

Depression and its determinants among adolescents in Jimma town, Southwest Ethiopia

By
Zhen Chen
Duru Demirbag
Fionn Johnson

1. Introduction

This report focuses on analysing factors associated with depression among adolescents in Jimma, a town in Southwest Ethiopia, based on collected data by Girma et al. (2021). The dataset used in this report consists of 546 observations of adolescents aged 14 to 19, with key variables such as age, sex, school grade level, school type, residence, health status, social support score(OSLO3), and Body Mass Index (BMI). The response variable is the PHQ-9 depression score, which ranges from 0 to 15, with higher scores representing high levels of depression. Moreover, the dataset includes a TRAINSET variable, which specifies whether an observation is designated for training (1) or testing (0) purposes. This separation allows for the evaluation of the model by comparing it with the test data, demonstrating how meaningful the created model is. By exploring the relationships between depression scores and various explanatory variables, such as health status, social support, and living conditions, we aim to shed light on the specific challenges faced by adolescents in Jimma. Additionally, this analysis seeks to identify whether the observed relationships are unique to this context or if they align with broader patterns of depression in other African countries or globally.

The main objective of this report is to determine how well we can predict the response variable (PHQ-9 depression score) using the explanatory variables in the dataset. We aim to discover whether all the explanatory variables are necessary to predict the depression score, and if so, are they all of equal importance or do some of the variables have a more significant relationship with the response variable than others? Are there any redundant variables, or variables that do not significantly contribute to the depression score? It is also important to explore the individual effects of predictor variables on the response variable. For example, does the relationship between depression scores and other factors vary by sex? Similarly, does the BMI variable have a direct relationship with depression scores or is its effect mediated by other factors, such as health status? Through the model analysis we aim to answer these questions and discover if there is a subset of explanatory variables that yields a model that best predicts this depression score variable. It's important to highlight the possibility that all the predictor variables in our dataset aren't the most significant for determining depression scores. There could be other confounding variables that we haven't considered that have a greater impact on the depression score. For example, variables such as parent occupation and educational background were also considered in the original report.

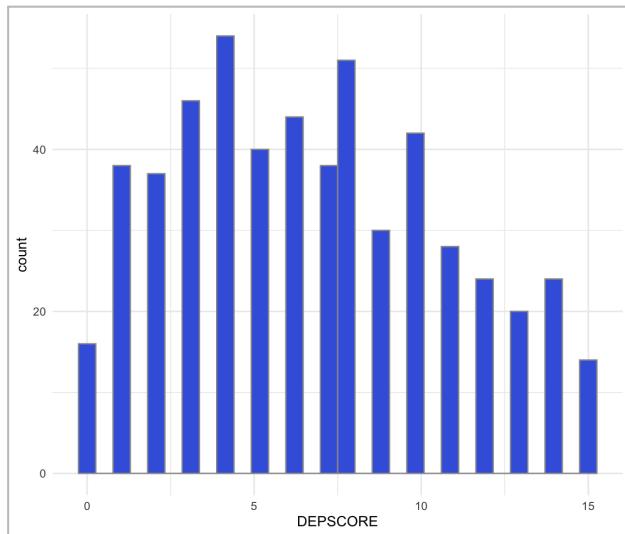
When constructing a linear regression model, several key assumptions must be satisfied to ensure the model's validity. The most important is the assumption of linearity, which requires a linear relationship between the response variable and the predictor variables. Additionally, the residuals of the model should follow a normal distribution, and the variance of the residuals should remain constant across all values of the predictor variables. To construct the model, we will employ linear regression alongside variable selection techniques such as backward elimination and subset selection. The significance of variables will be assessed through hypothesis testing of individual parameters and evaluation of diagnostic metrics, including residual plots and Q-Q plots, to ensure that model assumptions are satisfied.

The report is organized to provide statistical and graphical summaries that give an overview of the data, highlighting notable trends or patterns (Section 2). It details the linear models used to predict depression scores, presents the findings from these models (Section 3), and discusses their accuracy in predicting depression scores on the test data (Section 4). Finally, an appendix includes the R code used for data analysis and visualization in a detailed way (Section 5).

2.Exploratory Analysis through Descriptive statistics and Graphical summaries

This section introduces some important factors we believe to have significant relationships with the depression score. We will examine these factors and discuss their impact on depression levels. The R code used for this section is in Appendix A.2.

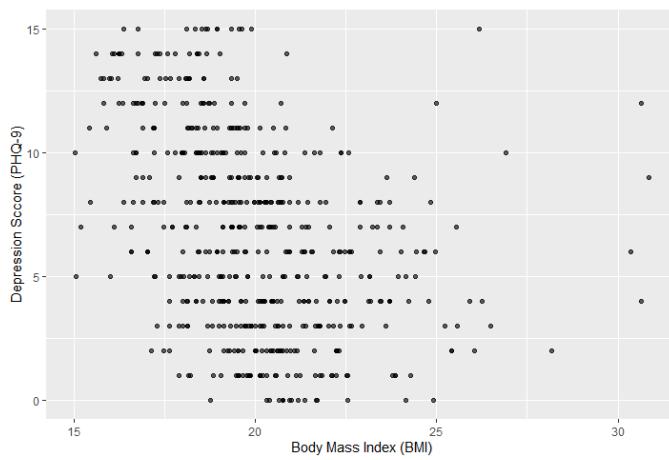
Figure 2.1 Depression Score Distribution



Appendix A.2.1

The distribution of depression scores is presented in the histogram in Figure 2.1. The majority of depression scores fall between 1 and 10 with the most common score amongst teenagers being 4. The skewness and lack of normality of the distribution could be due to external factors or the combined interactions and contributions of the predictor variables. For instance, the extent to which different variables such as Sex or BMI are associated with depression might differ. This highlights the importance of exploring the individual relationships between variables and the depression score to better understand the distribution plot.

Figure 2.2 Scatterplot of BMI against Depression Score

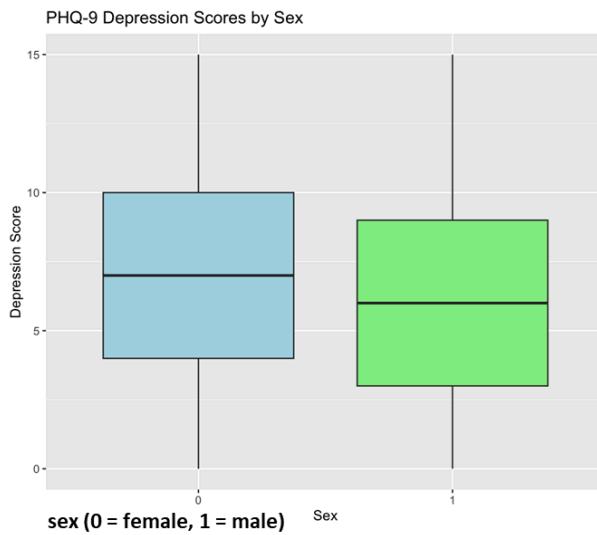


Appendix A.2.2

Figure 2.2 shows a scatterplot illustrating the relationship between the BMI and the Depression Score. There is a lack of linearity between the two variables suggesting that BMI alone doesn't directly predict the depression score. This implies that other factors might be influencing the relationship between BMI and Depression. For instance, a person's BMI could influence their health status or social support score which may subsequently affect their depression score. The shape of the plot indicates a quadratic relationship, where individuals with very low or high BMIs tend to have high depression scores whereas a moderate BMI score might indicate a lower depression score. This complex relationship

suggests that a simple linear model might not be appropriate and the exploration of BMI transformations might be beneficial.

Figure 2.3 Boxplot of Depression Scores by Sex



Appendix A.2.3

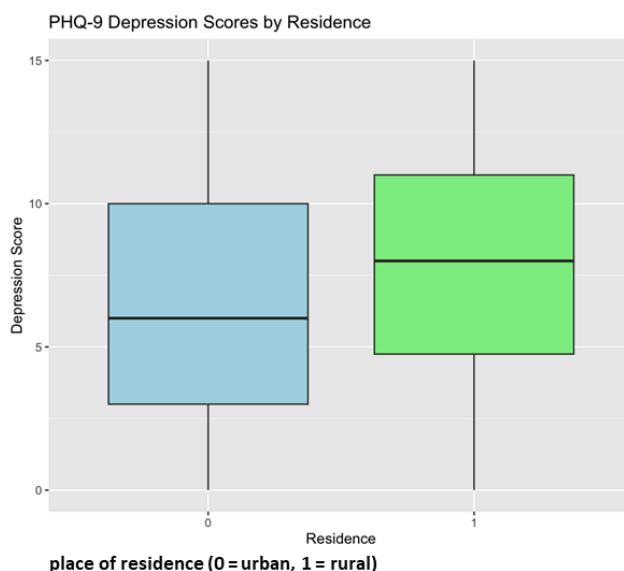
Table 2.1 Summary Statistics of PHQ-9 Depression Scores by Sex

Sex	Mean	Standard Deviation
0	7.030395	4.093308
1	6.373272	3.91505

sex (0 = female, 1 = male)

Figure 2.3 shows a boxplot of the distribution of PHQ-9 depression scores by sex, highlighting the median scores, interquartile range and variability for each group. We can infer that the depression levels in young females are higher than in young males, as clarified in table 2.1. While these differences in depression scores are not extreme, they provide evidence to include this variable in our model. Further investigations could be made into whether sex interacts with other variables to influence depression scores.

Figure 2.4 Boxplot of Depression Scores by Residence



Appendix A.2.4

The distribution of PHQ-9 depression scores by residence is displayed in the boxplot in Figure 2.4. The differences in depression levels between residences, urban and rural, are quite significant as also seen in table 2.2. This suggests that the Residence variable alone is an important predictor of depression scores and supports its inclusion in the final model. From this table, we observe that the average depression levels among adolescents in rural areas are higher than those in urban areas. However further analysis could be made to investigate whether the relationships between other variables and depression scores vary by residence.

Table 2.2 Summary Statistics of PHQ-9 Depression Scores by Residence

Residence	Mean	Standard Deviation
0	6.513333	3.980582
1	7.96875	4.079353

place of residence (0 = urban, 1 = rural)

3.Statistical Methods

In this section, our goal is to predict depression scores (the dependent variable) based on several factors (independent variables). Several linear regression models were developed and tested, each containing different sets of predictor variables. The accuracy and performance of the models were compared using performance metrics and by analysing the diagnostic plots. This approach allowed us to determine which variables were significant in their contribution towards predicting depression scores. All tested models are provided in Appendix A.3.

3.1 Fitting The First Model

To begin, we constructed an initial model that included all the variables from the dataset. The R code is Appendix A.3.1. The linear regression equation for this model is shown below:

$$E(DEPSCORE) = \beta_0 + \beta_1(AGE) + \beta_2(SEX) + \beta_3(GRADE) + \beta_4(SCHOOLTYPE) + \beta_5(RESIDENCE) \\ + \beta_6(HEALTH) + \beta_7(OSL03) + \beta_8(BMI)$$

In this equation β_0 (intercept) represents the expected depression score when all the other predictor numerical variables are equal to 0 and the categorical variables are at their reference levels. The other β terms represent the change in the depression score for a 1-unit increase in their corresponding predictor variable, holding all other variables constant. For example, β_8 represents the expected change in depression score for every 1-unit increase in BMI, assuming all other variables remain unchanged.

