Nutrition Study Data Report

1. One-way ANOVA

1.1 Descriptive Statistics

The *NutritionStudy* data is a 16-variable dataset with 315 records containing information obtained from medical records and observational self-report surveys of adults. One measure, *Cholesterol*, is a common response variable for predicting health outcomes. Descriptive statistics such as measures of Central Tendency, as well as the spread of data points were calculated for *Cholesterol* with across groups of *PriorSmoke* and reported in Table 1. *Cholesterol* steadily increases across the *PriorSmoke* categories, as the *PriorSmoke* category increases, there is a higher average of *Cholesterol*, suggesting that people in higher *PriorSmoke* category have higher *Cholesterol* levels. Additionally, the medians are lower than the means in all groups, while the max values are extremely high, indicating some right-skew distributions with potential outliers.

Table 1. Summary Statistics of Cholesterol by Prior Smoking Status

Summary Statistics by Prior Smoking Status

Cholesterol Levels

Prior Smoke	N	Prop (%)	Median	Mean	Min	Max	Variance	SD
1	157	49.8	195.8	228.4	37.7	900.7	18,017	134.2
2	115	36.5	216.7	250.4	46.3	747.5	14,809	121.7
3	43	13.7	239.2	272.5	78.3	718.8	21,292	145.9

N = sample size; Prop = proportion; SD = standard deviation.

1.2 One-way ANOVA Model

A one-way ANOVA model was fit to compare mean *Cholesterol* levels (Y) across the different *PriorSmoke* groups (X). The effect of prior smoking on cholesterol levels was found not to be statistically significant ($F_{(2, 312)} = 2.235$, $p = 0.109$). Although the mean *Cholesterol* increased across *PriorSmoke* groups, these differences were not significant, suggesting that cholesterol levels did not differ by prior smoking status. Figure A1 visualizes the difference between group means in *Cholesterol*.

2. MLR Regression (Model 1)

2.1 Model 1 Interpretation

A simple linear regression was fit to compare *Cholesterol* levels (Y) across *PriorSmoke* groups (X). Prior to the model fitting, dummy variables were created for the different *PriorSmoke* groups to compare group means of *Cholesterol* levels. The prediction equation for Model 1 is:

$$\widehat{Y} = 228.391 + 22.033(PriorSmoke\_2) + 44.141(PriorSmoke\_3)$$

The intercept coefficient ($\beta_0 = 228.391$) is the predicted cholesterol level for the reference group, specifically when *PriorSmoke* is 1 and was statistically significant ($t = 21.766$, $p < 0.0001$). This represents the baseline cholesterol level for individuals in the first *PriorSmoke* category. The coefficient for PriorSmoke group 2 ($\beta_1 = 22.033$) suggests that former smokers have cholesterol levels that are, on average, 22.033 higher than *PriorSmoke* group 1, though this difference is not statistically significant ($t = 1.365$, $p = 0.173$). The coefficient for PriorSmoke group 3 ($\beta_2 = 44.141$) indicates that current smokers have cholesterol levels that are, on average, 44.141 higher than PriorSmoke group 1, and this difference approaches statistical significance ($t = 1.951$, $p = 0.052$). Table 2 below reports on the coefficient table for Model 1.

Table 2. Coefficient table for Model 1

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	228.39	10.49	21.766	<2e-16
PriorSmoke_2	22.03	16.14	1.365	0.173
PriorSmoke_3	44.14	22.63	1.951	0.052

An ANOVA was performed to determine goodness-of-fit for Model 1, highlighted in Table 3. The model's overall fit is weak, with a coefficient of determination was 0.0141, indicating that *PriorSmoke* status only explains 1.41% of the variability in *Cholesterol*. Additionally, the overall model was not statistically significant ($F_{(2, 312)} = 2.235$, $p = 0.109$) as seen in Table 3.

Table 3. ANOVA table for Model 1

	DF	Sum Squares	Mean Square Err	F value	Pr (>F)
PriorSmoke_2	1	11487	11487	0.6645	0.4165
PriorSmoke_3	1	65771	65771	3.8049	0.052
Residuals	312	5393183	17286		

2.2 ANOVA and Model 1 Comparison

The ANOVA model from section 1.2 and the regression model (Model 1) produce equivalent overall results, which is expected since an ANOVA with categorical variables is mathematically identical to multiple regression with dummy variables. Both models yield the same overall F-statistic ($F = 2.235$), p-value ($p = 0.1087$), and degrees of freedom ($2, 312$). Additionally, both models' group means are identical: *PriorSmoke* group 1 (228.39), *PriorSmoke* group 2 (250.42), and *PriorSmoke* group 3 (272.53).

The difference lies in how the models present and test the results. The ANOVA model tests the omnibus null hypothesis that all group means are equal, providing a single test with a p-value of 0.109. This answers the broad question of whether smoking status has any effect on cholesterol levels. In contrast, the regression model breaks down the comparison into specific contrasts, testing whether *PriorSmoke* group 2 differs from *PriorSmoke* group 1 ($p = 0.173$), and *PriorSmoke* group 3 differs from *PriorSmoke* group 1 ($p = 0.052$). Additionally, the ANOVA table for Model 1 reveals an important nuance. When predictors are entered sequentially,

PriorSmoke_2 accounts for 11,487 in sum of squares (*p = 0.416*), while *PriorSmoke_3* accounts for an additional 65,771 in sum of squares (*p = 0.052*). The sequential decomposition indicates that most of the variation explained by smoking status comes from *PriorSmoke_3*, while *PriorSmoke_2* contributes relatively little to the *Cholesterol* variation.

In the end, both approaches lead to the same conclusion: smoking status does not significantly predict cholesterol levels at alpha of 0.05, though *PriorSmoke_3* warrants further investigation from its marginal significance.

