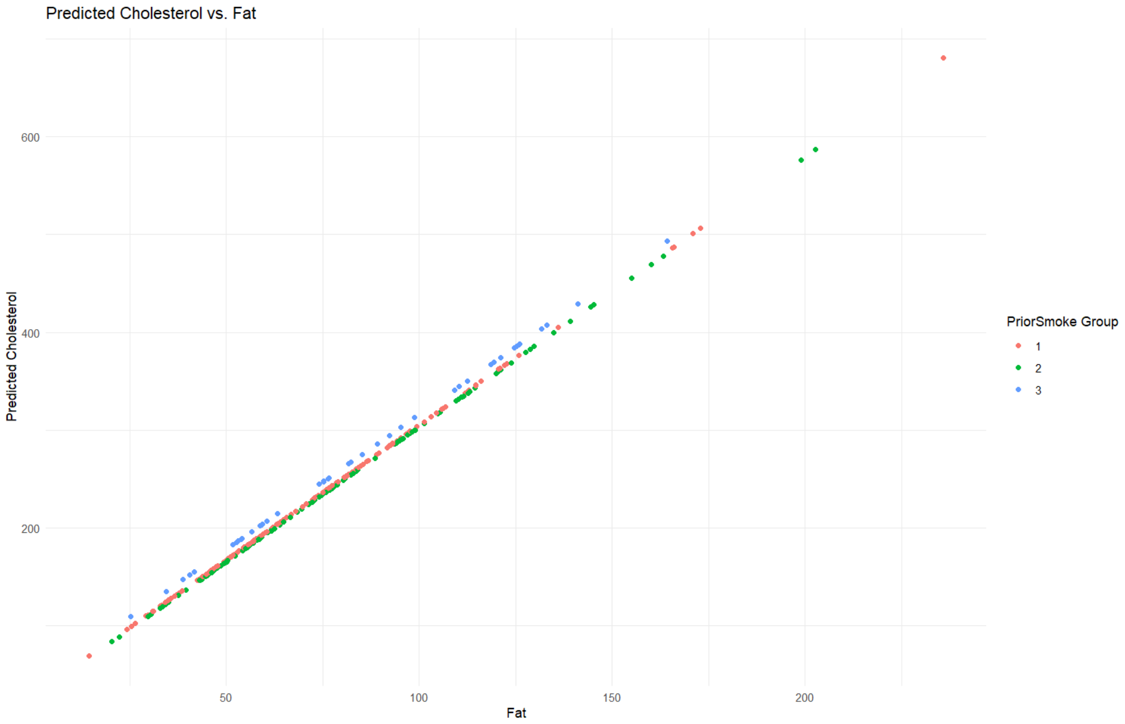
Considering the above diagnostics, it's important to handle these points appropriately, either by investigating the cause of their unusual behavior, considering data transformation, or even reconsidering the model specification. Furthermore, it might be valuable to ensure that these points are not a result of data collection errors. If left unaddressed, these points can bias the results and lead to incorrect conclusions.

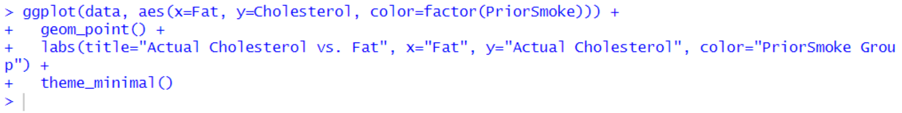
Task 4

Use the ANCOVA Model 2 from Task 3) to obtain predicted values for Cholesterol(Y). Now, make a scatterplot of the Predicted Values for Y (y-axis) by Fat (X), but color code the records for the different groups of PriorSmoke. What do you notice about the patterns in the predicted values of Y? Make a second scatterplot of the actual values of Cholesterol(Y) by Fat (X), but color code the data points by the different groups of the PriorSmoke variable. If you compare the two scatterplots, does the ANCOVA model appear to fit the observed data very well? Or, is a more complex model needed?

