525 Final Project

Daniel Cahill, Maxwell Owens, Cesar McElaney

May 2024

1 Introduction

1.0.1 Description of Data

The overall name of our group was ”Regression Analysis on Cancer”, as one can assume we worked with the cancer.csv dataset provided in class. This dataset includes measurements from 167 breast biopsies, each diagnosed as either ductal carcinoma in situ (DCIS) or benign usual ductal hyperplasia (UDH). The data was collected by segmenting biopsies into individual nuclei and recording each nucleus's area and color, with these measurements then averaged or subjected to standard deviation calculations at the biopsy level.

1.0.2 Research Questions

Overall we had two different research questions. The first was the goal of the project. The goal of this project is to explore the relationship between average nuclear area, which is known to be correlated with malignancy and the intensity of different colors in the biopsied tissue. Our second questions was what happens to the mean area when we fit a model with an interaction term between the average lightness across nuclei on the HSV scale and grade?

1.0.3 The Variables

When exploring the data set we had found that there were sensitive variables and obvious redundancies. For the sensitive variables we found that those included ones such as ClassGrade, Mean_mean_R, Mean_mean_G, Mean_mean_B, Mean_mean_HSV, Mean_mean_HE, Mean_mean_BR, Mean_mean_Lab, and the standard deviations of each variable as well. For example, SD_mean_R, SD_mean_HSV and so on. Later on we modified the ClassGrade variable to the Grade variable which is explained below. For the obvious redundancies we only found one variable that we thought was one: Mean_Area.

To create our first model, we first needed to understand what each variable was. Our response variable Mean_Area is the average nuclear area across nuclei. For our predictor variables, Mean_mean_R, Mean_mean_G, Mean_mean_B were the average red, green, and blue intensity across nuclei on the RGB scale respectively. For Mean_mean_HSV, Mean_mean_HE, Mean_mean_BR, these are the average lightness/value across nuclei on the HSV scale, the average lightness across nuclei from H&E color deconvolution, and average lightness across nuclei on BR grey scale respectively. Lastly, for Mean_mean_Lab and Grade, these are the average lightness across nuclei on Lab and Luv scale and the ClassGrade variable separated into either UDH or nUDH.

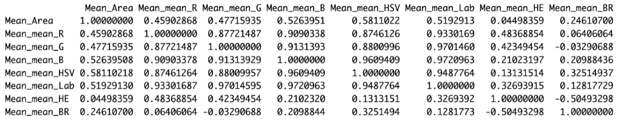
2 Methods and Results

2.1 Initial MLR Model

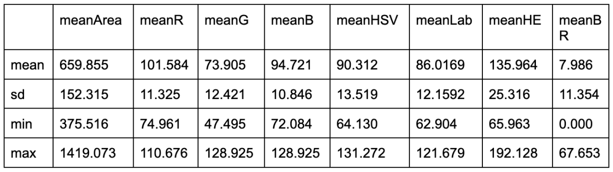
2.1.1 A clear description of the initial MLR model.

In our initial model, we have the response variable Mean_Area. We also have eight predictor variables: Grade, Mean_mean_R, Mean_mean_G, Mean_mean_B, Mean_mean_HSV, Mean_mean_HE, Mean_mean_BR, Mean_mean_Lab, and an interaction term between Mean_mean_HSV and Grade.