3. ANCOVA (Model 2)

3.1 Model 2 Interpretation

An ANOVA model, Model 2, was fitted that incorporates both categorical predictors from dummy coded *PriorSmoke* and the continuous predictor, *Fat*, to predict *Cholesterol* levels. This ANCOVA model allows us to examine the effects of smoking status on cholesterol levels while controlling for amount of fat. The prediction equation can be written as:

$$\widehat{Y} = 28.940 - 2.114(PriorSmoke\_2) + 10.636(PriorSmoke\_3) + 2.763(Fat)$$

The intercept ($\beta_0 = 28.94$) represents the predicted *Cholesterol* level for *PriorSmoke_1* (reference group) with zero *Fat*, though this is not meaningful in practice since it is unrealistic for a person to have zero fat. The coefficient for PriorSmoke_2 ($\beta_1 = -2.11$, $p = 0.855$) suggests that individuals in *PriorSmoke_2* have *Cholesterol* levels 2.114 lower than individuals in *PriorSmoke_1* when controlling for *Fat*, though this difference is not statistically significant. The coefficient for PriorSmoke_3 ($\beta_2 = 10.636$, p = *0.511*) indicates that individuals in *PriorSmoke_2* have *Cholesterol* levels 10.636 higher than individuals in *PriorSmoke_1*, though not statistically significant. Most importantly, the Fat coefficient ($\beta_3 = 2.763$, $p < 0.0001$) shows that for every 1-unit increase in *Fat*, *Cholesterol* increases by 2.763, holding smoking status constant. This is highly significant and represents the strongest predictor in the model.

An omnibus F-test was performed to determine if the overall model was statistically significant. The F-test revealed that the model was highly significant ($F_{(3, 311)} = 105.7$, $p < 0.0001$), indicating that the predictors collectively explain a substantial amount of variance in *Cholesterol*. An ANOVA was performed to determine goodness of fit for Model 2. The sequential ANOVA table in Table 4 uncovers Fat contributes the vast majority of explained variance ($F_{(1, 311)} = 308.21$, $p < 0.0001$), while smoking status variables contribute much less. Notably, when Fat is added to the model, the smoking status variables lose their already marginal significance, indicating that much of the apparent relationship between *Cholesterol* and *PriorSmoke* in Model 1 may have been confounded by differences in *Fat* across smoking groups.

Table 4. ANOVA table for Model 2

	DF	Sum Squares	Mean Square Err	F value	Pr (>F)
PriorSmoke_2	1	11487	11487	1.3189	0.252
PriorSmoke_3	1	65771	65771	7.5513	0.006
Fat	1	2684427	2684427	308.21	< 2.2e-16
Residuals	311	2708756	8710		

The R^2 value showed strong predictive performance with a value of 0.5048, meaning that 50.48% of the variance in *Cholesterol* was explained by *PriorSmoke* and *Fat*. The adjusted R^2 value of 0.5001 confirms that this is not due to overfitting. This represents a dramatic improvement over Model 1, which only explained 1.41% of the variance. Additionally, the residual standard error of 93.33 indicates typical prediction error, which is substantially lower than Model 1's residual standard error of 131.5, meaning that Model 2's predictions are closer to the true value of Cholesterol.

3.2 Diagnostic Assessment

A comprehensive diagnostic assessment was conducted to determine if Model 2 upholds underlying assumptions of an ANCOVA model, which include, normality, homoscedasticity, homogeneity of regression slopes, and outlier tests.

To assess normality, a Shapiro-Wilk Test was performed for a statistical assessment, and a Q-Q plot was used for visual assessment. The Shapiro-Wilk Test indicates a significant violation of the normality assumption ($W = 0.884$, $p < 0.0001$). The Q-Q plot confirms this, showing deviations from the normal line, particularly at the upper tail of the distribution. Observations 94 and 112 are identified as potential severe outliers. While this violation is concerning, with a sample size of 315, the Central Limit Theorem provides some robustness to moderate departures normal.

To assess homoscedasticity, a non-constant variance test was conducted for statistical analysis and scatterplot of the residuals and predicted values was produced for visual analysis. The non-constant variance test reveals a significant violation of the homoscedasticity assumption ($\chi^2 = 36.20$, $p < 0.0001$). The residuals vs. fitted values plot in Figure A3 shows that residual spread increases substantially as predicted cholesterol values exceed 300, creating a funnel-shaped pattern. This heteroscedasticity suggests that prediction precision decreases for individuals with higher cholesterol levels and may require transformation or a weighted analysis.

To assess homogeneity of regression slopes, or parallel slopes, an interaction test was conducted. The interaction test reveals the assumption of parallel slopes is violated ($F_{(2, 309)} = 3.668$, $p < 0.0001$). Specifically, the interaction between *PriorSmoke_2* and *Fat* is significant ($\beta = -0.684$, $p = 0.0431$), indicating that the relationship between Fat and Cholesterol differs for *PriorSmoke_2* compared to *PriorSmoke_1*. The slope for *PriorSmoke_2* (2.974 - 0.684 = 2.290) is less steep than for *PriorSmoke_1* (2.974), suggesting that *PriorSmoke_2* shows a weaker cholesterol response to *Fat*.

To assess potential outliers in Model 2, an outlier test was performed. The outlier test identified three significant outliers: observation 112 (studentized residual = 5.819, Bonferroni p < 0.0001), observation 94 (studentized residual = 5.758, Bonferroni p < 0.0001), and observation 257 (studentized residual = 4.654, Bonferroni p = 0.002). These individuals have cholesterol levels much higher than predicted by the model, even accounting for their smoking status and

fat. Further investigation of these cases is warranted to determine if they represent data entry errors, unique biological responses, or other unmeasured factors affecting cholesterol levels.

3.3 Summary

Model 2 demonstrates that *Fat* is by far the strongest predictor of cholesterol levels, rendering smoking status non-significant when fat is controlled. However, significant violations of the normality, homoscedasticity, and homogeneity of slopes assumptions raise concerns about the validity of the standard ANCOVA model. The significant interaction between *PriorSmoke_2* status and *Fat* suggests that a model including interaction terms may be more appropriate.