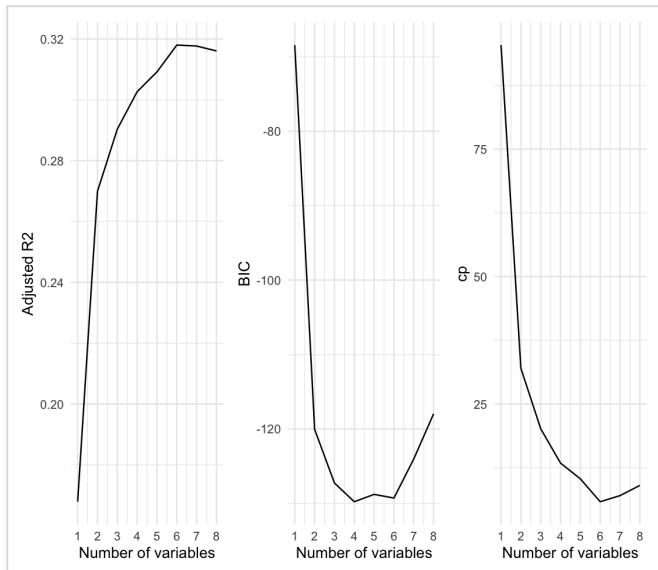
Figure 3.1 Regression Coefficients and Model Performance Metrics for Model 1

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	20.602281	3.365120	6.122	2.10e-09 ***
AGE	0.186083	0.208182	0.894	0.371911
SEX	-0.877978	0.339318	-2.587	0.010000 **
GRADE	0.356329	0.232209	1.535	0.125647
SCHOOLTYPE	-0.007134	0.434241	-0.016	0.986899
RESIDENCE	1.062144	0.432109	2.458	0.014369 *
HEALTH	-0.644225	0.178584	-3.607	0.000346 ***
OSL03	-0.459100	0.064313	-7.139	4.12e-12 ***
BMI	-0.526591	0.068996	-7.632	1.54e-13 ***
<hr/>				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 3.38 on 424 degrees of freedom				
Multiple R-squared: 0.3288, Adjusted R-squared: 0.3161				
F-statistic: 25.96 on 8 and 424 DF, p-value: < 2.2e-16				

Figure 3.1 displays the model's performance metrics for each predictor variable included. A hypothesis test was performed on each variable, where the null hypothesis states that the variable does not affect the response (i.e., the corresponding β coefficient is 0). The p-values represent the probability of obtaining the test statistic under the null hypothesis. From the figure, we observe that there is insufficient evidence that age, school type and grade are significant contributors towards the model. This is based on a significance level of $\alpha = 0.05$.

3.2 Variable-Based Model Selection and Analysis

Figure 3.2 Model Selection Criteria



Appendix A.3.4

To confirm these results and further refine our model, we used subset selection. This method evaluates all possible combinations of variables to produce the subset of variables with the lowest Residual Sum of Squares (RSS) value. The results from subset selection are visualised in plots of Adjusted R^2 , Bayesian Information Criterion (BIC) and Mallows Cp against the number of variables, shown in Figure 3.2. These plots evaluate how the number of variables affects performance of the model.

The Adjusted R^2 plot shows that model performance improves as more variables are added, reaching a peak at 6 variables. After this point, the Adjusted R^2 begins to decline, suggesting that including further variables does not improve the model's performance and may lead to overfitting. Similarly, both the BIC and Mallows' Cp decrease with more variables. For both metrics, the turning point occurs at 6 variables, where the scores are minimized. Beyond this, the values start increasing, indicating that including more than 6 variables worsens the model's performance.

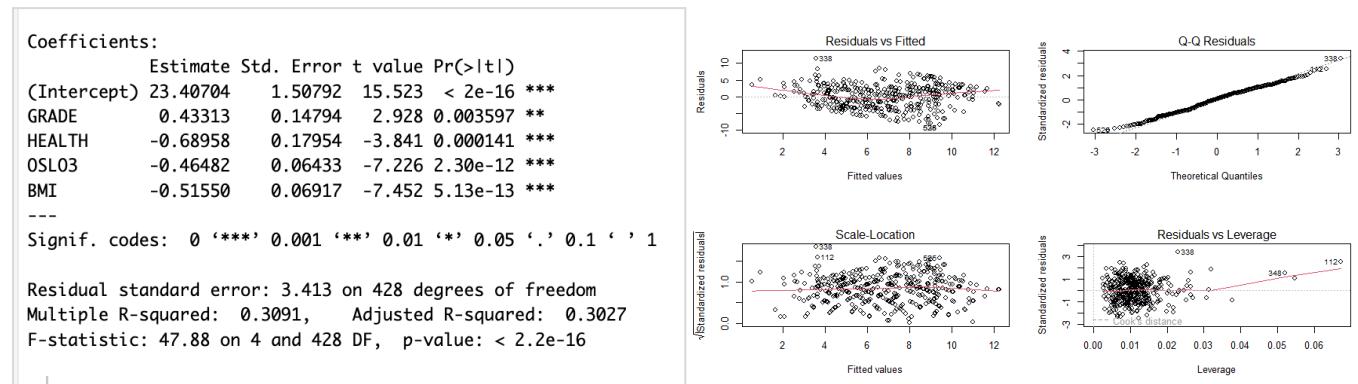
The hypothesis tests and the subset selection lead to the same conclusion: removing age and school-type variables would benefit the model. Furthermore, the BIC plot indicated that a model with four variables had the best BIC score. Based on these results, further investigations were made into the model with 6 variables and the model with 4 variables.

3.2.1 Model 4

The R code for Model 4 is in Appendix A.3.5.

$$E(DEPSCORE) = \beta_0 + \beta_1(GRADE) + \beta_2(HEALTH) + \beta_3(OSL03) + \beta_4(BMI)$$

Figure 3.3 Regression Coefficients and Model Performance Metrics for Model 4

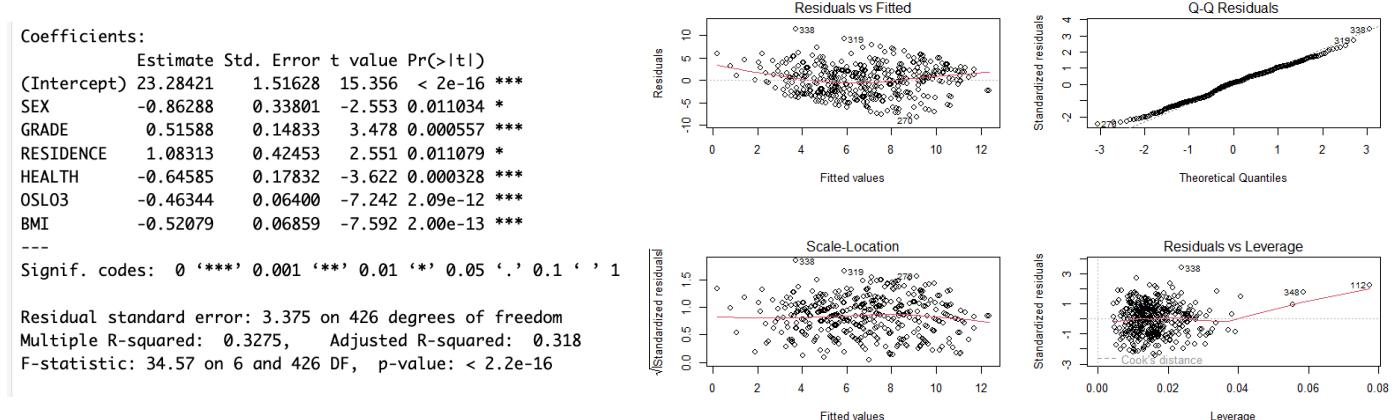


3.2.2 Model 6

The R code for Model 6 is in Appendix A.3.6.

$$E(DEPSCORE) = \beta_0 + \beta_1(SEX) + \beta_2(GRADE) + \beta_3(RESIDENCE) + \beta_4(HEALTH) + \beta_5(OSLO3) + \beta_6(BMI)$$

Figure 3.4 Regression Coefficients and Model Performance Metrics for Model 6



We concluded from the diagnostic plots for Model 4, that there was insufficient evidence to drop the Sex and Age variables, so decided to discontinue further investigations with this model. Based on Figure 3.4, both the variables Sex and Residence have p-values smaller than 0.05, providing significant evidence to reject the null hypothesis that they have no effect on the Depression Score. Consequently, the model with all six variables (Model 6) was selected.

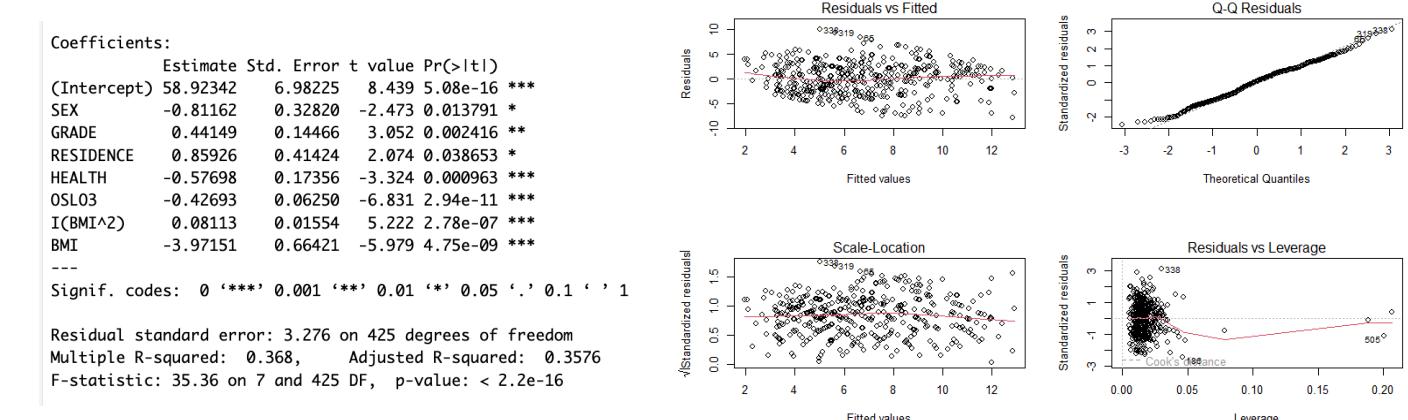
Examining the residual plot, there is no extreme funneling, and residuals do not form obvious clusters or distinct shapes, which suggests that homoscedasticity is mostly satisfied. However, the curvature of the red line indicates that the linearity assumption might not hold perfectly. Additionally, the Q-Q plot of Model 6 shows that data are not normally distributed, indicating the need for further improvements to Model 6.

3.2.3 Model 6.1

The R code for Model 6.1 is in Appendix A.3.8.

$$E(DEPSCORE) = \beta_0 + \beta_1(SEX) + \beta_2(GRADE) + \beta_3(RESIDENCE) + \beta_4(HEALTH) + \beta_5(OSLO3) \\ + \beta_6(BMI)^2 + \beta_7(BMI)$$

Figure 3.5 Regression Coefficients and Model Performance Metrics for Model 6.1



In Model 6.1, a polynomial term was added to address the curve seen in the red line of the residuals plot in Model 6. This change improved the model, as the red line became flatter in the residuals plot, and the Q-Q plot showed a more normal distribution.