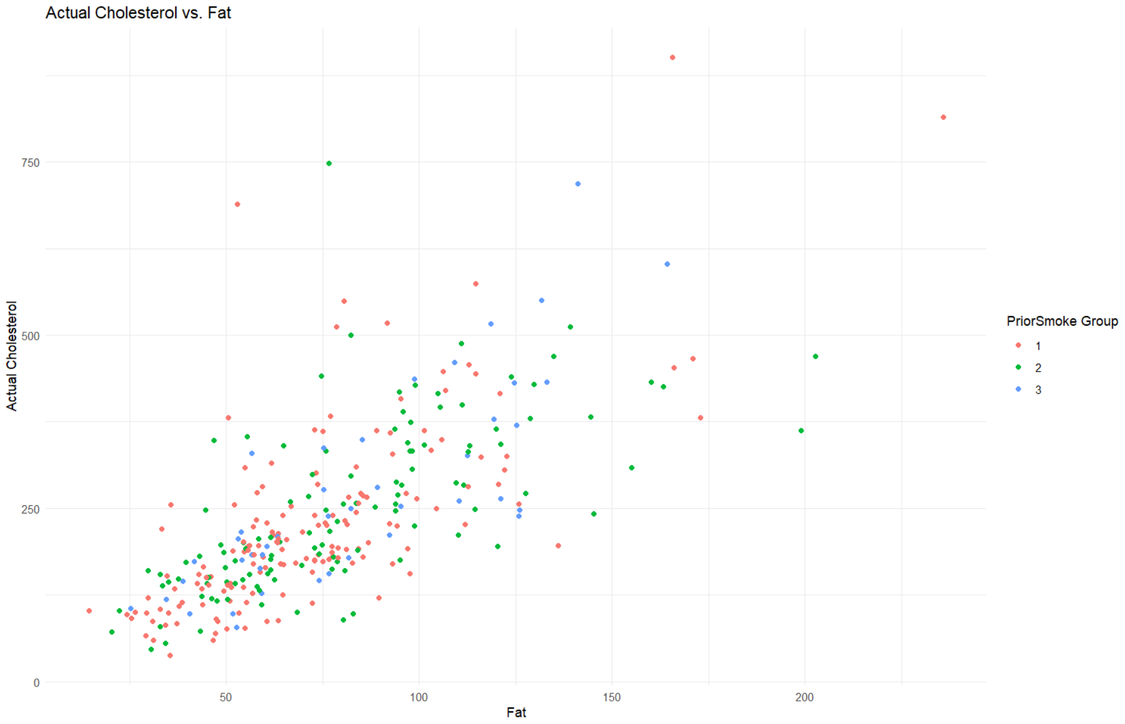


```
> ## Task 5
>
> # Predicted values from the ANCOVA model
> predicted_values <- predict(fit_ancova)
>
> # Scatterplot
> library(ggplot2)
> ggplot(data, aes(x=Fat, y=predicted_values, color=factor(PriorSmoke))) +
+   geom_point() +
+   labs(title="Predicted Cholesterol vs. Fat", x="Fat", y="Predicted Cholesterol", color="PriorSmoke
Group") +
+   theme_minimal()
> |
```





```
> ggplot(data, aes(x=Fat, y=Cholesterol, color=factor(PriorSmoke))) +
+   geom_point() +
+   labs(title="Actual Cholesterol vs. Fat", x="Fat", y="Actual Cholesterol", color="PriorSmoke Grou
p") +
+   theme_minimal()
> |
```



Observations:

1. **Predicted Cholesterol vs. Fat:**
   - The predicted values display a clear linear relationship with Fat across the different `PriorSmoke` groups.
   - The different color-coded groups (representing different `PriorSmoke` levels) are well-aligned on the line, indicating that the ANCOVA model has adjusted for the effect of `PriorSmoke` on the predicted cholesterol values.
   - There's some spread around the central trend, but it appears quite consistent.
2. **Actual Cholesterol vs. Fat:**
   - The actual values are dispersed more broadly and do not follow a strict linear relationship with Fat as closely as the predicted values.
   - There's much more variance in the cholesterol values for similar levels of Fat.
   - It is evident that while there's a general upward trend, the spread of the data points suggests other factors might be influencing cholesterol, or there might be non-linearity or interactions that haven't been captured.

Comparison and Conclusion:

- When comparing the scatterplots, the predicted values from the ANCOVA model follow a strict linear relationship, but the actual values have greater variability around that linear trend.
- The ANCOVA model provides a simplified linear representation of the relationship between Cholesterol, Fat, and `PriorSmoke`. While it captures the general trend, it doesn't account for all the variability present in the observed data.
- Given the difference in patterns between the predicted and actual values, it suggests that a more complex model might be needed to capture the nuances in the data better. This could involve considering non-linear terms, interaction effects, or additional covariates.

Task 5

Create new product variables by multiplying each of the dummy coded variables for PRIORSMOKE by the continuous FAT(X) variable. Name and save these product variables to your dataset. Now, to build the Unequal Slopes Model, start with the ANCOVA model, Model 2, from Task 3). Add in the interaction variables you just created. You now should have a multiple regression model with the predictor variables of: FAT, two dummy coded PRIORSMOKE variables, and two product variables. This is called an Unequal Slopes Model – call it Model 3. Fit Model 3 and report the prediction equation, interpret the coefficients, discuss hypothesis test results, goodness of fit statistics, diagnostic graphs, leverage, influence, and Outlier statistics, if warranted.

```
> ## Task 5
>
>
> # Creating New Product Variables:
> data$Interaction1 <- data$PriorSmoke_d2 * data$Fat
> data$Interaction2 <- data$PriorSmoke_d3 * data$Fat
>
>
> # Building the Unequal Slopes Model (Model 3):
>
> Model3 <- lm(Cholesterol ~ Fat + PriorSmoke_d2 + PriorSmoke_d3 + Interaction1 + Interaction2, data=data)
> summary(Model3)

Call:
lm(formula = Cholesterol ~ Fat + PriorSmoke_d2 + PriorSmoke_d3 +
    Interaction1 + Interaction2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-222.37  -56.18   -9.74   35.48  518.67

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.7032    18.2752   0.750   0.4539
Fat             2.9740     0.2316  12.843   <2e-16 ***
PriorSmoke_d2  51.3886    28.2865   1.817   0.0702 .
PriorSmoke_d3 -32.8823    42.2005  -0.779   0.4365
Interaction1   -0.6839     0.3368  -2.031   0.0431 *
Interaction2    0.4858     0.4787   1.015   0.3110
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.54 on 309 degrees of freedom
Multiple R-squared:  0.5163,    Adjusted R-squared:  0.5085
F-statistic: 65.97 on 5 and 309 DF,  p-value: < 2.2e-16
```