4. Scatterplot Comparison

4.1 Predicted Cholesterol Values vs. Fat

A scatterplot of the predicted *Cholesterol* values in Model 2 was plotted against *Fat* illustrated in Figure A3. A striking pattern emerges: the ANCOVA produces three parallel lines, one for each *PriorSmoke* group. Group 1 (shown in green) has the lowest predicted values of *Cholesterol* at any given *Fat* level, Group 2 (shown in orange) falls in the middle, and Group 3 (shown in blue) has the highest *Cholesterol* levels. Importantly, all three lines have an identical slope of 2.763 meaning the model assumes that a 1-unit increase in *Fat* raises *Cholesterol* by exactly 2.763, regardless of smoking status. The lines are simply vertically shifted versions of each other, with Group 2 lowered by -2.114 and Group 3 elevated by 10.636 relative to Group 1. This parallel pattern is a fundamental characteristic of the standard ANCOVA model, which assumes homogeneity of regression slopes.

4.2 Observed Cholesterol Values vs. Fat

A scatterplot of the observed *Cholesterol* values in was plotted against *Fat* seen in Figure A4. When comparing the actual *Cholesterol* values versus *Fat*, a notable discrepancy emerges. The observed data does not clearly exhibit parallel patterns across the three smoking groups. In the actual data, the relationship between *Fat* and *Cholesterol* appears to vary by smoking status. Group 2 shows what appears to be a less steep relationship between fat and cholesterol compared to Groups 1 and 3. Additionally, there is substantial scatter around any potential trend lines, with considerable overlap among the three groups, particularly at lower fat intake levels. The observed data also contains several extreme outliers with very high cholesterol values (above 700) that the parallel-lines model struggles to accommodate.

4.3 Model Adequacy Assessment

The ANCOVA model does not appear to fit the observed data very well, and a more complex model is needed. This conclusion is supported by three key observations. First, the assumption of parallel slopes is clearly violated, as confirmed by the significant interaction test ($p = 0.027$) from the diagnostic analysis. The PriorSmoke_2:Fat interaction ($\beta = -0.684$, $p = 0.043$) indicates that *PriorSmoke_2* have a different fat-cholesterol slope than *PriorSmoke_1*. Second, the simple

parallel-lines structure cannot capture the complexity visible in the actual data, where the groups appear to have genuinely different responses to *Fat*. Third, the heteroscedasticity detected in the residual diagnostics is visible in the raw data plot as well, variability increases dramatically at higher *Fat* levels, particularly for individuals with *Cholesterol* above 500.

5. Unequal Slopes Model (Model 3)

5.1 Model 3 Interpretation

Interaction terms *PriorSmoke_2:Fat* and *PriorSmoke_3:Fat* were manually calculated by taking the product of *PriorSmoke* dummy variables with *Fat* and added to the dataset to develop an unequal slopes model, deemed Model 3. Model 3 was fitted and the prediction equation written as:

$$\widehat{\boldsymbol{Y}} = \mathbf{13.703} + \mathbf{51.389}(\boldsymbol{PriorSmoke\_2}) - \mathbf{32.882}(\boldsymbol{PriorSmoke\_3}) + \mathbf{2.974}(\boldsymbol{Fat}) - \mathbf{0.684}(\boldsymbol{PriorSmoke\_2}:\boldsymbol{Fat}) + \mathbf{0.486}(\boldsymbol{PriorSmoke\_3}:\boldsymbol{Fat})$$

The intercept ($\beta_0 = 13.703$, $p = 0.4539$) represents the *Cholesterol* levels for *PriorSmoke_1* with zero *Fat*, which is not practical. The main effect of *Fat* ($\beta_3 = 2.974$, $p < 0.0001$) indicates that for the reference group (*PriorSmoke_1*), each additional increase in *Fat* increases *Cholesterol* by 2.974. The *PriorSmoke_2* coefficient *($\beta_2 = 51.389$, $p = 0.070$)* represents the difference in intercepts between *PriorSmoke_2* and *PriorSmoke_1* at zero *Fat*, which approaches but does not reach conventional significance. More importantly, the *PriorSmoke_2:Fat* interaction *($\beta_4 = -0.684$, $p = 0.043$)* is statistically significant, indicating that the slope of the fat-cholesterol relationship for *PriorSmoke_2* is 0.684 lower than for *PriorSmoke_1*. This means *PriorSmoke_2* shows a weaker *Cholesterol* response to *Fat* (slope = 2.290) compared to *PriorSmoke_1* (slope = 2.974). The *PriorSmoke_3* coefficient *($\beta_2 = -32.882$, $p = 0.437$)* and its interaction term *($\beta_5 = 0.486$, $p = 0.311$)* are both non-significant. While the interaction suggests PriorSmoke_3 might have a steeper slope (3.460) than PriorSmoke_1, this difference is not statistically reliable. The negative main effect coupled with the positive interaction creates a crossover pattern, where *PriorSmoke_3* have lower predicted cholesterol at low fat levels but potentially higher cholesterol at very high fat level.

An omnibus F-test was performed to determine if the overall model was statistically significant. The F-test determined that Model 3 was highly significant *($F_{(5, 309)} = 65.97$, $p < 0.0001$)*, indicating that the predictors collectively explain substantial variance in *Cholesterol*. An ANOVA was performed to determine goodness of fit for Model 3. The sequential ANOVA table in Table 5 shows that the interaction terms collectively add 62,817 in sum of squares beyond the main effects model. When compared to Model 2, the addition of the two interaction terms significantly improves model fit *($F_{(2, 309)} = 3.668$, $p = 0.027$)*, justifying the increased complexity of the unequal slopes model.