3.2.4 Model 6.2

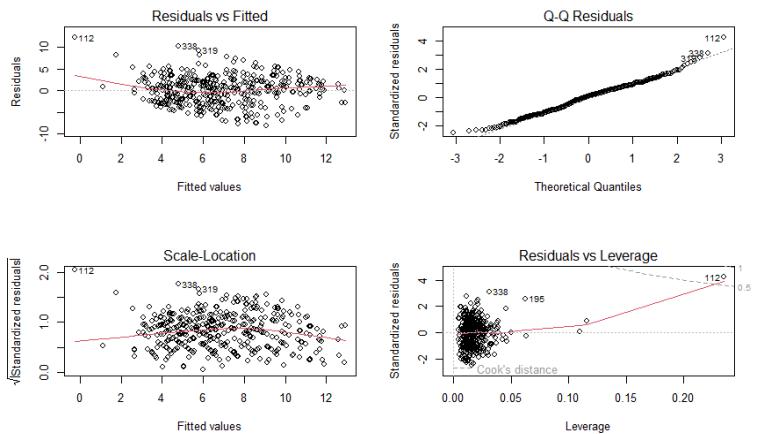
The R code for Model 6.2 is in Appendix A.3.8.

$$E(DEPSCORE) = \beta_0 + \beta_1(SEX) + \beta_2(GRADE) + \beta_3(RESIDENCE) + \beta_4(HEALTH) + \beta_5(OSLO3) \\ + \beta_6(BMI) + \beta_7(OSLO3:BMI)$$

Figure 3.6 Regression Coefficients and Model Performance Metrics for Model 6.2

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 41.89223   5.20780   8.044 8.74e-15 ***
SEX          -0.78998   0.33358  -2.368 0.018322 *  
GRADE        0.46528   0.14676   3.170 0.001632 ** 
RESIDENCE    1.15236   0.41865   2.753 0.006166 ** 
HEALTH       -0.68298   0.17596  -3.882 0.000120 *** 
OSLO3        -2.34017   0.50710  -4.615 5.22e-06 *** 
BMI          -1.45079   0.25834  -5.616 3.54e-08 *** 
OSLO3:BMI    0.09422   0.02526   3.730 0.000218 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.325 on 425 degrees of freedom
Multiple R-squared:  0.3488, Adjusted R-squared:  0.3381 
F-statistic: 32.52 on 7 and 425 DF,  p-value: < 2.2e-16
```



In Model 6.2, an interaction term was added because the GGPLOT (in the Appendix) showed some interaction between variables. This change improved the Q-Q plot, which became closer to normal, but the residual plot did not improve. When comparing Model 6.1 and Model 6.2, Model 6.1 has a higher multiple R-squared value (0.368) than Model 6.2 (0.3488). However, the difference is small, so further comparisons will be needed.

3.2.4 Model 6.12

The R code for Model 6.12 is in Appendix A.3.8.

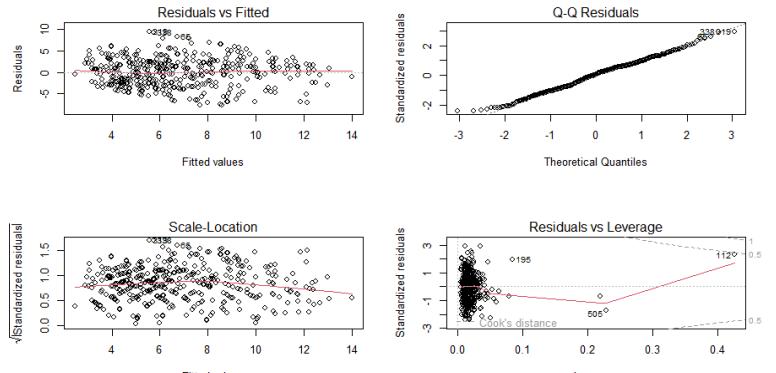
$$E(DEPSCORE) = \beta_0 + \beta_1(SEX) + \beta_2(GRADE) + \beta_3(RESIDENCE) + \beta_4(HEALTH) + \beta_5(OSLO3) \\ + \beta_6(BMI) + \beta_7(OSLO3:BMI) + \beta_8(BMI)^2$$

Figure 3.612 Regression Coefficients and Model Performance Metrics for Model 6.12

```
Residuals:
    Min      1Q      Median      3Q      Max
-7.6600 -2.4761  0.2012   2.1476  9.4197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 69.78034   7.85614   8.882 < 2e-16 ***
SEX          -0.76013   0.32580  -2.333 0.020110 *  
GRADE        0.40968   0.14381   2.849 0.004601 ** 
RESIDENCE    0.93575   0.41145   2.274 0.023450 *  
HEALTH       -0.61282   0.17248  -3.553 0.000424 *** 
OSLO3        -1.88953   0.50456  -3.745 0.000205 *** 
BMI          -4.34509   0.67071  -6.478 2.57e-10 *** 
I(BMI^2)     0.07291   0.01566   4.657 4.29e-06 *** 
OSLO3:BMI    0.07325   0.02508   2.921 0.003677 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.247 on 424 degrees of freedom
Multiple R-squared:  0.3805, Adjusted R-squared:  0.3688 
F-statistic: 32.55 on 8 and 424 DF,  p-value: < 2.2e-16
```



In Model 6.12, both polynomial term and interaction term were added based on the model 6.1 and 6.2. This change improved both residual plot and Q-Q plot. However, the multiple R-squared value (0.3805) indicates that only 38.05% of the variability in depression scores can be explained by the predictor variables. Therefore, further investigation is required.

3.2.4 Models Dropped

The R code is in Appendix A.3.8.1

Several alternative models were also tested; however, as they did not show notable improvements, they were excluded from further comparisons.

Model 6.3:

Figure 3.7.1 Regression Coefficients and Model Performance Metrics for Model 6.3

```
call:
lm(formula = DEPSCORE ~ RESIDENCE * GRADE + SEX * RESIDENCE +
RESIDENCE * HEALTH + RESIDENCE * OSLO3 + RESIDENCE * BMI +
RESIDENCE, data = modeldata)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.1351 -2.7036  0.2834  2.4519 11.0870 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 22.81289  1.68604 13.530 < 2e-16 ***
RESIDENCE   4.71518  4.03092  1.170 0.242762    
GRADE        0.52242  0.16168  3.231 0.001329 **  
SEX          -0.86182  0.37693 -2.286 0.022727 *   
HEALTH       -0.70624  0.19741 -3.578 0.000387 *** 
OSLO3        -0.41420  0.07048 -5.877 8.5e-09 *** 
BMI          -0.51116  0.07952 -6.428 3.5e-10 *** 
RESIDENCE:GRADE -0.21330  0.42431 -0.503 0.615436    
RESIDENCE:SEX   -0.22733  0.87534 -0.260 0.795218    
RESIDENCE:HEALTH 0.33994  0.48118  0.706 0.480280    
RESIDENCE:OSLO3 -0.33152  0.17665 -1.877 0.061252 .  
RESIDENCE:BMI    -0.06216  0.16190 -0.384 0.701232  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.38 on 421 degrees of freedom
Multiple R-squared:  0.3337,    Adjusted R-squared:  0.3162 
F-statistic: 19.16 on 11 and 421 DF,  p-value: < 2.2e-16
```

In this model, we investigated whether the dummy variable "Residence" interacts with other variables and influences the model. As shown in the results above, all the p-values for the interaction terms are greater than 0.05, indicating that the interactions between these variables may not significantly affect the dependent variable. Therefore, there is insufficient evidence to justify including these interaction terms in the model. As a result, we will no longer continue with this model.

Model 6.4:

The R code for Model 6.4 is in Appendix A.3.8.

Figure 3.7.2 Regression Coefficients and Model Performance Metrics for Model 6.4

```
call:
lm(formula = DEPSCORE ~ SEX * GRADE + SEX * RESIDENCE + SEX * 
    HEALTH + SEX * OSLO3 + SEX * BMI + SEX, data = modeldata)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0182 -2.6804  0.2788  2.3556 11.2399 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 22.53564  1.87287 12.033 < 2e-16 ***
SEX          1.10598  3.21105  0.344  0.73070    
GRADE        0.56772  0.19401  2.926  0.00362 **  
RESIDENCE   1.03417  0.59258  1.745  0.08168 .  
HEALTH      -0.49479  0.22869 -2.164  0.03106 *   
OSLO3        -0.56023  0.08434 -6.642 9.57e-11 *** 
BMI          -0.46888  0.08637 -5.429 9.62e-08 *** 
SEX:GRADE   -0.11927  0.30184 -0.395  0.69292    
SEX:RESIDENCE 0.10859  0.85763  0.127  0.89930    
SEX:HEALTH   -0.33362  0.36971 -0.902  0.36737    
SEX:OSLO3    0.22214  0.13068  1.700  0.08991 .  
SEX:BMI      -0.13154  0.14352 -0.917  0.35991  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.38 on 421 degrees of freedom
Multiple R-squared:  0.3336,    Adjusted R-squared:  0.3161 
F-statistic: 19.16 on 11 and 421 DF,  p-value: < 2.2e-16
```

In this model, we examined whether the dummy variable "Sex" interacts with other variables and influences the model. This investigation relates to one of the earlier questions posed back in the analysis about whether the relationship between predictors and depression scores varies with sex. The results conclude there is insufficient evidence to include these interaction terms in our model as the p-values exceed the significance level of 0.05. These findings suggest that even though sex is an important predictor on its own, its interactions with other variables do not significantly influence the depression scores. Moreover, it implies that all the predictor variables affect depression scores similarly across the sexes. No further investigations with this model took place due to the insignificance of the interaction terms.

Model 6.6:

The R code for Model 6.6 is in Appendix A.3.8.

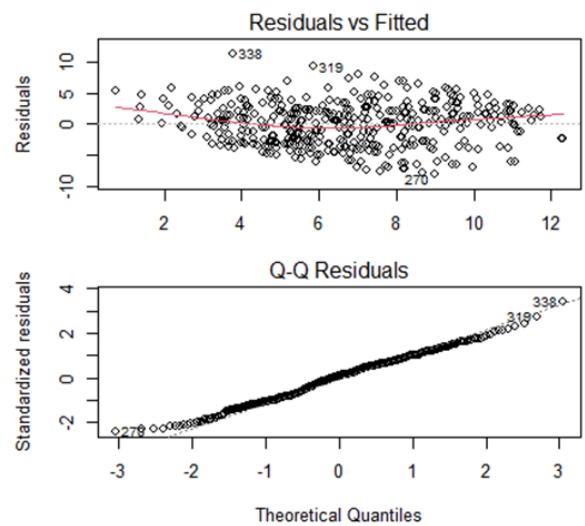
Figure 3.7.4 Regression Coefficients and Model Performance Metrics for Model 6.6

```
Call:
lm(formula = DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 +
log(BMI), data = modeldata)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.0069 -2.6873  0.3087  2.2602 11.2416 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 46.9840    4.1774 11.247 < 2e-16 ***
SEX         -0.8621    0.3352 -2.572 0.010452 *  
GRADE        0.5003    0.1472  3.398 0.000743 *** 
RESIDENCE   1.0400    0.4213  2.469 0.013947 *  
HEALTH       -0.6304    0.1770 -3.562 0.000409 *** 
OSLO3        -0.4532    0.0636 -7.126 4.44e-12 ***
log(BMI)    -11.4550   1.4145 -8.098 5.92e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 3.348 on 426 degrees of freedom
Multiple R-squared:  0.3383,    Adjusted R-squared:  0.329 
F-statistic: 36.31 on 6 and 426 DF,  p-value: < 2.2e-16
```



For this model we applied a log of "BMI" to see if it would improve Model 6. As shown in Figure 3.7.4, the Multiple R-squared value of 0.3383 is higher compared to Model 6. However, the residuals plot and Q-Q plot did not show any improvement. Therefore, we discarded this model.

3.3 Model Comparison and Selection

The R code for comparison and selection is in Appendix A.3.9.

After evaluating several models, three models stood out based on their R^2 values and diagnostic plots: Model 6, Model 6.1, Model 6.2 and Model 6.12. We assessed each model using the test dataset to determine which was best for predicting depression scores. The Root Mean Square Error (RMSE) was measured and compared for each model on the test data. The RMSE is an average measure of the difference between the predicted values obtained by the model and the actual values from the test data. It's a great way of evaluating a model's accuracy in predicting a response variable. Table 3.1 displays the results of the RMSE for each model.

Table 3.1 Root Mean Squared Error (RMSE) of the Models in test data.

Model	RMSE
6	3.211518
6.1	3.074088
6.2	3.105501
6.12	3.054343

Among the models, Model 6.12 had the lowest RMSE, indicating that this model had the highest accuracy in predicting depression scores.

4. Results and Conclusions

This section presents the results of the statistical analysis to investigate how well we can predict depression scores among adolescents with various factors.

After developing and testing several models, Model 6.12 produced the most accurate results. The model included the best subset of predictor variables based on hypothesis testing and subset selection which resulted in the highest R² value during training. Additionally, Model 6.12 displayed the best diagnostic plots, particularly the residuals vs fitted values and QQ plot, suggesting that the assumptions of linear regression, (such as homoscedasticity and linearity) were largely met. Furthermore, when applied to the test data, Model 6.12 produced the lowest RMSE value, confirming it as the better model for predicting depression scores.

Figure 4.1 Regression Coefficients and Model Performance Metrics for Model 6.12