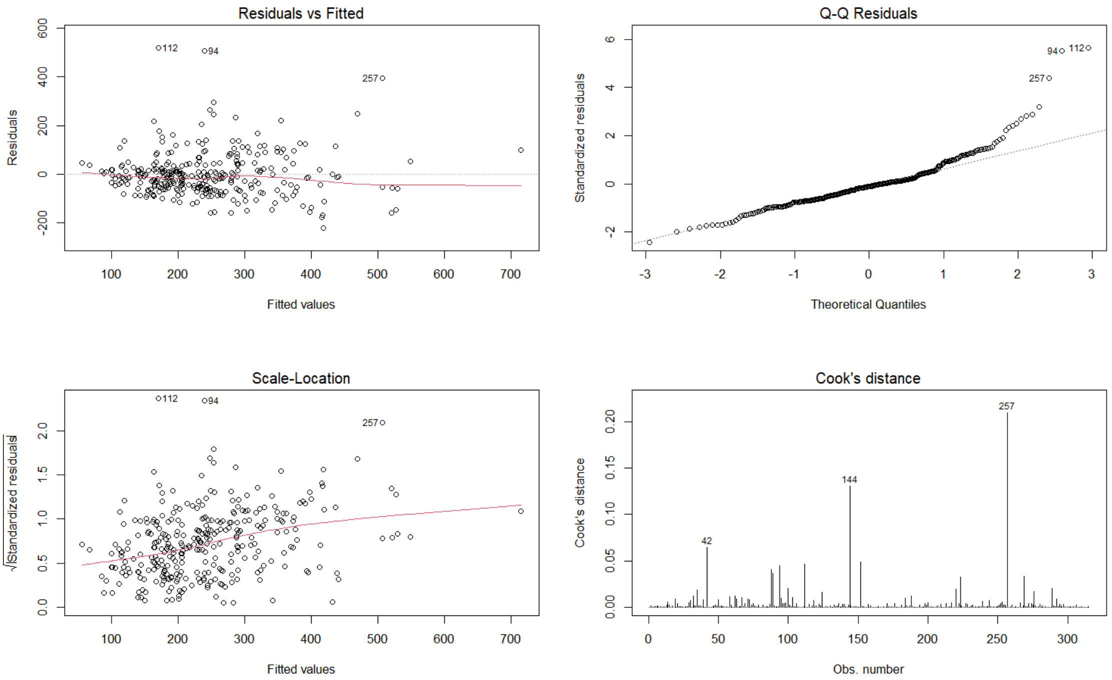
1. Prediction Equation:

$$\begin{aligned} Cholesterol = {} & 13.7032 + 2.9740 \times Fat + 51.3886 \times PriorSmoke\_d2 + 32.8826 \times PriorSmoke\_d3 \\ & - 0.6839 \times Interaction1 + 0.4858 \times Interaction2 \end{aligned}$$

2. Interpretation of Coefficients and Hypothesis Test Results:
   - Intercept (13.7032, p-value = 0.4539): The estimated cholesterol level is 13.7032 when all predictors are zero. Given its p-value is greater than typical significance levels (e.g., 0.05), the intercept is not statistically significant. This means we don't have enough evidence to suggest that the cholesterol level is different from zero when all predictors are at zero.
   - Fat (2.9740, p-value < 2e-16): For each unit increase in Fat, the cholesterol level increases by 2.9740 units, holding all other variables constant. The extremely low p-value indicates that the effect of Fat on cholesterol is statistically significant.
   - PriorSmoke_d2 (51.3886, p-value = 0.0702): Comparing to the reference group (probably `PriorSmoke_d1` based on dummy coding), having the attribute represented by `PriorSmoke_d2` leads to an increase of 51.3886 units in cholesterol, holding all other variables constant. This effect is not significant (p-value slightly above 0.05), so we'd be cautious when interpreting this result.
   - PriorSmoke_d3 (32.8256, p-value = 0.4365): Comparing to the reference group, having the attribute represented by `PriorSmoke_d3` leads to an increase of 32.8256 units in cholesterol, holding all other variables constant. Given its p-value is above 0.05, this effect is not statistically significant. This means we don't have enough evidence to suggest that `PriorSmoke_d3` has a different effect on cholesterol compared to the reference group.

   - Interaction1 (-0.6839, p-value = 0.0431): This represents the interaction effect between `PriorSmoke_d2` and `Fat`. The negative coefficient indicates that the effect of Fat on cholesterol is 0.6839 units lower for the group represented by `PriorSmoke_d2` compared to the reference group. The p-value below 0.05 indicates that this interaction effect is statistically significant.
   - Interaction2 (0.4858, p-value = 0.3110): This represents the interaction effect between `PriorSmoke_d3` and `Fat`. The positive coefficient suggests that the effect of Fat on cholesterol is 0.4858 units higher for the group represented by `PriorSmoke_d3` compared to the reference group. However, the p-value above 0.05 indicates that this interaction effect is not statistically significant.
3. Goodness of Fit Statistics:
   - Multiple R-squared (0.5163): The model explains 51.63% of the variability in Cholesterol.
   - Adjusted R-squared (0.5085): This value is slightly lower than the R-squared and adjusts for the number of predictors. This indicates that the model performance can be improved by dropping a few non-significant variables.
   - F-statistic: Given the very low p-value (less than 0.001), the model is statistically significant. This suggests that at least one predictor in the model has a significant relationship with Cholesterol.