Table 5. ANOVA Table for Model 3

	DF	Sum Squares	Mean Square Err	F value	Pr (>F)
PriorSmoke_2	1	11487	11487	1.3189	0.252
PriorSmoke_3	1	65771	65771	7.5513	0.006

Fat	1	2684427	2684427	313.4948	< 2.2e-16
PriorSmoke_2:Fat	1	53998	53998	6.3061	0.012543
PriorSmoke_3:Fat	1	8819	8819	1.0298	0.310988
Residuals	309	2645939	8563		

The R^2 value shows modest improvement over Model 2. The R^2 increased from 0.5048 to 0.5163, meaning the model now explains 51.63% of variance in *Cholesterol* levels. The adjusted R^2 of 0.5085 confirms this improvement is not merely due to adding parameters. The residual standard error decreased slightly from 93.33 to 92.54, indicating marginally better prediction precision. While these improvements are statistically significant, they are relatively modest in practical terms, suggesting that the interaction effects, while real, explain only a small additional portion of cholesterol variability.

5.2 Diagnostic Assessment

A comprehensive diagnostic assessment was conducted to determine if Model 3 upholds underlying assumptions of an Unequal Slopes model, which include, normality, homoscedasticity, and outlier tests.

A Q-Q plot was developed to assess normality assumption. The plot reveals that the normality assumption is violated. The plot shows clear deviations from the theoretical normal line, particularly at both tails. At the lower tail (left side), there's a slight departure suggesting some left skewness. More problematically, at the upper tail (right side), there are substantial deviations with observations pulling away from the reference line, indicating heavy right-tail behavior. Observations 94 and 112 are flagged as extreme outliers in the outlier test with studentized residuals around 6, far exceeding conventional cutoffs. The overall pattern suggests the residuals follow a right-skewed distribution rather than a normal distribution.

Homoscedasticity was tested by producing a scatterplot of the residuals versus predicted values seen in Figure A5. The plot reveals that heteroscedasticity persists in Model 3, though the pattern differs slightly from Model 2. For predicted values between approximately 100 and 400, the residuals show relatively constant spread around zero, which is encouraging. However, at the extremes of the prediction range, particularly for predicted values above 500, there are observations with very large positive residuals (exceeding 400), indicating the model severely underpredicts for some individuals. The funnel-shaped pattern is less pronounced than in Model 2, but the variance is still not constant across the range of fitted values.

5.3 Predicted Cholesterol Values vs. Fat Scatterplot Analysis

Figure A6 represents a scatterplot of predicted *Cholesterol* values versus *Fat* values in Model 3 to understand the differences between Model 2 and Model 3. The plot reveals the fundamental differences between the parallel slope model (Model 2) and the unequal slope model (Model 3). Unlike Model 2, which produced three parallel lines, Model 3 generates three non-parallel

regression lines, one for each smoking group, demonstrating that the relationship between *Fat* and *Cholesterol* varies by *PriorSmoke* status.

Group 1 (shown in blue) represents the baseline or reference relationship against which the other groups are compared. This group shows a moderately steep positive relationship, with the line starting at approximately 14 at zero Fat and rising with a slope of 2.974 per 1-unit increase in Fat. Group 2 (shown in red) has the flattest slope of the three groups (2.290). Notably, this line starts much higher than the Group 1's line (intercept around 65) but rises more gradually. This creates a convergence pattern where Group 2 have substantially higher predicted *Cholesterol* at low *Fat* compared to Group 1, but this difference diminishes as *Fat* increases. Group 3 (shown in yellow) shows the steepest slope (3.460) but starts with the lowest intercept (around -19). This creates a crossover pattern. At very low *Fat* (below approximately 40), Group 3 are predicted to have the lowest cholesterol of all three groups. However, because their slope is steepest, their line rapidly crosses both other lines. By moderate *Fat* (around 100), Group 3 have caught up to Group 1, and at high *Fat* (above 150), Group 3 are predicted to have the highest cholesterol levels of all three groups.

The diverging pattern is the most striking feature of this plot. At low *Fat* levels (20-60), the three lines are relatively close together and even cross over. As *Fat* increases beyond 100, the lines spread apart considerably, with Group 3 pulling away at the top and Group 2 maintaining intermediate predictions. The crossover point between Group 3 and Group 1 occurs at *Fat* levels around 40-50. Below this point, the negative intercept for Group 3 dominates, giving them lower predicted *Cholestero*l. Above this point, their steeper slope takes over, and their *Cholesterol* levels increase more rapidly with each 1-unit increase of *Fat*.

6. Nested Model

6.1 Hypothesis Testing

Model 2 and Model 3 are nested models, meaning that one model is a special case of the other. Model 3 is the full model because it contains all the predictors for Model 2 plus the additional interaction terms. Model 2 is the reduced model because it can be obtained from Model 3 by setting the interaction coefficients ($\beta_4 = \beta_5 = 0$) to zero. The null hypothesis for a nested F-test states that the regression slopes are homogeneous across smoking groups, indicating that the interaction terms do not significantly improve the model. This implies that the relationship between *Fat* and *Cholesterol* is the same for all *PriorSmoke* groups. The alternative hypothesis for a nested F-test states that the regression slopes are not homogeneous across smoking groups, meaning that at least one interaction term is significantly different from zero. This implies that the relationship between *Fat* and *Cholesterol* differs by *PriorSmoke* group. The hypothesis can be written as:

$$H_0: \beta_4 = \beta_5 = 0$$

$$H_A: \text{At least of } \beta_4 \text{ or } \beta_5 \neq 0$$

The nested F-test was statistically significant ($F_{(2, 309)} = 3.668$, $p = 0.0265$) at $\alpha = 0.05$, indicating that we can reject the null hypothesis. There is statistically significant evidence that the interaction terms improve the model fit, indicating that regression slopes are not homogeneous across smoking groups. Adding the two interaction terms to the model significantly reduces the sum of squares by 62,817, which is substantial enough to justify the increased model complexity despite adding the extra parameters.

6.2 Summary

The nested F-test provides strong evidence that the relationship between *Fat* and *Cholesterol* levels varies significantly by smoking status. The significant nested F-test confirms what we observed in the predicted values plot: the three PriorSmoke groups exhibit different fat-cholesterol relationships. Specifically, the significant *PriorSmoke_2:Fat* interaction *($\beta_4 = -0.684$, $p = 0.043$)* indicates that *PriorSmoke_2* show a significantly weaker *Cholesterol* response to *Fat* levels compared to *PriorSmoke_1*. For every 1-unit increase in *Fat*, individuals in *PriorSmoke_2* group's cholesterol increases by only 2.290 compared to 2.974 for *PriorSmoke_1*.