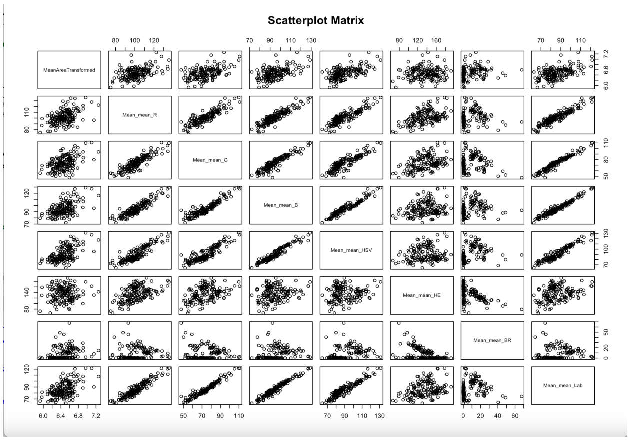
2.1.2 Summary, Correlation Matrix, and Scatter Plot Matrix



```
              Mean_Area Mean_mean_R Mean_mean_G Mean_mean_B Mean_mean_HSV Mean_mean_Lab Mean_mean_HE Mean_mean_BR
Mean_Area     1.00000000  0.45902868  0.47715935   0.5263951     0.5811022     0.5192913   0.04498359   0.24610700
Mean_mean_R   0.45902868  1.00000000  0.87721487   0.9090338     0.8746126     0.9330169   0.48368854   0.06406064
Mean_mean_G   0.47715935  0.87721487  1.00000000   0.9131393     0.8800996     0.9701460   0.42349454  -0.03290688
Mean_mean_B   0.52639508  0.90903378  0.91313929   1.0000000     0.9609409     0.9720963   0.21023197   0.20988436
Mean_mean_HSV 0.58110218  0.87461264  0.88009957   0.9609409     1.0000000     0.9487764   0.13131514   0.32514937
Mean_mean_Lab 0.51929130  0.93301687  0.97014595   0.9720963     0.9487764     1.0000000   0.32693915   0.12817729
Mean_mean_HE  0.04498359  0.48368854  0.42349454   0.2102320     0.1313151     0.3269392   1.00000000  -0.50493298
Mean_mean_BR  0.24610700  0.06406064 -0.03290688   0.2098844     0.3251494     0.1281773  -0.50493298   1.00000000
```



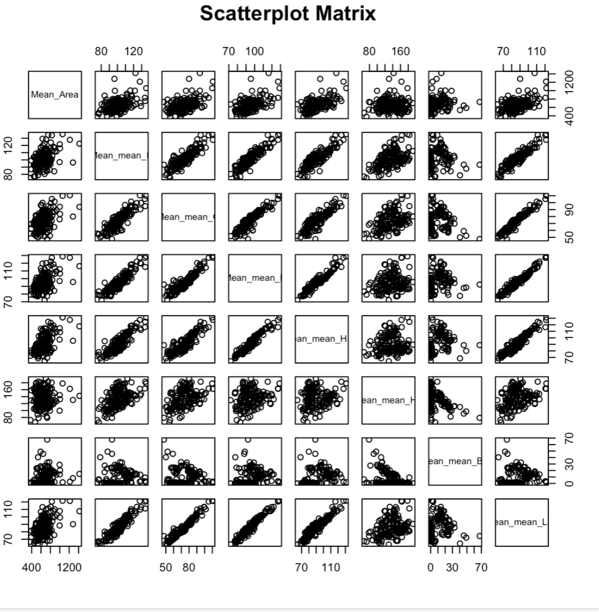
	meanArea	meanR	meanG	meanB	meanHSV	meanLab	meanHE	meanBR
mean	659.855	101.584	73.905	94.721	90.312	86.0169	135.964	7.986
sd	152.315	11.325	12.421	10.846	13.519	12.1592	25.316	11.354
min	375.516	74.961	47.495	72.084	64.130	62.904	65.963	0.000
max	1419.073	110.676	128.925	128.925	131.272	121.679	192.128	67.653

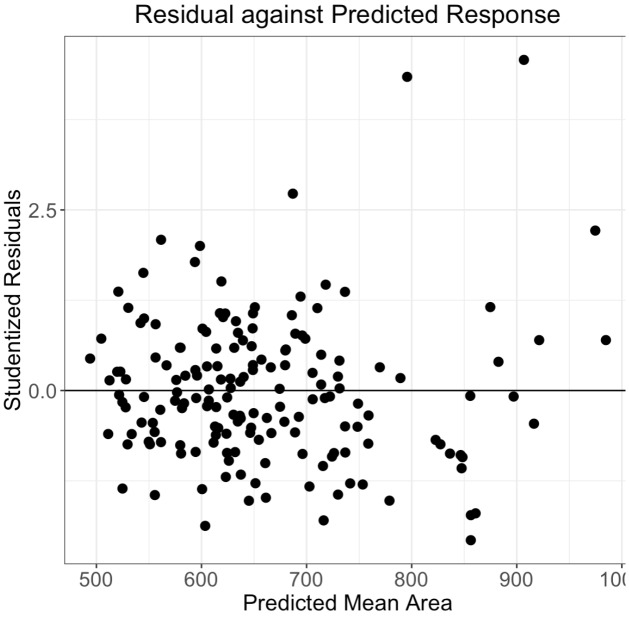


Provided above we can see the correlation matrix, summary of the response and the quantitative predictor variables, and the scatter plot matrix. The positive moderate linear relationship between our response variable Mean_Area and our predictor variables Mean_mean_R, Mean_mean_G, Mean_mean_B, and Mean_mean_HSV are moderate rather than strong. This is consistent with the correlation matrix. We can see in the scatter plots with Mean_Area and Mean_mean_B, we can see that there are potential outliers.