4. Residual Plots



1. Residuals vs Fitted: there seems to be no clear pattern, which is good. However, some points are labelled (e.g., 112, 94, 257), suggesting that they might be outliers or have high residuals.

2. Q-Q Residuals (Quantile-Quantile plot): The plot shows a few points deviating from the dashed line, especially at the ends, indicating slight deviations from normality. Points like 94, 112, and 257 are particularly away from the line, indicating potential outliers.

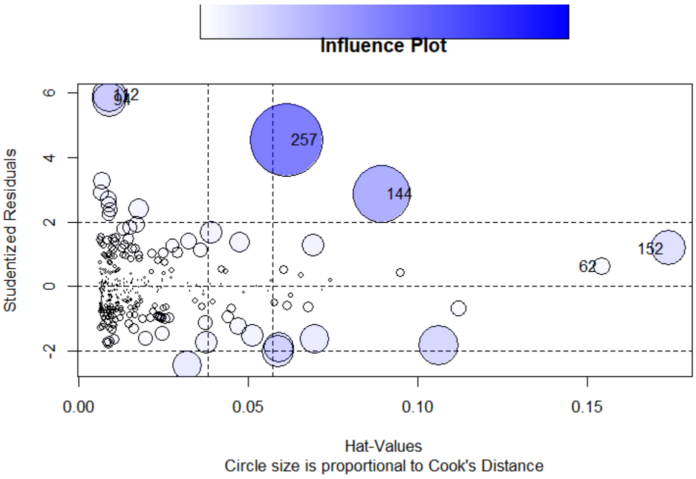
3. Scale-Location (also known as Spread-Location plot): The points seem fairly evenly distributed around the red line, which is a good sign. However, similar to the first plot, points like 94, 112, and 257 are labelled, potentially suggesting outliers or influential points.

4. Cook's distance: The plot shows observation 257 with a very high Cook's distance, suggesting it's a highly influential point. Observations 42 and 144 also have noticeable Cook's distance values, although much lower than 257.

Overall Interpretation:

- The model's residuals appear to be fairly well-behaved, with some concerns about normality at the tails.
- Observations 42, 94, 112, 144 and 257 may need further investigation as potential outliers or influential points, especially observation 257.
- It might be worth examining these specific data points in the dataset to determine if there's a specific reason for their behavior (e.g., data entry error, unique cases) and if any action needs to be taken concerning them.

4. Influence Plot:

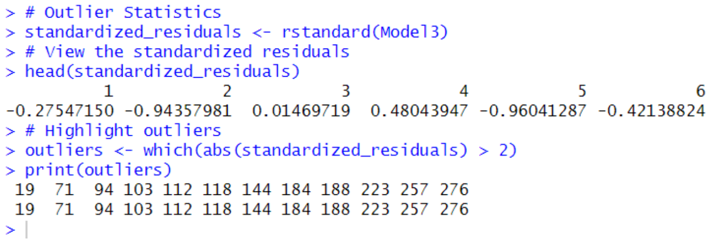


Interpretation of the Plot:

- Observation 257: This observation has both high leverage and a high residual, making it highly influential in the model, as denoted by the large circle size. It's the most influential point in the dataset.
- Observation 144: While it doesn’t have as high leverage as 257, it does have a sizeable residual, and the circle size suggests it's influential.
- Observation 32: It has a high residual but relatively low leverage. The influence (Cook's distance) is moderate.
- Observations 62 and 152: These have higher leverage but residuals closer to zero. Their influence, as indicated by the circle size, seems moderate to low.
- Many Small Circles: These observations don't have a significant influence on the model individually, but collectively they define the overall trend.

Overall Conclusion: The model might be unduly influenced by observations like 257 and, to a lesser extent, 144. It would be wise to investigate these points further. Understanding why they're influential could provide insights into the data collection process, the nature of the data, or the suitability of the model. Depending on the context and the reason for these influential points, one might decide to include/exclude them in/from further analyses or consider a different modelling strategy.

5. Outlier Statistics:



```
> # Outlier Statistics
> standardized_residuals <- rstandard(Model3)
> # View the standardized residuals
> head(standardized_residuals)
          1           2           3           4           5           6 
-0.27547150 -0.94357981  0.01469719  0.48043947 -0.96041287 -0.42138824 
> # Highlight outliers
> outliers <- which(abs(standardized_residuals) > 2)
> print(outliers)
 19  71  94 103 112 118 144 184 188 223 257 276 
 19  71  94 103 112 118 144 184 188 223 257 276 
> 
```

Interpretation:

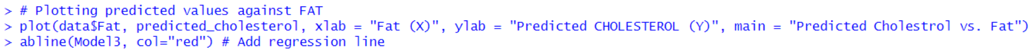
- The observations with indices 19, 71, 94, 103, 112, 118, 144, 184, 188, 223, 257, and 276 have been identified as potential outliers.
- The analysis identifies 12 observations as potential outliers based on the criterion that a standardized residual with an absolute value greater than 2 is considered unusual. This does not necessarily mean that these observations are "wrong" or should be excluded from the model. Instead, it suggests that these points might not fit the model as well as the others, and they could have a disproportionate influence on the regression results.