```
call:
lm(formula = DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 +
    BMI + I(BMI^2) + OSLO3 * BMI, data = modeldata)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.6600 -2.4761  0.2012  2.1476  9.4197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 69.78034   7.85614   8.882 < 2e-16 ***
SEX          -0.76013   0.32580  -2.333 0.020110 *  
GRADE        0.40968   0.14381   2.849 0.004601 ** 
RESIDENCE    0.93575   0.41145   2.274 0.023450 *  
HEALTH       -0.61282   0.17248  -3.553 0.000424 *** 
OSLO3        -1.88953   0.50456  -3.745 0.000205 *** 
BMI          -4.34509   0.67071  -6.478 2.57e-10 *** 
I(BMI^2)      0.07291   0.01566   4.657 4.29e-06 *** 
OSLO3:BMI    0.07325   0.02508   2.921 0.003677 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.247 on 424 degrees of freedom
Multiple R-squared:  0.3805, Adjusted R-squared:  0.3688 
F-statistic: 32.55 on 8 and 424 DF,  p-value: < 2.2e-16
```

Based on figure 4.1, the final linear regression equation (with predictor coefficients rounded to 2dp) will be:

$$E(DEPSCORE) = 69.78 - 0.76(SEX) + 0.41(GRADE) + 0.94(RESIDENCE) - 0.61(HEALTH) \\ - 1.89(OSLO3) - 4.35(BMI) + 0.07(OSLO3:BMI) + 0.07(BMI)^2$$

From this equation we can observe the numerical relationships between the predictor variables and the depression scores. Each coefficient represents the expected change in depression score associated with a one-unit change in the corresponding predictor variable, assuming all the other variables remain unchanged. For example, girls have PHQ-9 Depression scores that are on average 0.76 points higher compared to boys. Similarly, teenagers living in rural areas have depression scores that are on average 0.94 points higher than those in urban areas. Each one point increase in health score is associated with a 0.61 point decrease in depression score. However, moving up a grade corresponds to an average increase of 0.41 points in depression score.

In our model, the BMI score and OSLO3 score have the most interesting relationship with depression score. When the squared BMI term and interaction term (OSLO3: BMI) was introduced, the model's accuracy in predicting depression scores improved, suggesting a non-linear relationship between the two terms. This discovery links back to one of the questions posed in the introduction, whether BMI has a direct relationship with depression scores or if its effect is mediated by other factors. Now, significant interactions between BMI and OSLO3 were identified, applying both squared BMI term and interaction term (OSLO3: BMI) to our model did significantly improve the performance metrics and diagnostic plots, showing it better reflects the relationship between BMI, OSLO3 and depression score.

Table 4.1 Performance Metrics of Model 6.12[Appendix A.4](#)

Model	RMSE	Correlation	R-squared
6.12	3.054343	0.6133986	0.3762578

RMSE measures the average difference between the actual and predicted values. A lower RMSE indicates better model performance. In this case, an RMSE of 3.05 suggests that, on average, the model's predictions typically deviate from the actual depression scores by around 3.05 units. The R² value measures the proportion of variance in the response variable that is explained by the predictor variables in the model. Based on our results displayed in Table 4.1, an R² of 0.376, means that 37.6% of the variability in depression scores can be explained by the predictors included in the model, while the remaining 62.4% of the variance is due to factors not captured by the model. This suggests that while the model does provide some insights into the factors influencing depression scores, there is still room for further improvement.

In the introduction, we highlighted the possibility of confounding variables influencing depression scores, as the predictor variables in the dataset might not contribute to all the variability in these scores. Based on the low R² value, it seems increasingly likely that our initial assumptions were true and that there are other factors influencing the depression scores that haven't been taken into account. Additionally, it's important to remember that the model was developed using data collected from Jimma in Southwest Ethiopia and its predictive accuracy of depression scores will differ when applied to populations from other countries. These limitations emphasise the need for further research to include other predictor variables and a more diverse population data to improve the model's performance.

Appendix A: R Code and Output

>#A.1 R code for loading data

```
> group3<-read.csv("Group3.csv",header = T)
> names(group3)
> [1] "ID"           "AGE"          "SEX"          "GRADE"        "SCHOOLTYPE"
  "RESIDENCE"     "HEALTH"
[8] "OSLO3"         "BMI"          "DEPSCORE"      "TRAINTEST"
> #split date to test data and train data.
> modeldata<-subset(group3,TRAINTEST==1)
>testdata<-subset(group3,TRAINTEST==0)
```

>#Appendix A.2 R code for creating boxplots and histograms

>#Appendix A.2.1 R code for plotting the histogram of depression scores.

```
> library(ggplot2)
histogram shows the distribution of depression scores of the observation, but
the graph didn't indicate there is normal distribution.
>ggplot(modeldata)+theme_minimal()+geom_histogram(aes(x=DEPSCORE),fill="royalB
lue",color="grey60")
```

>#Appendix A.2.2 R code for plotting BMI against Depression Score

```
# Scatter plot of BMI against Depression score
>ggplot(group3, aes(x = BMI, y = DEPSCORE)) +
  geom_point(color = "black", alpha = 0.6) +
  labs(
    title = "Relationship between BMI and Depression Score",
    x = "Body Mass Index (BMI)",
    y = "Depression Score (PHQ-9)"
  ) +
  theme_gray()
```

>#Appendix A.2.3 R code for box plot to examine the relationship between sex and depression scores.

#There is a box plot between sex factor and shows that sex factor may impact the DEPSCORE.

```
> ggplot(modeldata, aes(x = factor(SEX), y = DEPSCORE)) +
  geom_boxplot(fill = c("lightblue", "lightgreen")) +
  labs(title = "PHQ-9 Depression Scores by Sex", x = "Sex", y = "Depression
Score")
```

>#Appendix A.2.4 R code for box plot to examine the relationship between residence and depression scores

#box plot between Residence factor and shows the residence factor may impact the DEPSCORE.

```
> ggplot(modeldata, aes(x = factor(RESIDENCE), y = DEPSCORE)) +
  geom_boxplot(fill = c("lightblue", "lightgreen")) +
  labs(title = "PHQ-9 Depression Scores by Residence", x = "Residence", y =
"Depression Score")
```

>#Appendix A.3: R Code for Linear Regression Models

>#Appendix A.3.1: R Code for Correlation Analysis

```
>library(GGally)
```

#ggpairs indicate that the same variable is affected by each other. eg,grade and age,grade and id,residence and school type,oslo3 and health state,bmi and oslo3, and residence,health, oslo3,bmi have effect on depscore.
ggpairs(modeldata)

>#Appendix A.3.2: R Code for Initial Linear Model

```

#build a linear model with all the variables which shows SCHOOL TYPE, Age and
Grade have high P-value, there is no significant evidence to reject that they
have no effect on DEPSCORE.
> model1<-lm(DEPSCORE~AGE+SEX+GRADE+SCHOOLTYPE+RESIDENCE+HEALTH+OSLO3+BMI,data
= modeldata)
> summary(model1)

Call:
lm(formula = DEPSCORE ~ AGE + SEX + GRADE + SCHOOLTYPE + RESIDENCE +
    HEALTH + OSLO3 + BMI, data = modeldata)

Residuals:
      Min        1Q    Median        3Q       Max
-8.1808 -2.6984  0.3277  2.2673 11.3963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.602281   3.365120   6.122 2.10e-09 ***
AGE          0.186083   0.208182   0.894 0.371911
SEX          -0.877978   0.339318  -2.587 0.010000 **
GRADE         0.356329   0.232209   1.535 0.125647
SCHOOLTYPE   -0.007134   0.434241  -0.016 0.986899
RESIDENCE     1.062144   0.432109   2.458 0.014369 *
HEALTH        -0.644225   0.178584  -3.607 0.000346 ***
OSLO3         -0.459100   0.064313  -7.139 4.12e-12 ***
BMI           -0.526591   0.068996  -7.632 1.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.38 on 424 degrees of freedom
Multiple R-squared:  0.3288, Adjusted R-squared:  0.3161
F-statistic: 25.96 on 8 and 424 DF, p-value: < 2.2e-16

> round(coef(summary(model1)),3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.602      3.365      6.122      0.000
AGE          0.186      0.208      0.894      0.372
SEX          -0.878      0.339     -2.587      0.010
GRADE         0.356      0.232      1.535      0.126
SCHOOLTYPE   -0.007      0.434     -0.016      0.987
RESIDENCE     1.062      0.432      2.458      0.014
HEALTH        -0.644      0.179     -3.607      0.000
OSLO3         -0.459      0.064     -7.139      0.000
BMI           -0.527      0.069     -7.632      0.000

```

>**#Appendix A.3.3: R Code for Subset Selection Using Exhaustive Search**

```

#running a double check to make sure about our result.
#Both adjusted R2 and CP plots indicate we should keep 6 variables. BIC
indicates about 4 variables.
> library(leaps)
> check_model<-regsubsets(DEPSCORE~AGE+SEX+GRADE+SCHOOLTYPE+RESIDENCE+HEALTH+OS
LO3+BMI,data = modeldata)
> summary(check_model)
Subset selection object
Call: regsubsets.formula(DEPSCORE ~ AGE + SEX + GRADE + SCHOOLTYPE +
    RESIDENCE + HEALTH + OSLO3 + BMI, data = modeldata)
8 Variables (and intercept)
      Forced in Forced out
AGE             FALSE      FALSE
SEX             FALSE      FALSE

```

```

GRADE      FALSE    FALSE
SCHOOLTYPE FALSE    FALSE
RESIDENCE   FALSE    FALSE
HEALTH     FALSE    FALSE
OSLO3      FALSE    FALSE
BMI        FALSE    FALSE
1 subsets of each size up to 8
Selection Algorithm: exhaustive
      AGE SEX GRADE SCHOOLTYPE RESIDENCE HEALTH OSLO3 BMI
1 ( 1 ) " " " " " " " "
2 ( 1 ) " " " " " " " "
3 ( 1 ) " " " " " " " "
4 ( 1 ) " " " " * " " " "
5 ( 1 ) " " " * " * " " " "
6 ( 1 ) " " " * " * " " " "
7 ( 1 ) " * " " * " * " " " "
8 ( 1 ) " * " " * " * " " * "

```

#It shows the variables we need to select based on how many variables we will use to create the model.

>#Appendix A.3.4: R Code for Visualizing Model Selection Metrics

```

> library(patchwork)
> results_metrics<-summary(check_model)
> library(patchwork)
> results_metrics<-summary(check_model)
>
data_plot<-data.frame(adjr2=results_metrics$adjr2,bic=results_metrics$bic,cp=r
esults_metrics$cp)
> g<-ggplot(data_plot)+theme_minimal()+aes(x=1:8)+labs(x="Number of
variables")+scale_x_continuous(breaks = 1:8)
> g1<-g+geom_line(aes(y=adjr2))+labs(y="Adjusted R2")
> g2<-g+geom_line(aes(y=bic))+labs(y="BIC")
> g3<-g+geom_line(aes(y=cp))+labs(y="cp")
> g1+g2+g3

```

>#Appendix A.3.5: R Code for Building and Evaluating the Linear Regression Model with Four Variables

```

#build a linear regression model for 4 variables, and they all have very small
p-value which small that 0.05, pleased to keep them.
> best_model4<-lm(DEPSCORE~GRADE+HEALTH+OSLO3+BMI,data = modeldata)
> round(coef(summary(best_model4)),3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.407     1.508  15.523  0.000
GRADE        0.433     0.148   2.928  0.004
HEALTH       -0.690     0.180  -3.841  0.000
OSLO3        -0.465     0.064  -7.226  0.000
BMI          -0.515     0.069  -7.452  0.000

```

>#Appendix A.3.6: R Code for Building and Evaluating the Linear Regression Model with Six Variables

```

#build a linear regression model for 6 variable, and they all have very small
p-value which small that 0.05, pleased to keep them.

```

```

> best_model6<-lm(DEPSCORE~SEX+GRADE+RESIDENCE+HEALTH+OSLO3+BMI,data =
modeldata)
> round(coef(summary(best_model6)),3)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.284     1.516  15.356  0.000
SEX         -0.863     0.338  -2.553  0.011
GRADE        0.516     0.148   3.478  0.001
RESIDENCE   1.083     0.425   2.551  0.011
HEALTH       -0.646     0.178  -3.622  0.000

```

OSLO3	-0.463	0.064	-7.242	0.000
BMI	-0.521	0.069	-7.592	0.000

>#Appendix A.3.7: R Code for Backward Elimination

>#Running backward regression and giving 6 variable linear regression model same as the best_model6 we built before.