From a practical standpoint, these findings have important implications for cholesterol management. The parallel slopes assumption of Model 2 would lead to identical fat recommendations for all individuals regardless of smoking status. However, Model 3 reveals that a one-size-fits-all approach may be inappropriate. The statistical significance of the nested F-test, combined with the practical differences in predicted slopes (ranging from 2.290 to 3.460), strongly supports the use of Model 3 over Model 2. While Model 2 is more parsimonious with fewer parameters, it fails to capture important heterogeneity in how different smoking groups metabolize fat.

7. Best Predictor of Cholesterol

A comprehensive analysis of predicting *Cholesterol* levels extending the methodology previously reported on with *Fat* and other categorical variables such as *Smoke*, *Gender*, and *Alcohol*. Alcohol is a continuous variable but was discretized into meaningful categories: "None" (Alcohol = 0), "Some" (0 < Alcohol < 10), and "A lot" (Alcohol ≥ 10). This categorization facilitates interpretation and allows for testing of non-linear relationships. Dummy variables were created for current smoking status (Smoke_Yes: 0 = non-smoker, 1 = current smoker), gender (Gender_Male: 0 = female, 1 = male), and alcohol consumption levels, with "None" serving as the reference category.

A hierarchical modeling approach was employed, building progressively more complex models to isolate the effects of each categorical variable, starting with *Smoke* (Model 4), *Gender* (Model 5), and *Alcohol* (Model 6). For each categorical predictor, a fitted parallel slopes model (ANCOVA assuming homogeneous slopes) and an unequal slopes model (allowing interactions between the categorical variable and fat) was performed. Nested F-tests were used to determine whether interaction terms significantly improved model fit. Models were compared using R^2

values to determine goodness-of-fit to the data. Finally, a full comprehensive model with all categories was fitted (Model 7).

The parallel slopes model (Model 4A) revealed that current smoking status did not significantly predict *Cholesterol* levels when controlling for *Fat* levels ($\beta = 11.56$, $p = 0.452$). The model explained 50.48% of variance in cholesterol ($R^2 = 0.5048$). *Fat* remained a highly significant predictor ($\beta = 2.760$, $p < 0.0001$), indicating that for every additional unit of *Fat*, *Cholesterol* increases by 2.76 regardless of current smoking status. The unequal slopes model (Model 4B) suggested a marginally significant interaction between smoking and fat ($\beta = 0.811$, $p = 0.074$), with current smokers showing a steeper fat-cholesterol slope (3.46) compared to non-smokers (2.65). However, the nested F-test did not reach conventional significance ($F_{(1, 311)} = 3.21$, $p = 0.074$), indicating that the parallel slopes assumption is appropriate.

Gender emerged as a highly significant predictor of cholesterol levels. The parallel slopes model (Model 5A) showed that males have cholesterol levels 46.76 higher than females after controlling for *Fat* ($\beta = 46.76$, $p = 0.003$). This model explained 51.79% of variance ($R^2 = 0.5179$), indicating superior performance. The unequal slopes model (Model 5B) tested whether the fat-cholesterol relationship differed by gender. While the Gender × Fat interaction coefficient was negative ($\beta = -0.482$), suggesting males might have a slightly flatter slope, this interaction was not statistically significant ($p = 0.294$). The nested F-test confirmed that slopes are homogeneous across genders ($F_{(1, 311)} = 1.11$, $p = 0.294$), supporting the parallel slopes model. Both males and females show similar cholesterol responses to fat levels (approximately 2.7), but males consistently maintain higher baseline cholesterol levels across all fat levels.

Alcohol consumption categories showed no significant relationship with cholesterol levels in the parallel slopes model (Model 6A). Neither moderate drinkers ($\beta = -8.22$, $p = 0.468$) nor heavy drinkers ($\beta = -9.83$, $p = 0.633$) differed significantly from non-drinkers in cholesterol levels when controlling *Fat* levels. The model explained 50.48% of variance ($R^2 = 0.5048$), performing no better than the smoking model and worse than the gender model. The unequal slopes model (Model 6B) revealed an interesting pattern: the interaction between moderate alcohol consumption and *Fat* levels was statistically significant ($\beta = -0.832$, $p = 0.026$), suggesting that moderate drinkers have a reduced cholesterol response to *Fat* levels (slope = 2.55) compared to non-drinkers (slope = 3.38). However, the overall nested F-test for both alcohol interactions was only marginally significant ($F_{(2, 309)} = 2.52$, $p = 0.082$), and the improvement in R^2 was modest (0.5048 to 0.5128). Given the marginal significance and the fact that the main effects of alcohol consumption were non-significant, these findings should be interpreted cautiously and may represent chance findings requiring replication.

The full parallel slopes model (Model 7A) included all three categorical variables simultaneously along with *Fat*. This model explained 52.02% of variance in cholesterol ($R^2 = 0.5202$). *Fat* remained the strongest predictor ($\beta = 2.686$, $p < 0.0001$), and gender remained highly significant ($\beta = 48.84$, $p = 0.002$), with males having approximately 49 higher cholesterol levels than females. Current smoking status ($\beta = 10.75$, $p = 0.481$) and both alcohol consumption categories (Some: $\beta = -4.70$, $p = 0.676$; A lot: $\beta = -20.14$, $p = 0.328$) were not

significant predictors. The unequal slopes model (Model 7B) included all four possible interactions between categorical variables and fat intake. None of the individual interaction terms reached statistical significance: Smoke × Fat ($p = 0.198$), Gender × Fat ($p = 0.223$), Alcohol-Some × Fat ($p = 0.087$), and Alcohol-A lot × Fat ($p = 0.259$). The nested F-test comparing the full parallel and unequal slopes models was not significant ($F_{(4, 305)} = 1.66$, $p = 0.159$), confirming that the simpler parallel slopes model (Model 7A) is preferred. The minimal increase in R^2 from 0.5202 to 0.5304 does not justify adding four additional parameters to the model.