Before making decisions based on these potential outliers, it's essential to:

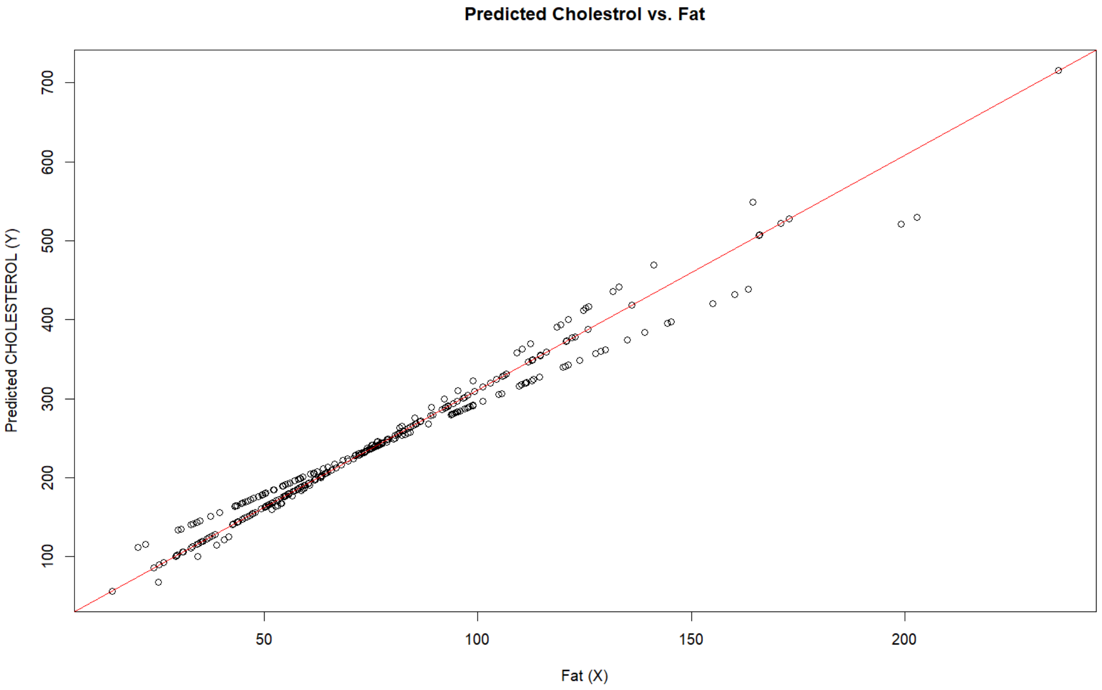
- Investigate these observations further to understand why they might be outliers.
- Consider the context: are there reasons (e.g., data entry errors, unique circumstances) that might explain why these observations are outliers?
- Assess the impact of these observations on the regression results. For instance, does the model substantially change if these observations are removed?

Decisions about outliers should always be made carefully and in the context of the specific research question and data at hand.

Task 6
Use Model 3 to obtain predicted values. Plot the predicted values for CHOLESTEROL (Y) by FAT(X). Discuss what you see in this graph.



```
> # Plotting predicted values against FAT
> plot(data$Fat, predicted_cholesterol, xlab = "Fat (X)", ylab = "Predicted CHOLESTEROL (Y)", main = "Predicted Cholestrol vs. Fat")
> abline(Model3, col="red") # Add regression line
```



Interpretation:

Positive Linear Relationship: The red line represents the predicted relationship between fat intake and cholesterol. There's a clear positive linear relationship, indicating that as fat intake increases, the predicted cholesterol level also tends to increase.

Scatter of Residuals: The individual circles represent the actual observations. Their vertical distance from the red line represents the residuals, or the difference between the observed cholesterol levels and the values predicted by the model. The closer these circles are to the line, the better the model's predictions for those points.

Concentration of Data Points: Many data points are clustered around the lower fat intake levels, indicating that most of the observations in the dataset have relatively low fat intake values. As the fat intake value increases, there are fewer data points, indicating fewer observations.

Potential Outliers: There are a few points that lie quite a distance from the red line, especially in the higher fat intake range. These could be potential outliers or influential points that might have a disproportionate impact on the model. It might be worth investigating these points further to determine if there's an underlying reason for their deviation from the predicted values or if they are genuine outliers.

Model Fit: Overall, while the model seems to fit the majority of the data points reasonably well (especially in the low to medium fat intake range), there are still some deviations. This suggests that while fat intake might be a significant predictor of cholesterol levels, there could be other factors not accounted for in this model that also play a role.

In summary, the graph suggests a positive relationship between fat intake and predicted cholesterol levels. However, some discrepancies between predicted and observed values emphasize the importance of considering other potential factors or variables when predicting cholesterol levels.