```
> model_backward<-step(model1,direction = "backward")
```

Start: AIC=1063.57

DEPSCORE ~ AGE + SEX + GRADE + SCHOOLTYPE + RESIDENCE + HEALTH + OSLO3 + BMI

	Df	Sum of Sq	RSS	AIC
- SCHOOLTYPE	1	0.00	4843.7	1061.6
- AGE	1	9.13	4852.9	1062.4
<none>			4843.7	1063.6
- GRADE	1	26.90	4870.6	1064.0
- RESIDENCE	1	69.02	4912.8	1067.7
- SEX	1	76.48	4920.2	1068.3
- HEALTH	1	148.66	4992.4	1074.7
- OSLO3	1	582.15	5425.9	1110.7
- BMI	1	665.44	5509.2	1117.3

Step: AIC=1061.57

DEPSCORE ~ AGE + SEX + GRADE + RESIDENCE + HEALTH + OSLO3 + BMI

	Df	Sum of Sq	RSS	AIC
- AGE	1	9.14	4852.9	1060.4
<none>			4843.7	1061.6
- GRADE	1	26.90	4870.6	1062.0
- RESIDENCE	1	70.91	4914.6	1065.9
- SEX	1	76.63	4920.4	1066.4
- HEALTH	1	148.69	4992.4	1072.7
- OSLO3	1	582.77	5426.5	1108.8
- BMI	1	665.46	5509.2	1115.3

Step: AIC=1060.38

DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 + BMI

	Df	Sum of Sq	RSS	AIC
<none>			4852.9	1060.4
- RESIDENCE	1	74.15	4927.0	1065.0
- SEX	1	74.24	4927.1	1065.0
- GRADE	1	137.80	4990.7	1070.5
- HEALTH	1	149.44	5002.3	1071.5
- OSLO3	1	597.41	5450.3	1108.7
- BMI	1	656.67	5509.5	1113.3

```
> summary(model_backward)
```

Call:

```
lm(formula = DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 +
    BMI, data = modeldata)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.1160	-2.7404	0.2698	2.3102	11.2970

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	23.28421	1.51628	15.356	< 2e-16 ***
SEX	-0.86288	0.33801	-2.553	0.011034 *
GRADE	0.51588	0.14833	3.478	0.000557 ***

```

RESIDENCE    1.08313    0.42453    2.551 0.011079 *
HEALTH      -0.64585    0.17832   -3.622 0.000328 ***
OSLO3       -0.46344    0.06400   -7.242 2.09e-12 ***
BMI         -0.52079    0.06859   -7.592 2.00e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 3.375 on 426 degrees of freedom
 Multiple R-squared: 0.3275, Adjusted R-squared: 0.318
 F-statistic: 34.57 on 6 and 426 DF, p-value: < 2.2e-16

```

> summary(model_backward)
Call:
lm(formula = DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 +
    BMI, data = modeldata)
Residuals:
    Min      1Q      Median      3Q      Max 
-8.1160 -2.7404   0.2698   2.3102  11.2970 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.28421   1.51628  15.356 < 2e-16 ***
SEX          -0.86288   0.33801  -2.553 0.011034 *  
GRADE        0.51588   0.14833   3.478 0.000557 *** 
RESIDENCE    1.08313   0.42453   2.551 0.011079 *  
HEALTH       -0.64585   0.17832  -3.622 0.000328 *** 
OSLO3        -0.46344   0.06400  -7.242 2.09e-12 *** 
BMI          -0.52079   0.06859  -7.592 2.00e-13 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 3.375 on 426 degrees of freedom
 Multiple R-squared: 0.3275, Adjusted R-squared: 0.318
 F-statistic: 34.57 on 6 and 426 DF, p-value: < 2.2e-16

```

> #Summary our best_model4, and compare the multiple R-squared with
best_model6, best_model 6 have higher R2 than 0.3275
> summary(best_model4)

```

```

Call:
lm(formula = DEPSCORE ~ GRADE + HEALTH + OSLO3 + BMI, data = modeldata)

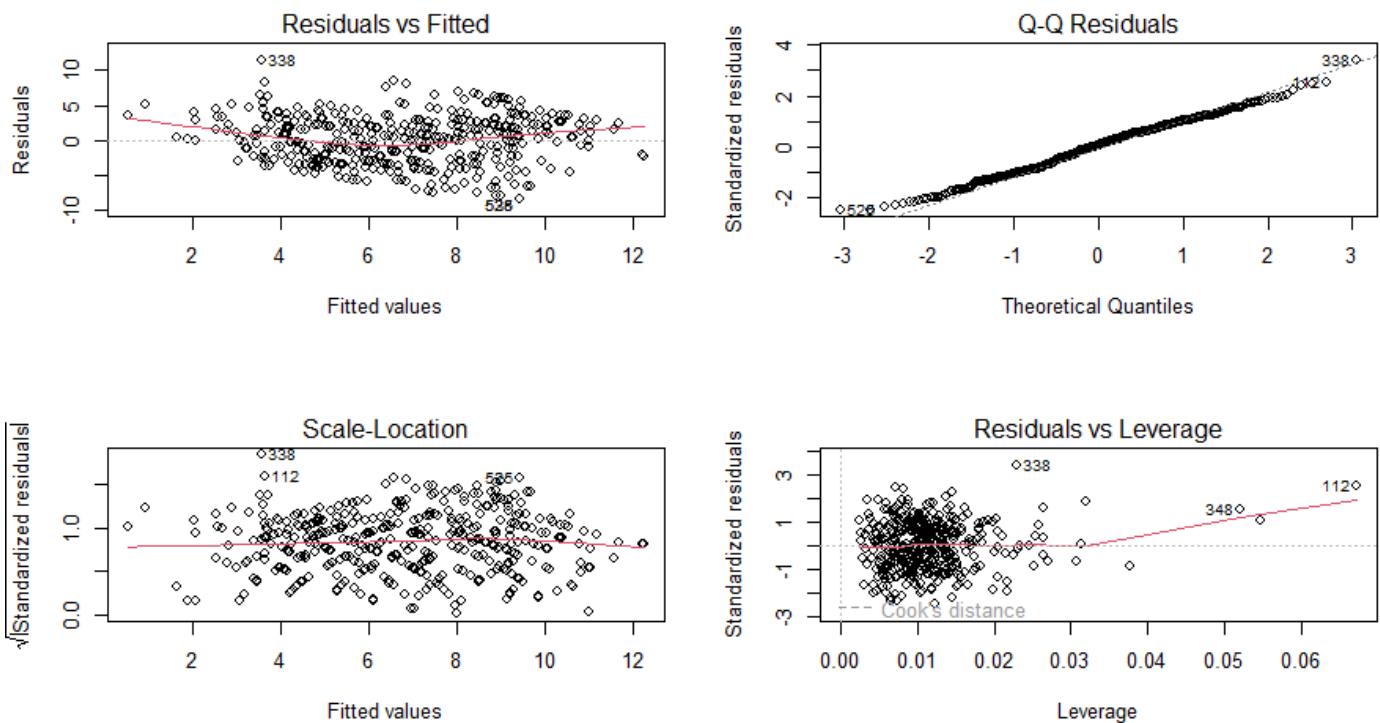
Residuals:
    Min      1Q      Median      3Q      Max 
-8.4345 -2.7068   0.1876   2.3642  11.4318 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 23.40704   1.50792  15.523 < 2e-16 ***
GRADE        0.43313   0.14794   2.928 0.003597 ** 
HEALTH       -0.68958   0.17954  -3.841 0.000141 *** 
OSLO3        -0.46482   0.06433  -7.226 2.30e-12 *** 
BMI          -0.51550   0.06917  -7.452 5.13e-13 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

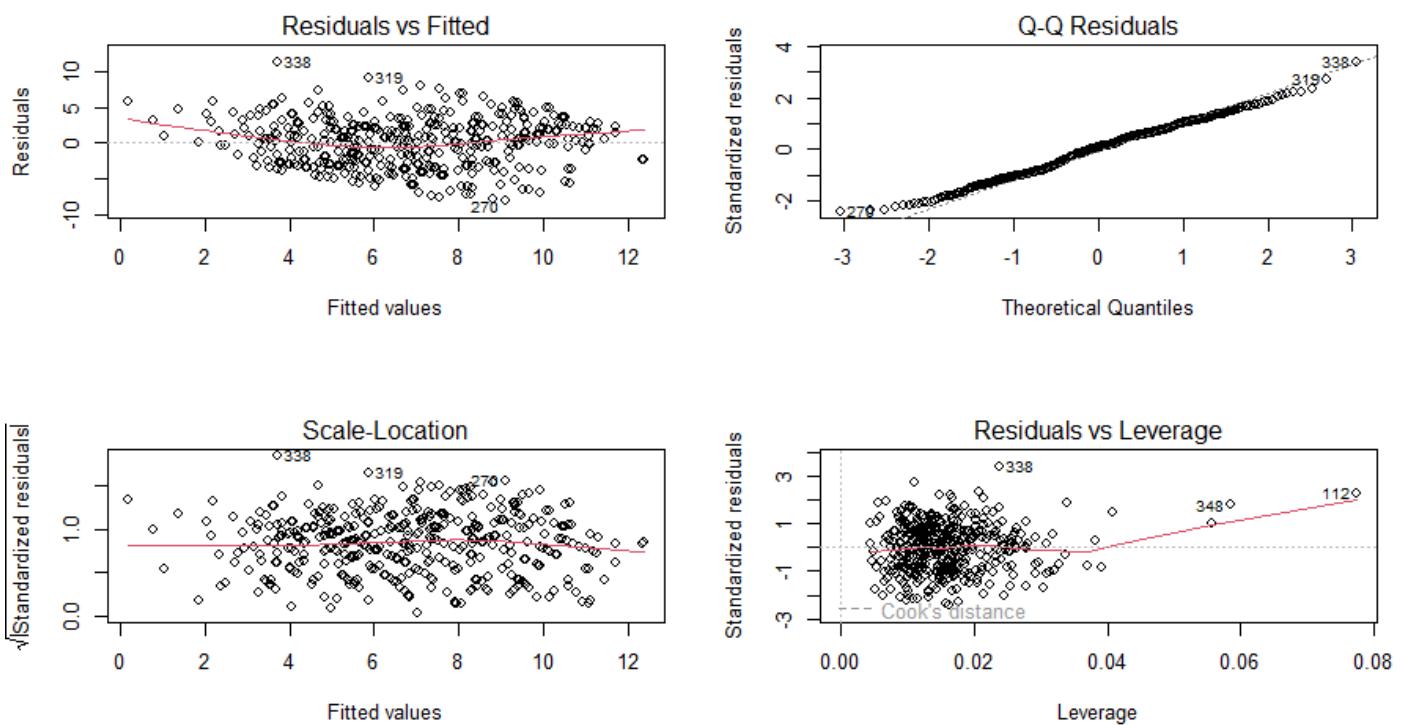
```

Residual standard error: 3.413 on 428 degrees of freedom
 Multiple R-squared: 0.3091, Adjusted R-squared: 0.3027
 F-statistic: 47.88 on 4 and 428 DF, p-value: < 2.2e-16

```
> par(mfrow=c(2,2))
> plot(best_model4)
```



```
> plot(best_model6)
```



```
> #compare the plots between two models, not too much difference.
```

#Appendix A.3.8: R Code for Testing and Comparing Advanced Linear Regression Models

```
> #decide to continue with best_model6 as it has a higher R2. but the QQ plot
```

of best_model6 didn't not shows Normal distribution , so tried different ways to improve it.