2.2 Checking Assumptions

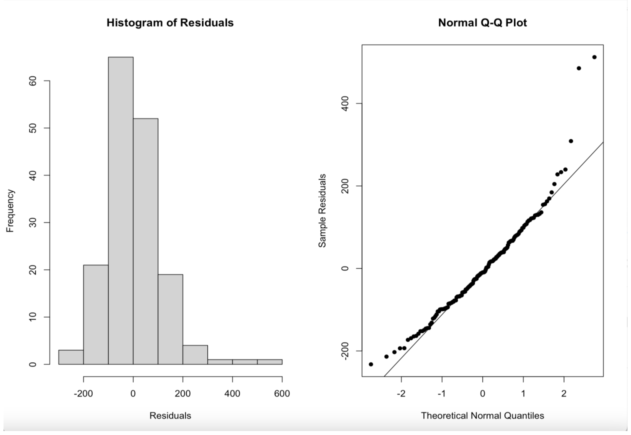
2.2.1 Before Log Transformation





In order to check linearity we look at the scatter plot matrix and residual plot. We see no curve in the scatter plot matrix and the residual plot has points scattered above and below the 0 residual line with no curves, all this together shows that the assumption of linearity is reasonably met.

When checking our constant variance assumption we used a residual plot and Levene Test. In the residual plot you can see that the variability is different for for larger predicted values. When we ran the Levene test, we calculated a p-value of 0.03114. This means that the p-value is less than $\alpha = 0.05$. These observations together mean that the constant variance assumption was violated.



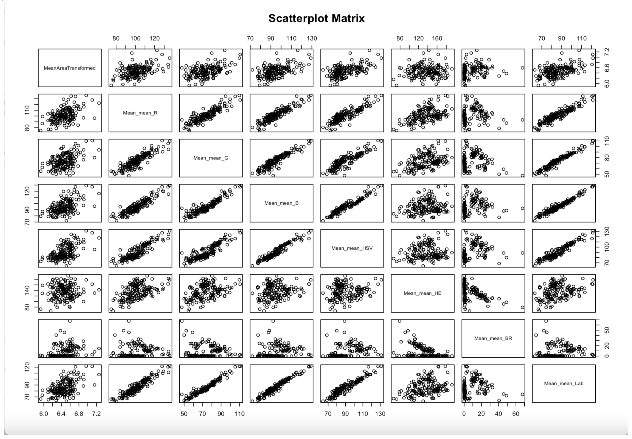
When checking normality, we used a histogram of residuals, a normal Q-Q plot, and a Shapiro-Wilk test. In the Q-Q plot, we can see that the data is close to the straight line in the middle part, but there is departure from the straight line at the tails. When we ran the Shapiro-Wilk test, we obtained a p-value of $2.408 * 10^{-6}$. This is less than $\alpha = 0.05$. Both of these observations together means that the normality assumption is violated.

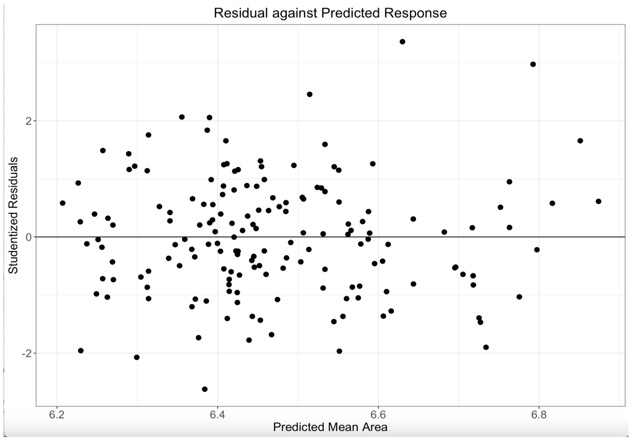
We assume a random sample for this dataset. So the independence assumption is reasonably met.