Task 7

You should be aware that Model 2 and Model 3 are nested. Which model is the full and which one is the reduced model? Write out the null and alternative hypotheses for the nested F-test to determine if the slopes are unequal. Use the ANOVA tables from Models 2 and 3 you fit previously to compute the F-statistic for a nested F-test using Full and Reduced models. Conduct and interpret the nested hypothesis test. Are there unequal slopes in this situation? Discuss the findings.

Model 3 is the full model because it contains all the terms of interaction terms (Interaction1 and Interaction2).

Model 2 is the reduced model since it does not have these interaction terms.

For the nested F-test to determine if the slopes are unequal (i.e., if the interaction terms are significant), the null and alternative hypotheses are:

Null Hypothesis ($H0$): The interaction terms (slopes) do not significantly improve the model. In other words, the reduced model (Model 2) is sufficient.

$\beta Interaction1 = \beta Interaction2 = 0$

Alternative Hypothesis (Ha): At least one of the interaction terms is significant, suggesting unequal slopes and that the full model (Model 3) provides a better fit.

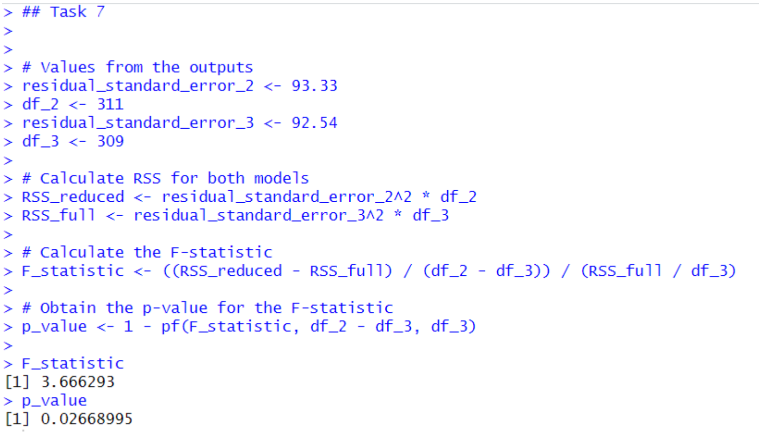
$\beta Interaction1 \neq 0 \text{ } or \text{ } \beta Interaction2 \neq 0$

Where:

$\beta Interaction1$ is the coefficient for the Interaction1 term.

$\beta Interaction2$ is the coefficient for the Interaction2 term.

If the F-test rejects the null hypothesis, it indicates that there are unequal slopes, and at least one of the interaction terms significantly improves the model fit.



```
> ## Task 7
>
>
> # Values from the outputs
> residual_standard_error_2 <- 93.33
> df_2 <- 311
> residual_standard_error_3 <- 92.54
> df_3 <- 309
>
> # Calculate RSS for both models
> RSS_reduced <- residual_standard_error_2^2 * df_2
> RSS_full <- residual_standard_error_3^2 * df_3
>
> # Calculate the F-statistic
> F_statistic <- ((RSS_reduced - RSS_full) / (df_2 - df_3)) / (RSS_full / df_3)
>
> # Obtain the p-value for the F-statistic
> p_value <- 1 - pf(F_statistic, df_2 - df_3, df_3)
>
> F_statistic
[1] 3.666293
> p_value
[1] 0.02668995
```

Interpretation:

- F-statistic Value: The computed F-statistic is 3.666293. This value represents the ratio of the improvement in fit by the full model (Model 3) over the reduced model (Model 2), relative to the variability not explained by the full model.
- p-value: The p-value is 0.02668995. p-value is less than 0.05, the null hypothesis is rejected in favor of the alternative hypothesis. This means there is evidence to suggest that at least one of the interaction terms in the full model (Model 3) is statistically significant and improves the fit compared to the reduced model (Model 2).

Conclusion:

The test results indicate that the full model with the interaction terms provides a significantly better fit to the data than the reduced model. This provides evidence of unequal slopes in the relationship between fat and cholesterol when accounting for prior smoking habits. In other words, the effect of fat on cholesterol seems to differ based on prior smoking habits, suggesting that the interaction between these factors is significant.

The evidence suggests that there are unequal slopes, meaning that the relationship between fat intake and cholesterol levels is not consistent across different levels of prior smoking habits. This underscores the importance of considering interactions in regression models, as they can reveal more nuanced relationships between predictor variables and the outcome of interest.

Task 8

Now that you've been exposed to these modelling techniques, it is time for you to use them in practice. Let's examine more of the NutritionStudy data. Use the above modelling approach to determine if the categorical variables SMOKE, ALCOHOL CONSUMPTION or GENDER, along with the continuous variables FAT variable are predictive of CHOLESTEROL. Formulate hypotheses, construct essential variables (as necessary), conduct the analysis and report on the results. Which categorical variables are most predictive of CHOLESTEROL?