#test model with polynomial terms.

```
> best_model61<-lm(DEPSCORE~SEX+GRADE+RESIDENCE+HEALTH+OSLO3+I(BMI^2)+BMI,data = modeldata)
> summary(best_model61)
```

Call:

```
lm(formula = DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 +
    I(BMI^2) + BMI, data = modeldata)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.8347	-2.4971	0.2353	2.2926	9.9799

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	58.92342	6.98225	8.439	5.08e-16 ***
SEX	-0.81162	0.32820	-2.473	0.013791 *
GRADE	0.44149	0.14466	3.052	0.002416 **
RESIDENCE	0.85926	0.41424	2.074	0.038653 *
HEALTH	-0.57698	0.17356	-3.324	0.000963 ***
OSLO3	-0.42693	0.06250	-6.831	2.94e-11 ***
I(BMI^2)	0.08113	0.01554	5.222	2.78e-07 ***
BMI	-3.97151	0.66421	-5.979	4.75e-09 ***

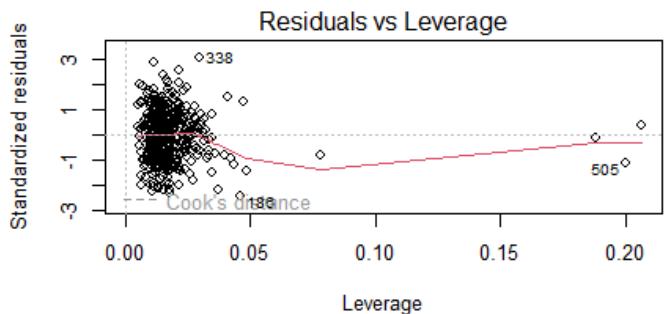
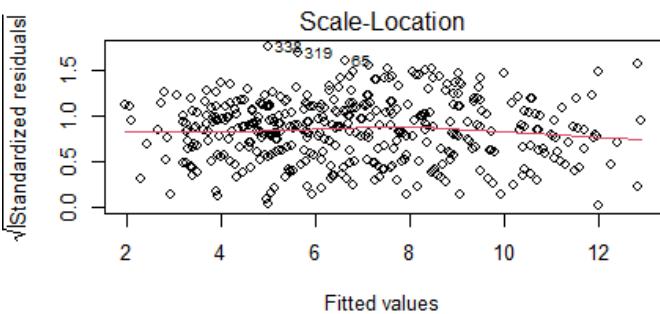
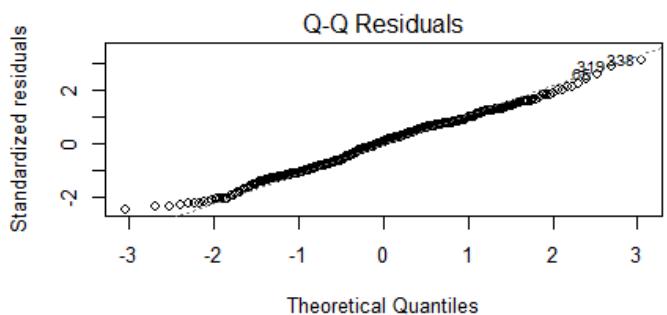
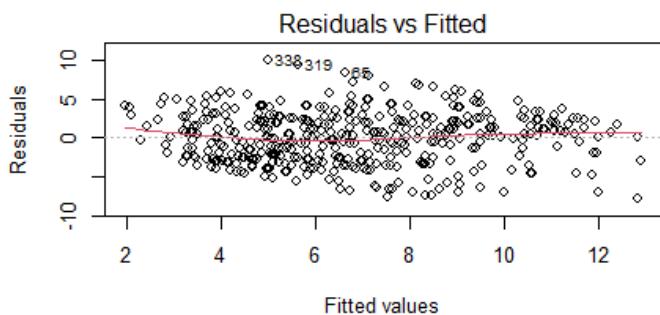
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.276 on 425 degrees of freedom

Multiple R-squared: 0.368, Adjusted R-squared: 0.3576

F-statistic: 35.36 on 7 and 425 DF, p-value: < 2.2e-16

```
> plot(best_model61)
```



#test model with the interaction term.

```
>best_model62<-lm(DEPSCORE~SEX*BMI+GRADE*BMI+RESIDENCE*BMI+HEALTH*BMI+OSLO3*BMI+OSLO3*BMI+BMI,data = modeldata)
> summary(best_model62)
```

```
Call:  
lm(formula = DEPSCORE ~ SEX * BMI + GRADE * BMI + RESIDENCE *  
    BMI + HEALTH * BMI + OSLO3 * BMI + OSLO3 * BMI + BMI, data = modeldata)
```

Residuals:

Min	1Q	Median	3Q	Max
-7.9608	-2.5644	0.3854	2.2088	10.7397

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	36.16243	7.89870	4.578	6.18e-06	***
SEX	1.18218	2.78243	0.425	0.67115	
BMI	-1.16822	0.39541	-2.954	0.00331	**
GRADE	0.98777	1.33658	0.739	0.46030	
RESIDENCE	-0.99184	3.37129	-0.294	0.76875	
HEALTH	1.24170	1.51963	0.817	0.41433	
OSLO3	-2.63896	0.56840	-4.643	4.60e-06	***
SEX:BMI	-0.09864	0.13889	-0.710	0.47796	
BMI:GRADE	-0.02607	0.06697	-0.389	0.69730	
BMI:RESIDENCE	0.10521	0.16936	0.621	0.53480	
BMI:HEALTH	-0.09646	0.07585	-1.272	0.20415	
BMI:OSLO3	0.10965	0.02852	3.844	0.00014	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.329 on 421 degrees of freedom
Multiple R-squared: 0.3536, Adjusted R-squared: 0.3367
F-statistic: 20.94 on 11 and 421 DF, p-value: < 2.2e-16

```
> #Drop variable according to result.  
>  
best_model62<-lm(DEPSCORE~SEX+GRADE+RESIDENCE+HEALTH+OSLO3+OSLO3*BMI+BMI,data  
= modeldata)  
> summary(best_model62)
```

Call:

```
lm(formula = DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 +  
    OSLO3 * BMI + BMI, data = modeldata)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.0881	-2.5810	0.3295	2.2023	12.2463

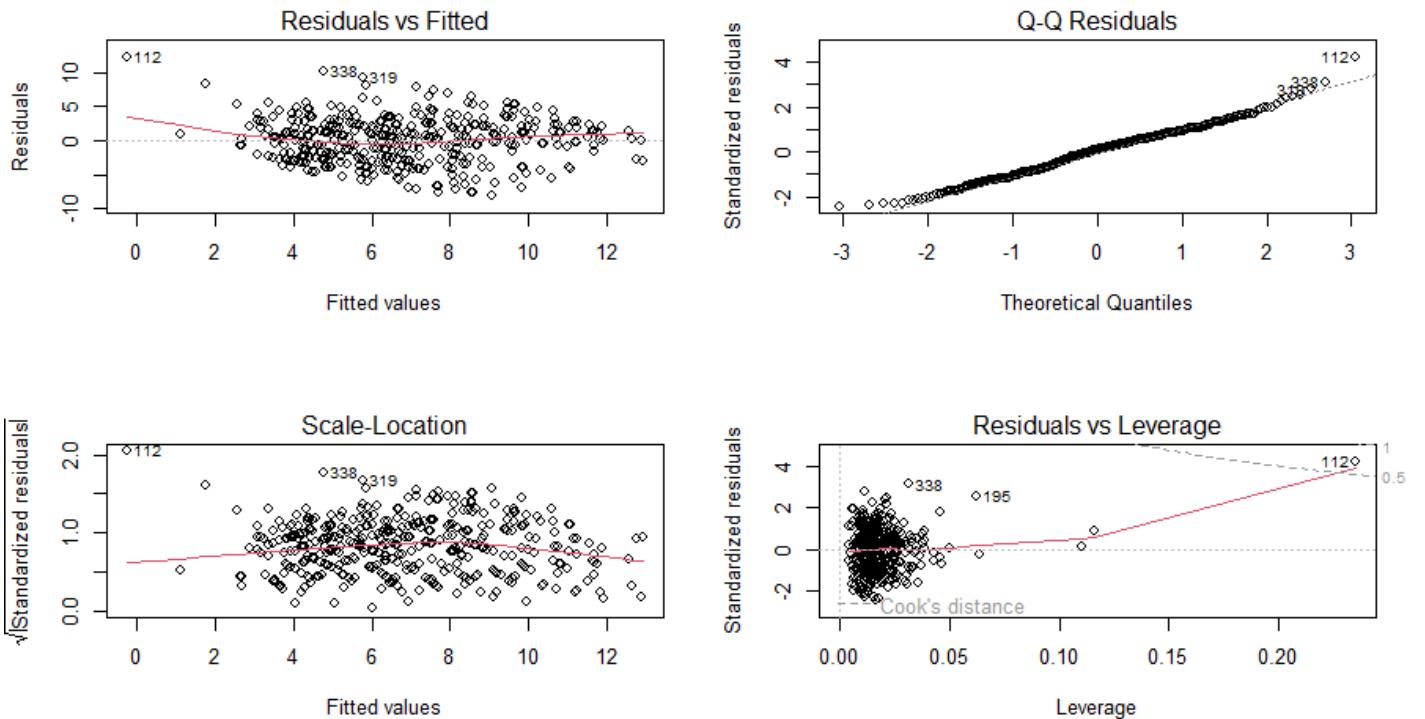
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	41.89223	5.20780	8.044	8.74e-15	***
SEX	-0.78998	0.33358	-2.368	0.018322	*
GRADE	0.46528	0.14676	3.170	0.001632	**
RESIDENCE	1.15236	0.41865	2.753	0.006166	**
HEALTH	-0.68298	0.17596	-3.882	0.000120	***
OSLO3	-2.34017	0.50710	-4.615	5.22e-06	***
BMI	-1.45079	0.25834	-5.616	3.54e-08	***
OSLO3:BMI	0.09422	0.02526	3.730	0.000218	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.325 on 425 degrees of freedom
Multiple R-squared: 0.3488, Adjusted R-squared: 0.3381
F-statistic: 32.52 on 7 and 425 DF, p-value: < 2.2e-16

```
> plot(best_model62)
```



```

> #test model with both interaction and polynomial term.
>
best_model612<-lm(DEPSCORE~SEX+GRADE+RESIDENCE+HEALTH+OSLO3+BMI+I(BMI^2)+OSLO3
*BMI,data = modeldata)
> summary(best_model612)

Call:
lm(formula = DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 +
    BMI + I(BMI^2) + OSLO3 * BMI, data = modeldata)

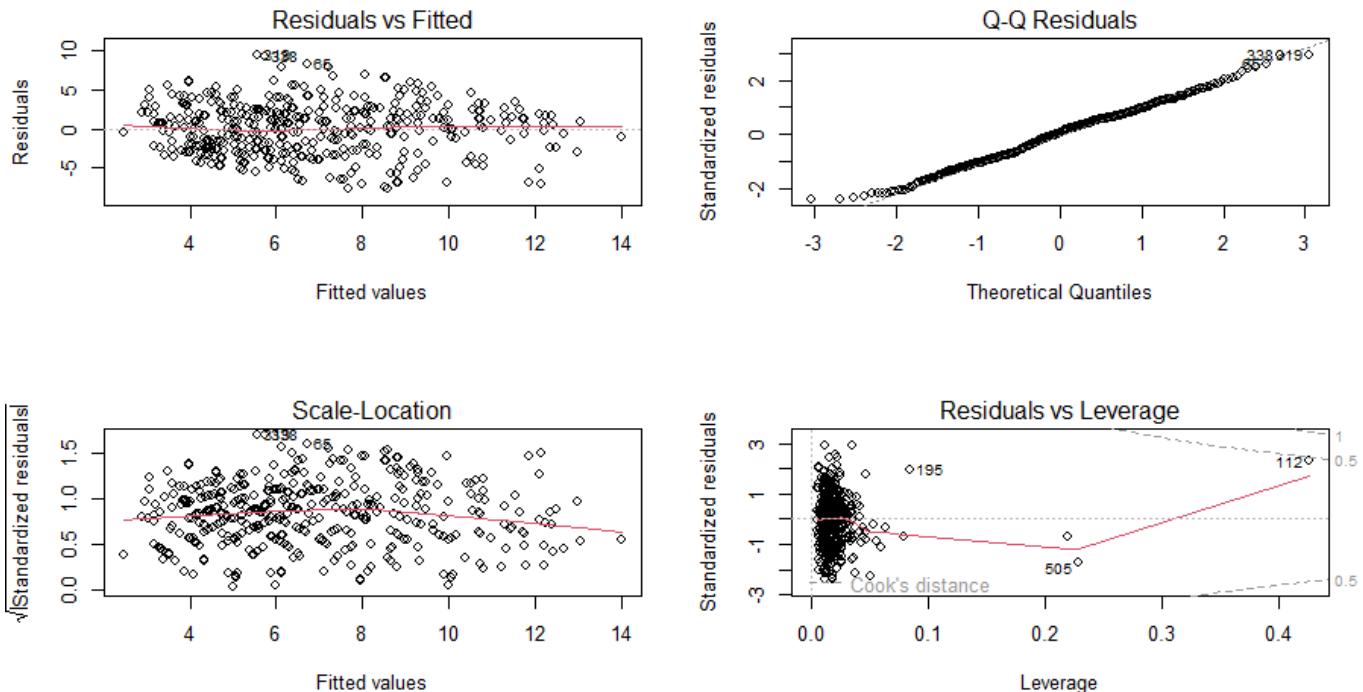
Residuals:
    Min      1Q  Median      3Q     Max 
-7.6600 -2.4761  0.2012  2.1476  9.4197 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 69.78034   7.85614   8.882 < 2e-16 ***
SEX          -0.76013   0.32580  -2.333 0.020110 *  
GRADE         0.40968   0.14381   2.849 0.004601 ** 
RESIDENCE     0.93575   0.41145   2.274 0.023450 *  
HEALTH        -0.61282   0.17248  -3.553 0.000424 *** 
OSLO3         -1.88953   0.50456  -3.745 0.000205 *** 
BMI           -4.34509   0.67071  -6.478 2.57e-10 *** 
I(BMI^2)       0.07291   0.01566   4.657 4.29e-06 *** 
OSLO3:BMI     0.07325   0.02508   2.921 0.003677 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.247 on 424 degrees of freedom
Multiple R-squared:  0.3805,    Adjusted R-squared:  0.3688 
F-statistic: 32.55 on 8 and 424 DF,  p-value: < 2.2e-16