Comparing all parallel slopes models revealed a clear hierarchy of predictive power. Model 5A (Gender + Fat) demonstrated the best balance of parsimony and fit with an R^2 of 0.5179. Model 7A (all categorical variables + Fat) achieved the highest R^2 (0.5202), indicating superior fit. The minimal difference in R^2 between Model 5A (0.5179) and Model 7A (0.5202) indicates that adding smoking status and alcohol consumption to a model already containing gender and fat provides negligible additional explanatory power. Based on the principle of parsimony, Model 5A represents the optimal model, though Model 7A provides the most comprehensive assessment of all variables of interest.

Gender is unequivocally the most predictive categorical variable for cholesterol levels. Moreover, Gender demonstrated the strongest and most consistent statistical significance across all models ($p = 0.002$-0.003); while smoking and alcohol consumption were never significant predictors. Additionally, gender contributed substantially to model performance: adding gender to a fat-only model increased R^2 by approximately 0.028 (from ~0.49 to 0.518), whereas smoking and alcohol added virtually no additional explanatory power. Lastly, the gender effect remained robust even when controlling for smoking status and alcohol consumption simultaneously, demonstrating its independent predictive value.

8. Conclusions

This assignment provided an extensive exploration of Analysis of Covariance (ANCOVA) and regression modeling techniques, progressing from basic one-way ANOVA to complex models incorporating multiple categorical predictors, continuous covariates, and their interactions. The progression from Model 1 (ANOVA with smoking groups) through Model 7 (full model with all categorical predictors and interactions) illustrated how different analytical approaches can reveal or obscure important patterns in data.

One of the most important conceptual insights from this assignment was recognizing that ANOVA is simply a special case of multiple regression using dummy-coded categorical variables. Model 1 demonstrated this perfectly: the ANOVA F-test and the regression F-test were identical ($F_{(2, 312)} = 2.235$, $p = 0.109$), and the group means calculated from both approaches were the same. This equivalence reinforces that these are not fundamentally different techniques but rather different frameworks for understanding the same underlying linear model.

The examination of interaction terms provided valuable insights into how relationships can differ across subgroups. Model 3 revealed that *PriorSmoke_2* showed a significantly reduced cholesterol response to fat levels (slope = 2.290) compared to *PriorSmoke_1* (slope = 2.974),

with the interaction term reaching statistical significance ($p = 0.043$). However, when examining current smoking status, gender, and alcohol consumption, most interactions were non-significant, suggesting that the fat-cholesterol relationship is relatively consistent across these demographic subgroups. Learning to test for, interpret, and visualize interactions enhanced my understanding of when parallel slopes models are appropriate versus when more complex unequal slopes models are needed.