The goal is to investigate the relationships between cholesterol levels and several independent variables: `Smoke`, `Alcohol_d2`, `Alcohol_d3`, `Gender_d`, `PriorSmoke_d2`, and `PriorSmoke_d3`, including their interactions with `Fat`. To help formulate hypotheses for this mixed model, let's break down the potential relationships:

1. Main Effects:

   - Smoke_d: The effect of smoking on cholesterol levels.
   - Alcohol_d2: The effect of moderate alcohol consumption on cholesterol levels.
   - Alcohol_d3: The effect of high alcohol consumption on cholesterol levels.
   - Gender_d: The effect of gender (potentially male if coded as 1) on cholesterol levels.
   - PriorSmoke_d2: The effect of having smoked in the past (but no longer currently smoking) on cholesterol levels.
   - PriorSmoke_d3: The effect of having smoked for a longer duration in the past on cholesterol levels.
   - Fat: The amount of Fat in the diet.

2. Interactions:
   - Alcohol_d2Alcohol_d3: This checks if the combined effects of moderate and high alcohol consumption on cholesterol levels are different than what we would expect based on their separate effects.
   - Gender_dPriorSmoke_d2: The interaction effect of gender and having smoked in the past on cholesterol levels.
   - PriorSmoke_d2PriorSmoke_d3: The interaction effect of having smoked in the past and the duration of past smoking on cholesterol levels.
   - Main effects & Interactions with Fat: This investigates if the effect of each of the main predictors on cholesterol levels changes depending on the amount of fat in the individual's diet.

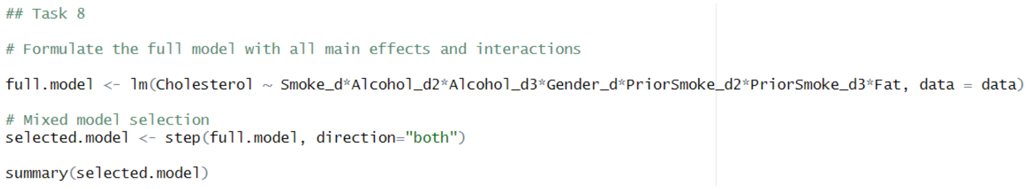
Hypotheses: Potential null and alternative hypotheses for each main effect and interaction:

1. Smoke_d:
   - $H0$: Smoking does not have an effect on cholesterol levels.
   - $H1$: Smoking has an effect on cholesterol levels.
2. Alcohol_d2:
   - $H0$: Moderate alcohol consumption does not affect cholesterol levels.
   - $H1$: Moderate alcohol consumption affects cholesterol levels.
3. Alcohol_d3:
   - $H0$: High alcohol consumption does not affect cholesterol levels.
   - $H1$: High alcohol consumption affects cholesterol levels.

4. Gender_d:
   - $H0$: Gender does not have an effect on cholesterol levels.
   - $H1$: Gender has an effect on cholesterol levels.
5. PriorSmoke_d2:
   - $H0$: Having smoked in the past does not affect cholesterol levels.
   - $H1$: Having smoked in the past affects cholesterol levels.
6. PriorSmoke_d3:
   - $H0$: The duration of past smoking does not affect cholesterol levels.
   - $H1$: The duration of past smoking affects cholesterol levels.
7. Fat:

- $H0$: The amount of fat in the diet does not affect cholesterol levels.
- $H1$: The amount of fat in the diet affects cholesterol levels.

8. Interactions: For each interaction effect, the null hypothesis (H0) would state that there is no interaction effect (i.e., the combined effect of the two variables is simply the sum of their individual effects), while the alternative hypothesis (H1) would suggest there is an interaction effect (i.e., the combined effect of the two variables is different than the sum of their individual effects).



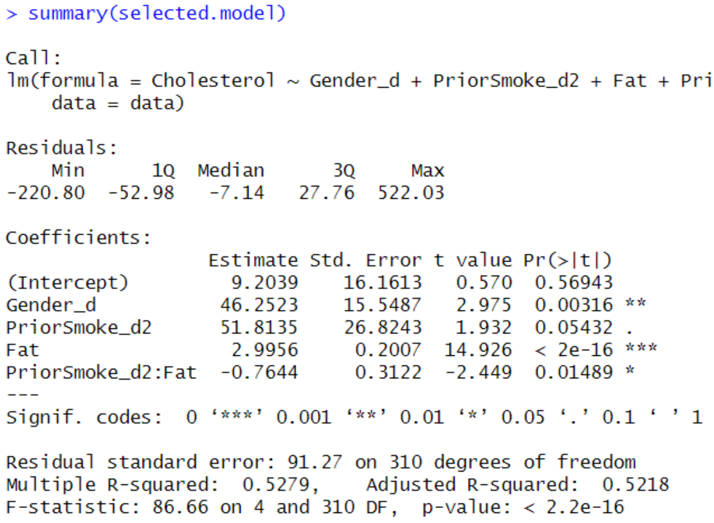
```
## Task 8

# Formulate the full model with all main effects and interactions

full.model <- lm(Cholesterol ~ Smoke_d*Alcohol_d2*Alcohol_d3*Gender_d*PriorSmoke_d2*PriorSmoke_d3*Fat, data = data)

# Mixed model selection
selected.model <- step(full.model, direction="both")

summary(selected.model)
```

Formulation of the Full Model:

- The code defines a full linear regression model named full.model that aims to predict Cholesterol based on several predictors.
- Main effects: Smoke_d, Alcohol_d2, Alcohol_d3, Gender_d, PriorSmoke_d2, PriorSmoke_d3, and Fat.
- Interaction terms: The model also considers interactions between all the variables.