```

```
> plot(best_model612)
```



>#Appendix A.3.8.1: R code for the models that were dropped.

```
>#test model with dommy variable "residence".
>best_model63<-lm(DEPSCORE~RESIDENCE*GRADE+SEX*RESIDENCE+RESIDENCE*HEALTH+RESIDENCE*OSLO3+RESIDENCE*BMI+RESIDENCE,data=modedata)
> summary(best_model63)
```

Call:

```
lm(formula = DEPSCORE ~ RESIDENCE * GRADE + SEX * RESIDENCE +
    RESIDENCE * HEALTH + RESIDENCE * OSLO3 + RESIDENCE * BMI +
    RESIDENCE, data = modedata)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.1351	-2.7036	0.2834	2.4519	11.0870

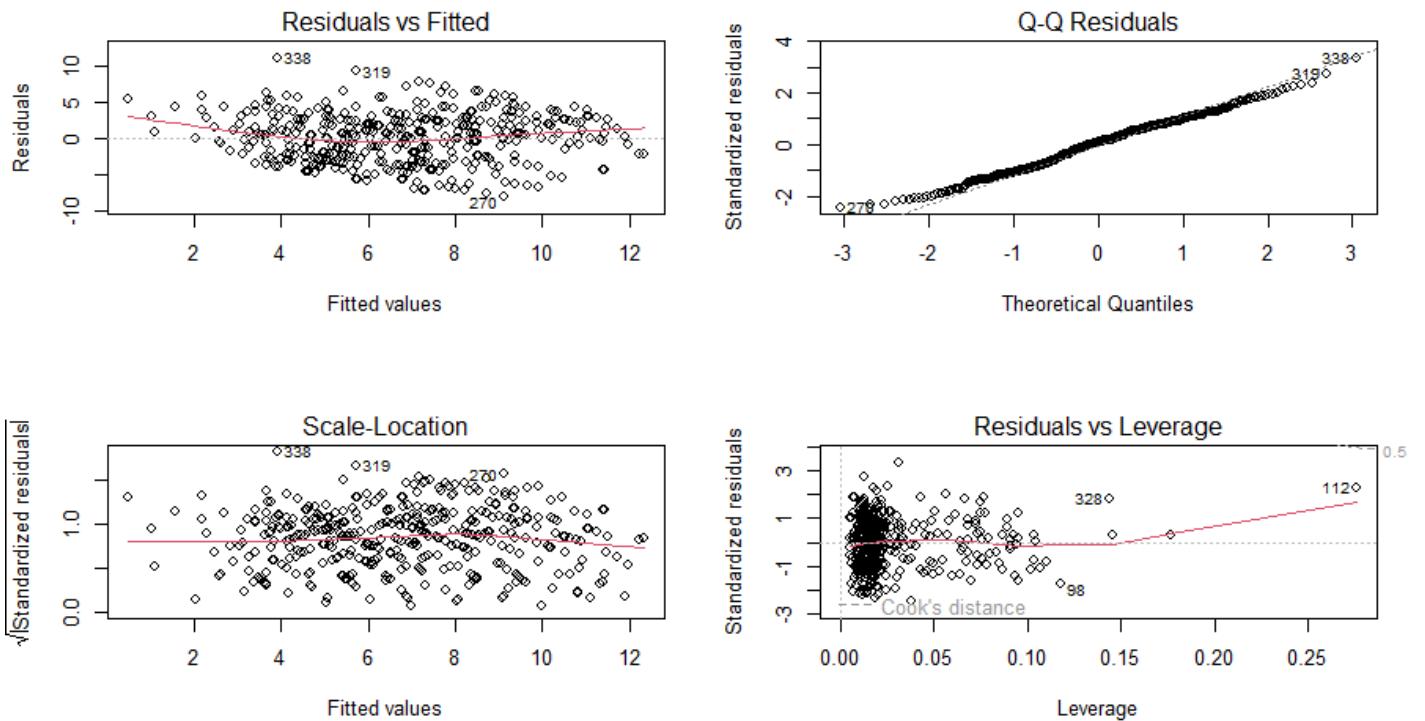
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.81289	1.68604	13.530	< 2e-16 ***
RESIDENCE	4.71518	4.03092	1.170	0.242762
GRADE	0.52242	0.16168	3.231	0.001329 **
SEX	-0.86182	0.37693	-2.286	0.022727 *
HEALTH	-0.70624	0.19741	-3.578	0.000387 ***
OSLO3	-0.41420	0.07048	-5.877	8.5e-09 ***
BMI	-0.51116	0.07952	-6.428	3.5e-10 ***
RESIDENCE:GRADE	-0.21330	0.42431	-0.503	0.615436
RESIDENCE:SEX	-0.22733	0.87534	-0.260	0.795218
RESIDENCE:HEALTH	0.33994	0.48118	0.706	0.480280
RESIDENCE:OSLO3	-0.33152	0.17665	-1.877	0.061252 .
RESIDENCE:BMI	-0.06216	0.16190	-0.384	0.701232

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.38 on 421 degrees of freedom
 Multiple R-squared: 0.3337, Adjusted R-squared: 0.3162
 F-statistic: 19.16 on 11 and 421 DF, p-value: < 2.2e-16

```
> plot(best_model63)
```



```
#test model with dummy variable "Sex".
>best_model64<-lm(DEPSCORE~SEX*GRADE+SEX*RESIDENCE+SEX*HEALTH+SEX*OSLO3+SEX*BMI+SEX, data=m
odeldata)
> summary(best_model64)
```

Call:
`lm(formula = DEPSCORE ~ SEX * GRADE + SEX * RESIDENCE + SEX *
 HEALTH + SEX * OSLO3 + SEX * BMI, data = modeldata)`

Residuals:

Min	1Q	Median	3Q	Max
-8.0182	-2.6804	0.2788	2.3556	11.2399

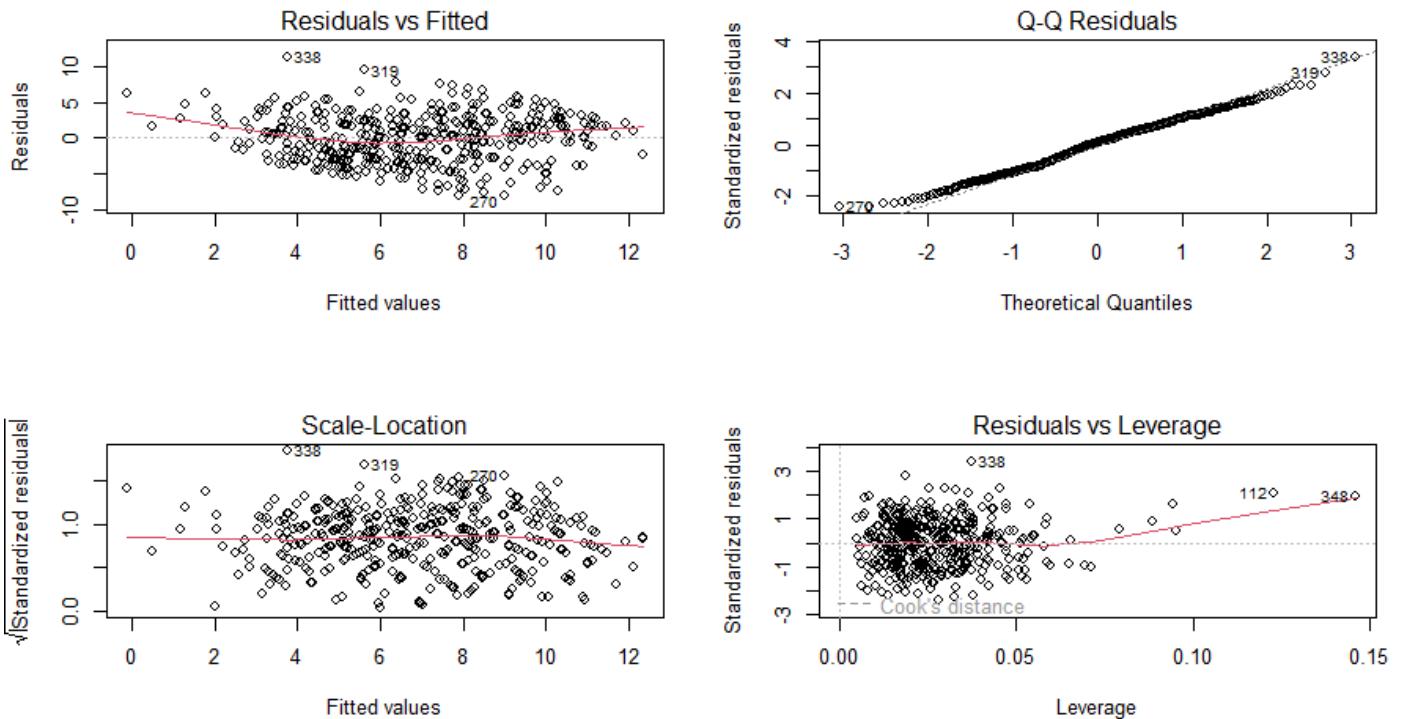
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.53564	1.87287	12.033	< 2e-16 ***
SEX	1.10598	3.21105	0.344	0.73070
GRADE	0.56772	0.19401	2.926	0.00362 **
RESIDENCE	1.03417	0.59258	1.745	0.08168 .
HEALTH	-0.49479	0.22869	-2.164	0.03106 *
OSLO3	-0.56023	0.08434	-6.642	9.57e-11 ***
BMI	-0.46888	0.08637	-5.429	9.62e-08 ***
SEX:GRADE	-0.11927	0.30184	-0.395	0.69292
SEX:RESIDENCE	0.10859	0.85763	0.127	0.89930
SEX:HEALTH	-0.33362	0.36971	-0.902	0.36737
SEX:OSLO3	0.22214	0.13068	1.700	0.08991 .
SEX:BMI	-0.13154	0.14352	-0.917	0.35991

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.38 on 421 degrees of freedom
 Multiple R-squared: 0.3336, Adjusted R-squared: 0.3161
 F-statistic: 19.16 on 11 and 421 DF, p-value: < 2.2e-16