A. Appendix

Figure 1. Cholesterol Levels by PriorSmoke Group

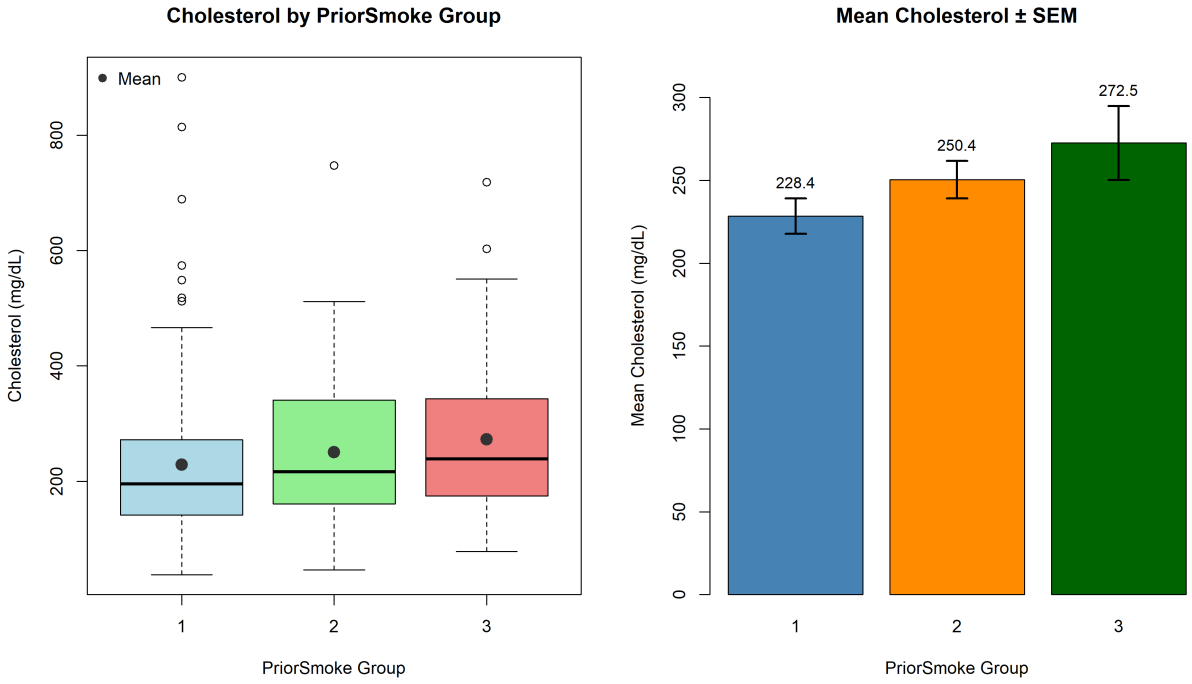

Figure 2. Scatterplot of Residuals and Predicted Values for Model 2

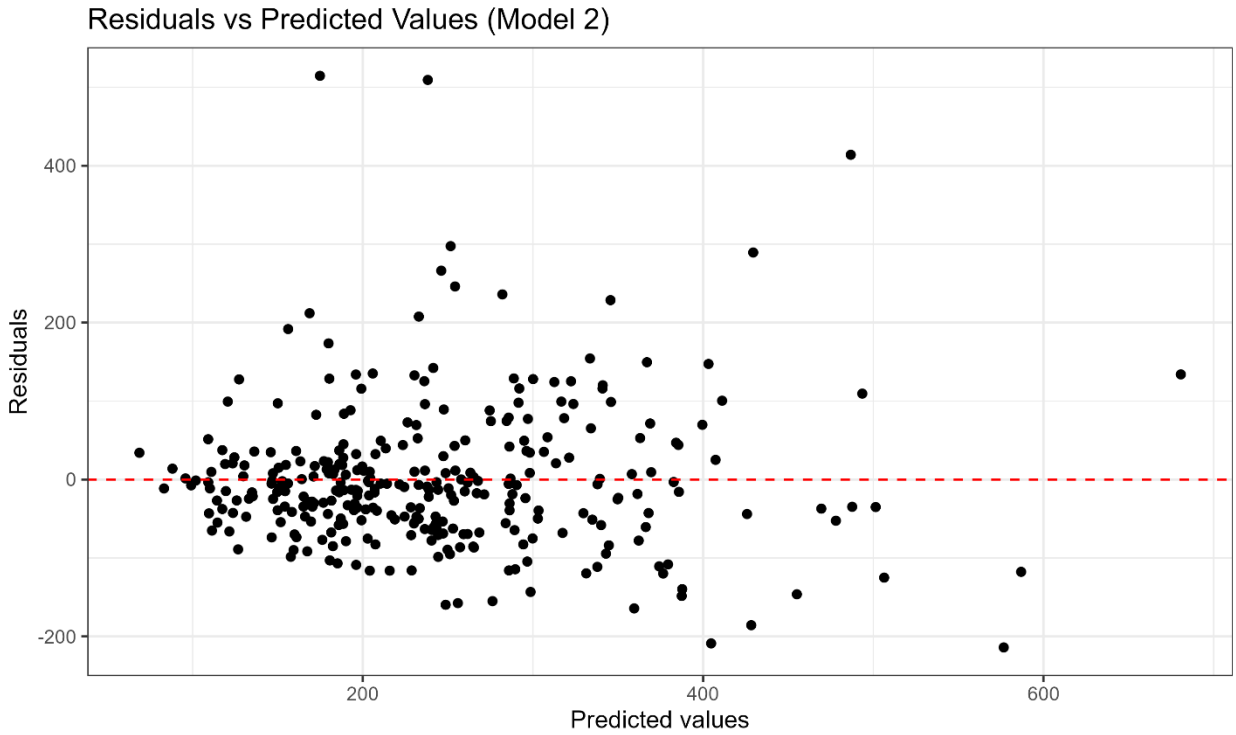

Figure 3. Scatter plot of Predicted Cholesterol Values in Model 2 vs Fat