2.2.2 After Log Transformation

Since we had assumption violations in our initial model, we performed a log transformation on our response variable Mean_Area. In our new model, our response variable is MeanAreaTransformed, which is equivalent to log(Mean_Area). We kept all of the same predictors and our interaction term. Once reinitialized, we rechecked the assumptions.

We assume a random sample for this dataset. So the independence assumption is reasonably met.

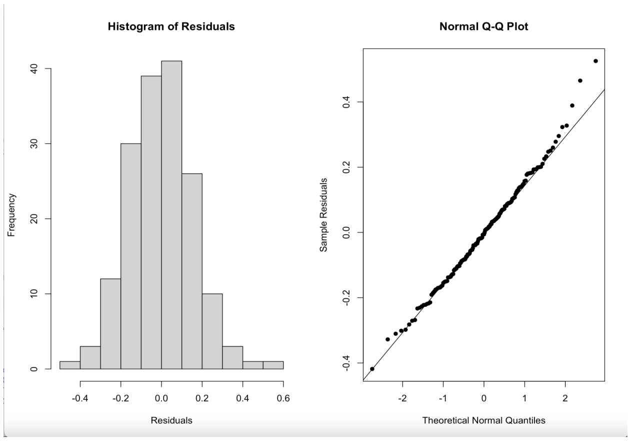




To check the assumptions of linearity, we again use a scatter plot matrix and a residual plot. In the scatter plot matrix, we saw there is no curve in the data. The residual plot has points scattered above and below the 0 residual line with no curves, providing the evidence that linearity is reasonably met.

For the assumption of constant variance, we also use the residual plot. In our initial model, this assumption was violated. Using the log transformation for our response variable, we refit the model and checked this assumption again. Once doing this, we can see that the residual plot shows similar variability in residuals for various predicted mean areas, which means the assumption of constant variance is reasonably met. After running a Levene test, we obtained a p-value of 0.6985. This is larger than $\alpha = 0.05$ so the assumption is reasonably met.

Once again, we assume a random sample for this dataset. So the independence assumption is reasonably met.



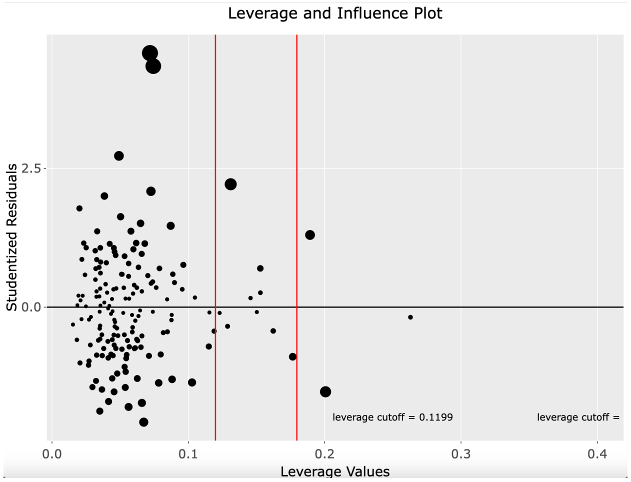
For the normality assumption, to check if it is met or not, we use a histogram of residuals and a normal Q-Q plot. We can see in the histogram that the data is normally distributed around 0.

In the normal Q-Q plot, we can see with the exception of some outliers, the Q-Q plot is close to the line. These observations lead us to believe that the normality assumption is reasonably met. We also ran a Shapiro-Wilk test. We got a calculated p-value of 0.9969. This value is greater than $\alpha = 0.05$ so this means that the normality assumption is reasonably met.

Further, none of our plots indicate any violation of independence, linearity, constant variance, or normality, so all of the MLR model assumptions are reasonably met.