```
> plot(best_model64)
```



```
#test model with log transformation.
>best_model65<-lm(log(DEPSCORE+1)~SEX+log(GRADE)+RESIDENCE+log(HEALTH)+log(OSLO3)+log(BMI), data = modedata)
> summary(best_model65)

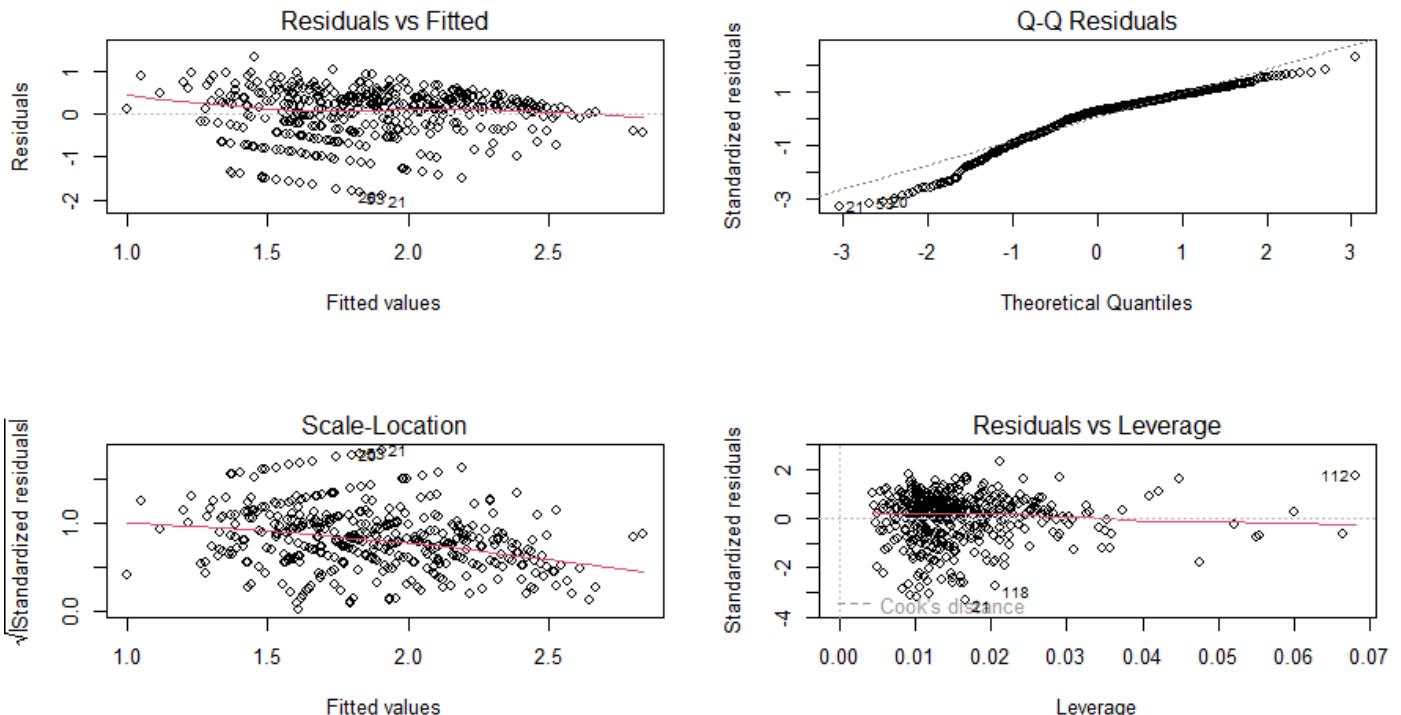
Call:
lm(formula = log(DEPSCORE + 1) ~ SEX + log(GRADE) + RESIDENCE +
    log(HEALTH) + log(OSLO3) + log(BMI), data = modedata)

Residuals:
    Min      1Q  Median      3Q      Max 
-1.9049 -0.3158  0.1584  0.3766  1.3176 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 8.29264   0.71975 11.522 < 2e-16 ***
SEX         -0.11556   0.05790 -1.996 0.046594 *  
log(GRADE)  0.19992   0.05345  3.741 0.000209 *** 
RESIDENCE   0.18595   0.07260  2.561 0.010774 *  
log(HEALTH) -0.31332   0.10000 -3.133 0.001848 ** 
log(OSLO3)  -0.50884   0.09997 -5.090 5.39e-07 *** 
log(BMI)     -1.66895   0.24359 -6.851 2.57e-11 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5784 on 426 degrees of freedom
Multiple R-squared:  0.2577, Adjusted R-squared:  0.2472 
F-statistic: 24.64 on 6 and 426 DF,  p-value: < 2.2e-16

> plot(best_model65)
```



```
> # test model with a single log term.
> best_model66<-lm(DEPSCORE~SEX+GRADE+RESIDENCE+HEALTH+OSLO3+log(BMI), data =
modeldata)
> summary(best_model66)
```

Call:

```
lm(formula = DEPSCORE ~ SEX + GRADE + RESIDENCE + HEALTH + OSLO3 +
log(BMI), data = modeldata)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.0069	-2.6873	0.3087	2.2602	11.2416

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	46.9840	4.1774	11.247	< 2e-16 ***
SEX	-0.8621	0.3352	-2.572	0.010452 *
GRADE	0.5003	0.1472	3.398	0.000743 ***
RESIDENCE	1.0400	0.4213	2.469	0.013947 *
HEALTH	-0.6304	0.1770	-3.562	0.000409 ***
OSLO3	-0.4532	0.0636	-7.126	4.44e-12 ***
log(BMI)	-11.4550	1.4145	-8.098	5.92e-15 ***

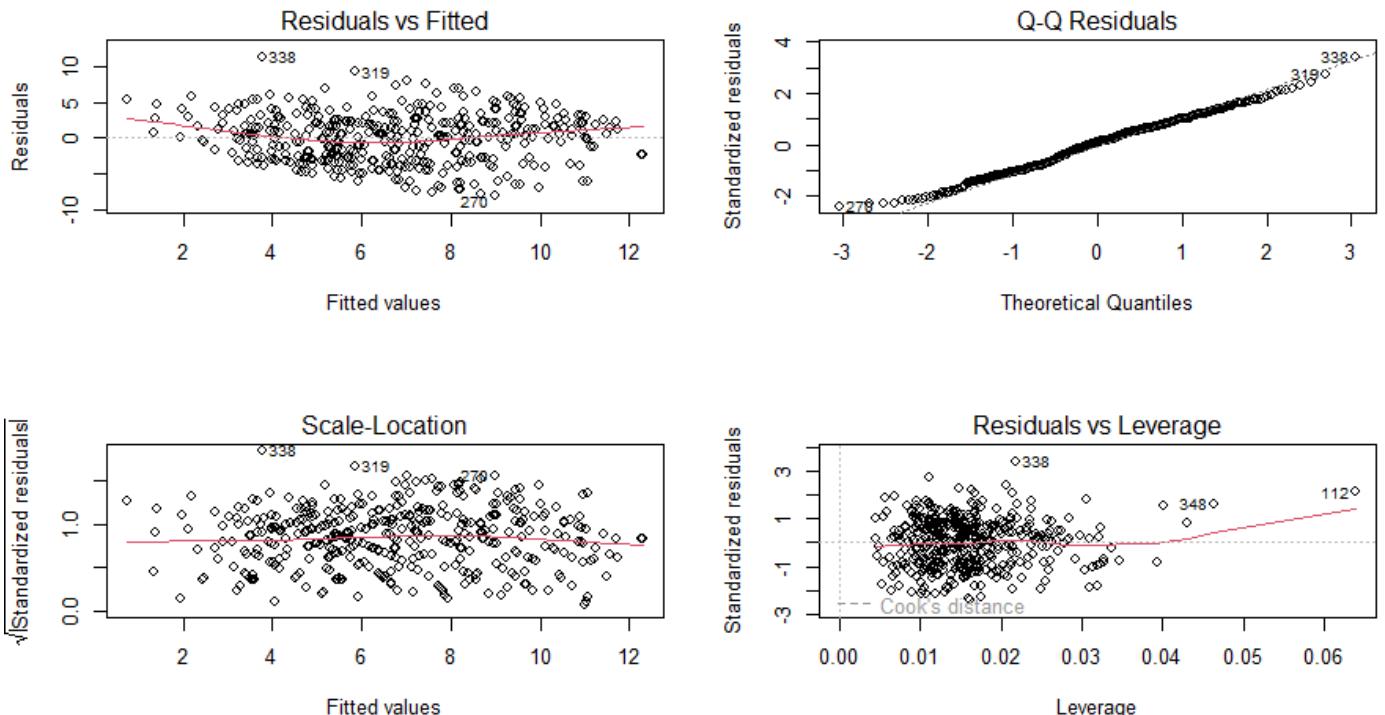
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.348 on 426 degrees of freedom

Multiple R-squared: 0.3383, Adjusted R-squared: 0.329

F-statistic: 36.31 on 6 and 426 DF, p-value: < 2.2e-16

```
> plot(best_model66)
```



```
> #compare all the summary information and plot of models above, best_model61, best_model62 and best_model612 give better R2 and plot. So we decided to continue with them.
```

#Appendix A.3.9: R Code for Evaluating Models Using RMSE on Test Data

```
> #best_model612 gives the smallest RMSE 0.3054343, so continue with model612.
>
> predictions <- predict(best_model6, newdata = modeldata)
> rmse <- sqrt(mean((predictions - testdata$DEPSCORE)^2))
> rmse
[1] 3.211518
> predictions <- predict(best_model61, newdata = modeldata)
> rmse <- sqrt(mean((predictions - testdata$DEPSCORE)^2))
> rmse
[1] 3.074088

> predictions <- predict(best_model62, newdata = modeldata)
> rmse <- sqrt(mean((predictions - testdata$DEPSCORE)^2))
> rmse
[1] 3.105501
> predictions <- predict(best_model612, newdata = testdata)
> rmse <- sqrt(mean((predictions - testdata$DEPSCORE)^2))
> rmse
[1] 3.054343
```

#Appendix A.4: R Code for Residual Analysis, Correlation, and Visualization of Prediction

```
> library(dplyr)
> library(caret)
> library(rpart)
> #Using best_model612 to predict PHQ-9 Depression Scores based on the test data, and add it to the test data table.
> testdata$fitted_values <- predict(best_model612,newdata = testdata)
>
```

```

> #get rmse for model612 based on test data.
> predictions <- predict(best_model612, newdata = testdata)
> rmse <- sqrt(mean((predictions - testdata$DEPSCORE)^2))
> rmse
[1] 3.054343
>
>
> #Add columns of residuals of predicted value and real data to test data.
> testdata<-testdata%>%mutate(residuals=DEPSCORE-fitted_values)
>
> #summary the data and get correlation of it.
> testdata%>%summarise(cor=cor(DEPSCORE,fitted_values))%>%
+   mutate(R2=cor^2)
      cor          R2
1 0.6133986 0.3762578
> #plot the scatter plot of observed PHQ-9 Depression Scores against Predict
PHQ-9 Depression Scores
> ggplot(testdata)+geom_point(aes(x=fitted_values,y=DEPSCORE))+labs(x="Predict
PHQ-9 Depression Scores ",y ="Observed PHQ-9 Depression Scores",title =
"scatter plot of observed PHQ-9 Depression Scores against Predict PHQ-9
Depression Scores")
scatterplot of observed PHQ-9 Depression Scores against Predict PHQ-9 Depression Scores

The scatter plot displays the relationship between observed and predicted PHQ-9 depression scores. The x-axis represents the predicted score, and the y-axis represents the observed score. Both axes range from 0 to 15. The data points generally follow a positive linear trend, indicating that predicted scores tend to be similar to observed scores. There are some notable outliers at higher values, such as a point at (8, 15) and others near (11, 14) and (15, 13).


>
> ggplot(testdata)+aes(x=GRADE,y=fitted_values,color=SEX)+geom_point()+facet_wra
p(~RESIDENCE)+theme_bw()

```