Model Selection:

- The code then performs a model selection process using the step function. This function uses a stepwise algorithm to select a more parsimonious model that retains significant predictors while removing non-significant ones.
- The direction="both" argument indicates that the stepwise procedure can add or remove terms in the model (i.e., it's bidirectional - both forward and backward selection).



```
> summary(selected.model)

Call:
lm(formula = Cholesterol ~ Gender_d + PriorSmoke_d2 + Fat + PriorSmoke_d2:Fat,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-220.80  -52.98   -7.14   27.76  522.03

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.2039    16.1613   0.570  0.56943
Gender_d           46.2523    15.5487   2.975  0.00316 **
PriorSmoke_d2      51.8135    26.8243   1.932  0.05432 .
Fat                 2.9956     0.2007  14.926  < 2e-16 ***
PriorSmoke_d2:Fat  -0.7644     0.3122  -2.449  0.01489 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91.27 on 310 degrees of freedom
Multiple R-squared:  0.5279,	Adjusted R-squared:  0.5218
F-statistic: 86.66 on 4 and 310 DF,  p-value: < 2.2e-16
```

1. Model:
   - Dependent variable: `Cholesterol`
   - Predictors: `Gender_d`, `PriorSmoke_d2`, `Fat`, and the interaction term `PriorSmoke_d2:Fat`.

2. Equation:

$$Cholesterol = 9.2039 + 46.2523 \times Gender\_d + 51.8135 \times PriorSmoke\_d2 + 2.9956 \times Fat - 0.7644 \times PriorSmoke\_d2{:}\,Fat$$

3. Residuals:

- These provide insight into the model's fit by showing the difference between observed and predicted values. The median of the residuals is -7.14 (close to 0), which indicates that on average, the model's predictions are fairly accurate.

4. Coefficients:
   - Intercept: The average value of cholesterol when all predictors are 0. It's estimated at 9.2039, but it's not statistically significant (p=0.56943).
   - Gender_d: On average, the `Gender_d` variable contributes an additional 46.2523 units to the cholesterol level, and it's statistically significant (p=0.00316). Since Gender_d = 1 was set for Male, according to the model Males have, on average, 46.2523 units higher cholesterol levels than Females when all other variables are held constant.
   - PriorSmoke_d2: On average, `PriorSmoke_d2` contributes an additional 51.8135 units to the cholesterol level. This predictor could be marginally significant (p=0.05432).
   - Fat: Each unit increase in `Fat` results in a 2.9956 unit increase in cholesterol. This predictor is highly significant (p < 2e-16).
   - PriorSmoke_d2:Fat Interaction: The interaction term indicates how the effect of `Fat` on cholesterol varies depending on the value of `PriorSmoke_d2`. The negative coefficient of -0.7644 suggests that the effect of fat on cholesterol is reduced by this amount when `PriorSmoke_d2` is present. This interaction is statistically significant (p=0.01489).

4. Goodness of fit:

- Residual Standard Error: The model's predictions are, on average, about 91.27 units away from the actual observed values.
- Multiple R-squared: Approximately 52.79% of the variability in cholesterol can be explained by the predictors in the model.
- Adjusted R-squared: After accounting for the number of predictors in the model, about 52.18% of the variability in cholesterol is explained.
- F-statistic: The overall model is highly significant (p-value < 2.2e-16), meaning that it's very likely that at least one of the predictors has a real effect on the dependent variable.

Overall: The model suggests that `Gender_d`, `Fat`, and the interaction between `PriorSmoke_d2` and `Fat` have significant effects on cholesterol. However, while `PriorSmoke_d2` shows an effect, it's marginally significant and should be interpreted with caution.

Task 9
Please write a conclusion / reflection on your experiences in this assignment.

The analytical journey undertaken to understand the predictors of cholesterol levels underscores the intricacies inherent in data-driven insights. Starting from a simplistic model focusing solely on prior smoking status, we progressively introduced more variables and interactions, shedding light on the multifaceted nature of cholesterol determinants.

The inclusion of the `Fat` variable as a covariate in the ANCOVA model highlighted the intertwined relationship between diet and cholesterol, and its significance underscored the need to factor in dietary habits when examining health outcomes. Furthermore, the interaction terms introduced in the Unequal Slopes Model hinted at the nuanced effects that lifestyle choices, such as smoking and fat intake, might have on cholesterol when combined.

However, the Full Model served as a reminder that a more comprehensive model isn't necessarily a better one. Complexity can lead to overfitting and reduced interpretability. This necessitated the stepwise regression process, ultimately yielding the Selected Model. This refined model not only ensured statistical robustness but also enhanced the clarity of insights by focusing on the most influential predictors.

This exercise emphasizes that statistical modeling is as much an art as it is a science. Balancing comprehensiveness with simplicity, ensuring robustness while avoiding overfitting, and constantly validating model assumptions are critical steps in the modeling process. Above all, the journey reiterates the importance of iterative refinement in data analytics – beginning with a foundational understanding and progressively building upon it to achieve nuanced and actionable insights.