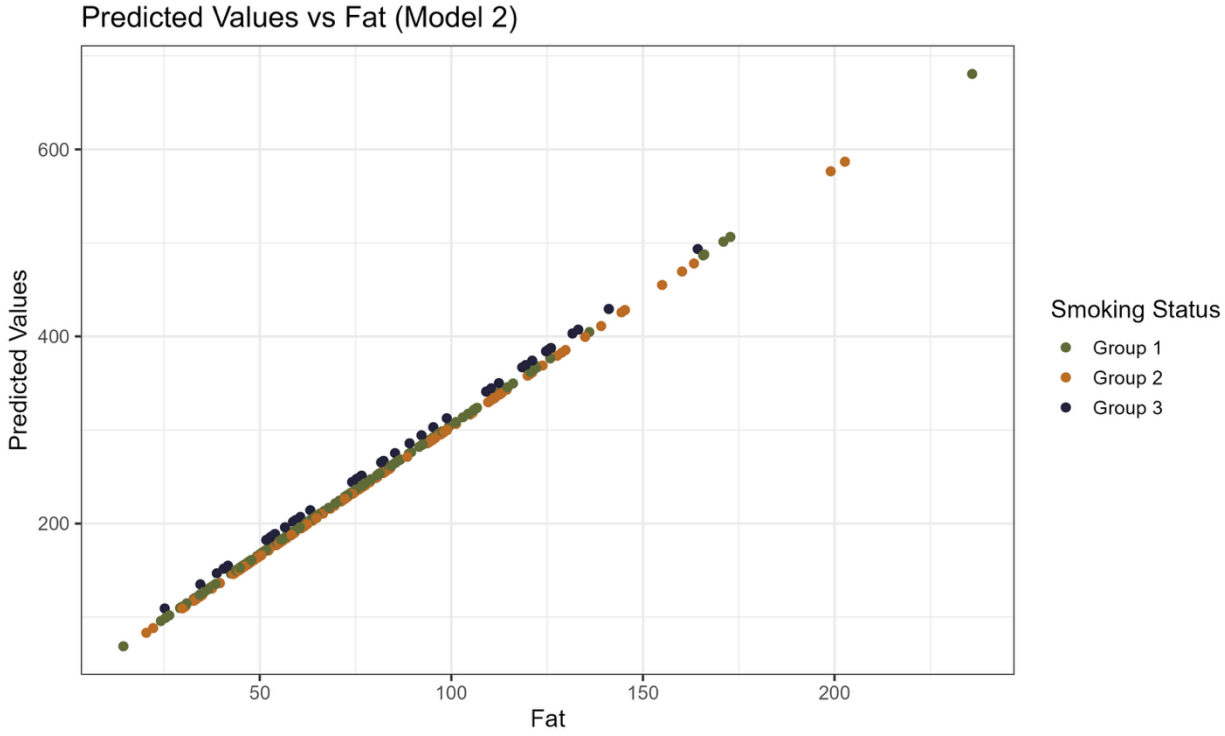

Figure 4. Scatterplot of Actual Cholesterol Values vs Fat

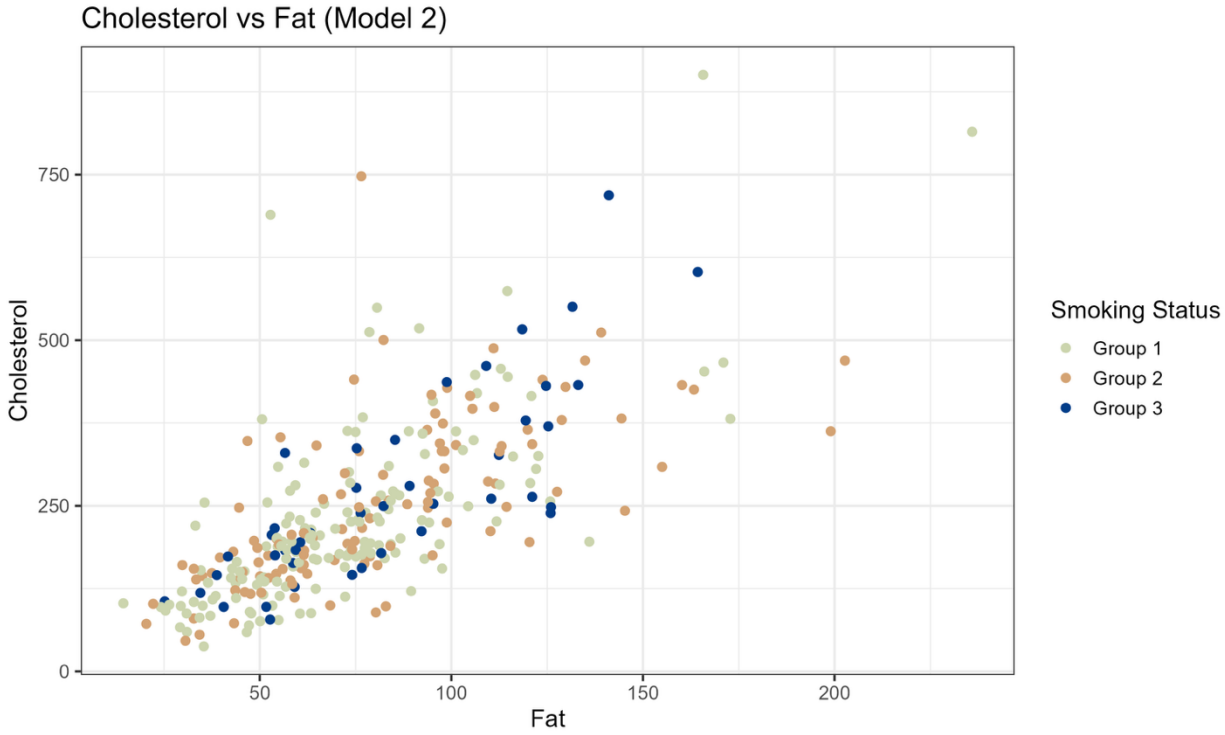

Figure 5. Scatterplot of Residuals and Predicted Values for Model 3

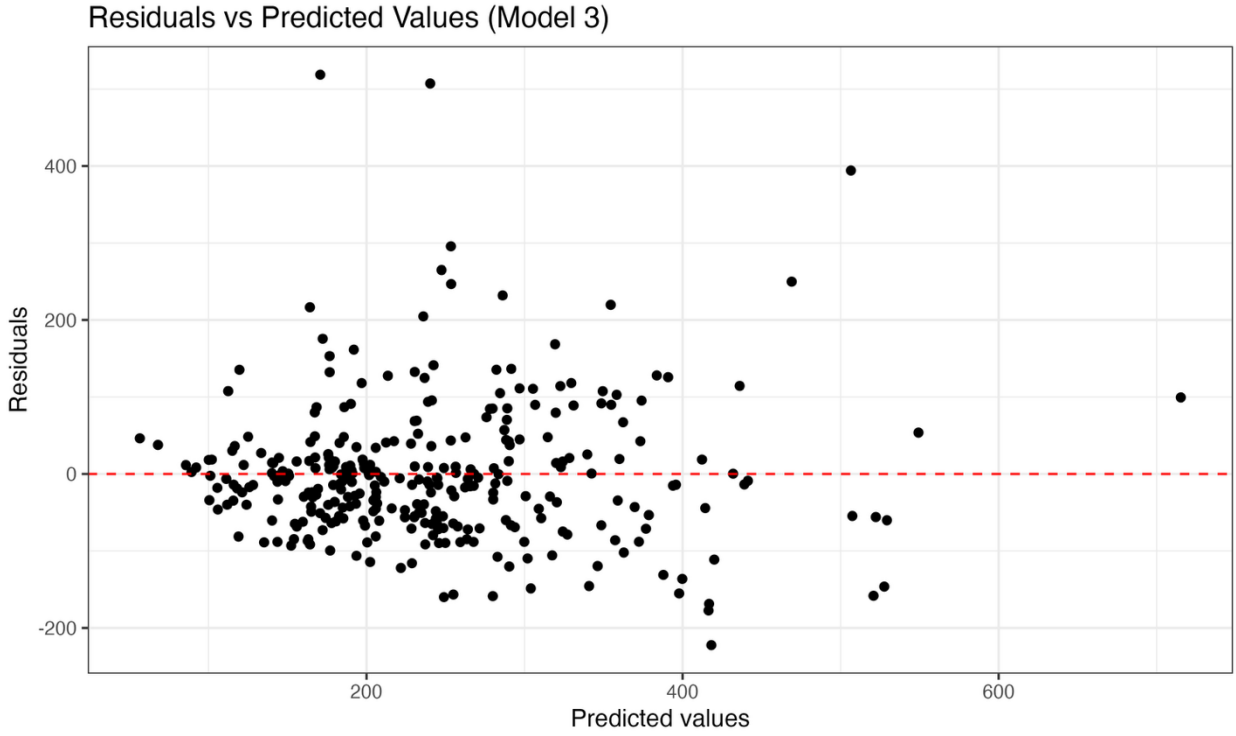

Figure 6. Scatter plot of Predicted Cholesterol Values in Model 3 vs Fat