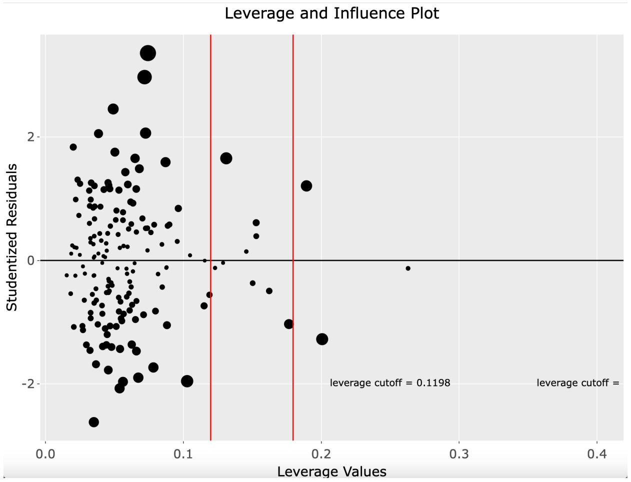
2.3 Leverage and Influence Plot

2.3.1 Before Log Transformation



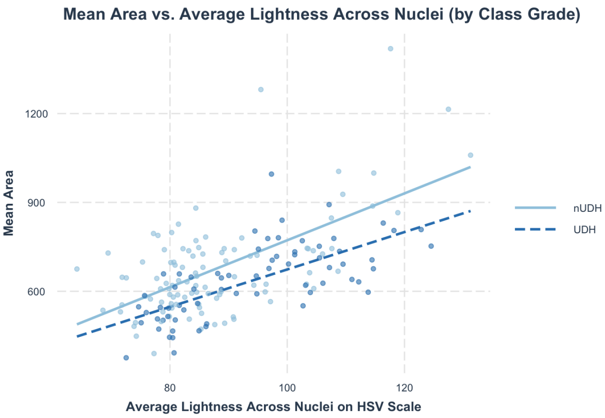
This is our leverage and influence plot for our model we initially set. As we can see, there is some concern for outliers. Due to this, we used our new model with the log transformation of our response variable and recreated the plot.

2.3.2 After Log Transformation



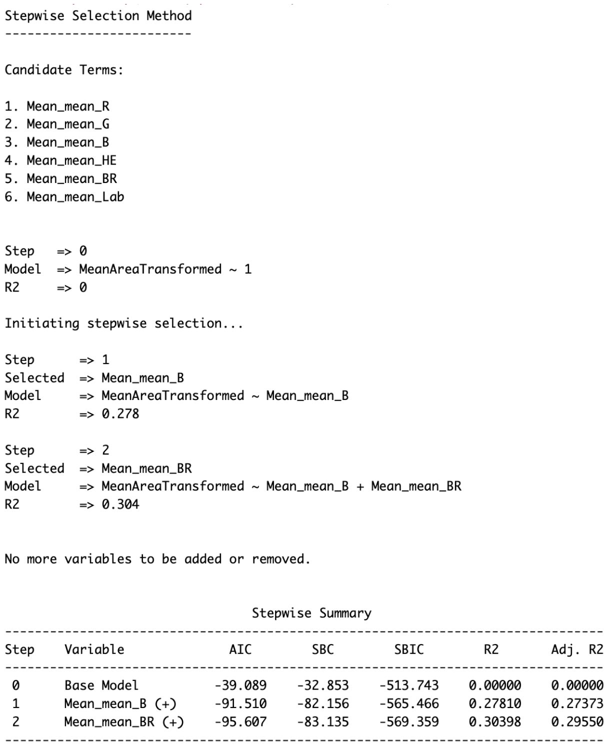
Once the log transformation is performed, we do not have concern for outliers.

2.4 Plot for Checking an Interaction Term

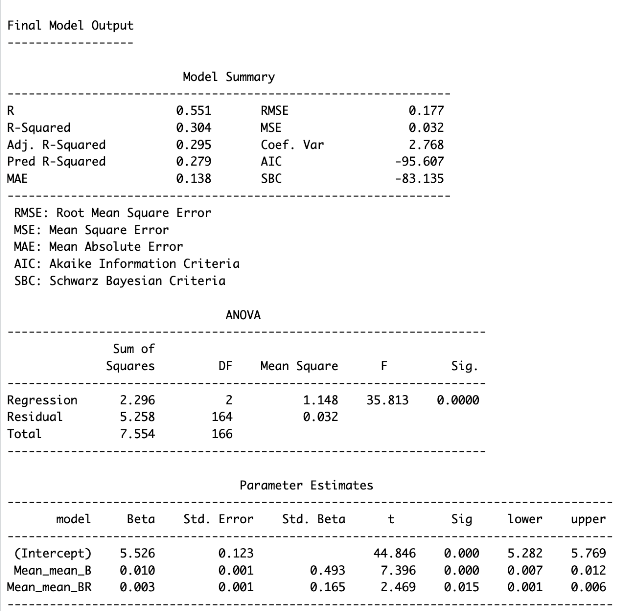


This interaction plot displays the relation shape between the mean lightness across nuclei on the HSV scale and the total mean area, categorized into two categories: UDH and not UDH. We can see by looking at our plot that the two lines are not parallel. This indicates that we can consider the interaction term between Mean_mean_HSV and Grade.

2.5 Sequential Model Selection



```
Stepwise Selection Method
-------------------------

Candidate Terms:

1. Mean_mean_R
2. Mean_mean_G
3. Mean_mean_B
4. Mean_mean_HE
5. Mean_mean_BR
6. Mean_mean_Lab


Step   => 0
Model  => MeanAreaTransformed ~ 1
R2     => 0

Initiating stepwise selection...

Step      => 1
Selected  => Mean_mean_B
Model     => MeanAreaTransformed ~ Mean_mean_B
R2        => 0.278

Step      => 2
Selected  => Mean_mean_BR
Model     => MeanAreaTransformed ~ Mean_mean_B + Mean_mean_BR
R2        => 0.304


No more variables to be added or removed.


                              Stepwise Summary
-----------------------------------------------------------------------------------
Step    Variable             AIC        SBC        SBIC        R2        Adj. R2
-----------------------------------------------------------------------------------
 0      Base Model         -39.089    -32.853    -513.743    0.00000    0.00000
 1      Mean_mean_B (+)    -91.510    -82.156    -565.466    0.27810    0.27373
 2      Mean_mean_BR (+)   -95.607    -83.135    -569.359    0.30398    0.29550
-----------------------------------------------------------------------------------
```



```
Final Model Output
------------------

                        Model Summary
--------------------------------------------------------------
R                       0.551       RMSE                0.177
R-Squared               0.304       MSE                 0.032
Adj. R-Squared          0.295       Coef. Var           2.768
Pred R-Squared          0.279       AIC               -95.607
MAE                     0.138       SBC               -83.135
--------------------------------------------------------------
 RMSE: Root Mean Square Error
 MSE: Mean Square Error
 MAE: Mean Absolute Error
 AIC: Akaike Information Criteria
 SBC: Schwarz Bayesian Criteria

                               ANOVA
-------------------------------------------------------------------
              Sum of
             Squares        DF    Mean Square       F         Sig.
-------------------------------------------------------------------
Regression     2.296         2          1.148    35.813    0.0000
Residual       5.258       164          0.032
Total          7.554       166
-------------------------------------------------------------------

                                 Parameter Estimates
------------------------------------------------------------------------------------
       model      Beta    Std. Error    Std. Beta       t        Sig     lower    upper
------------------------------------------------------------------------------------
 (Intercept)     5.526         0.123                 44.846    0.000    5.282    5.769
 Mean_mean_B     0.010         0.001        0.493     7.396    0.000    0.007    0.012
Mean_mean_BR     0.003         0.001        0.165     2.469    0.015    0.001    0.006
------------------------------------------------------------------------------------
```

Here we used the step-wise selection method with $\alpha = 0.05$ for sequential model selection. To perform the selection, we first created a model without the predictors that are necessary to answer our key objective, which were Mean_mean_HSV, Grade, and their interaction term. This means our full model before the selection was the response variable log(Mean_Area) with the predictors Mean_mean_R, Mean_mean_G, Mean_mean_B, Mean_mean_HE, Mean_mean_BR, and

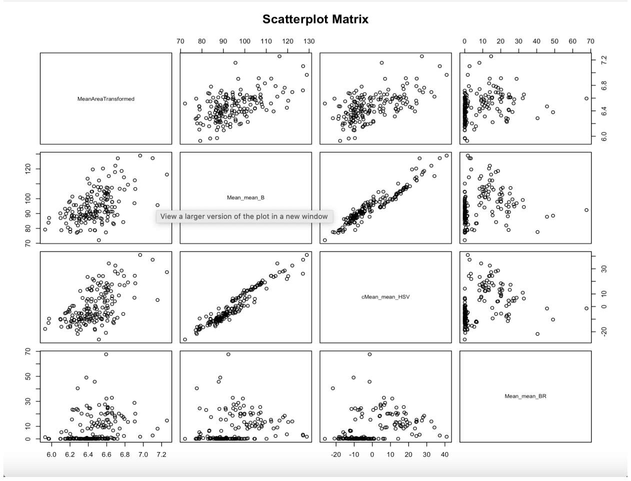
Mean_mean_Lab. Once the selection was performed, it removed everything but Mean_mean_B and Mean_mean_BR. This means that the model we obtained had the response variable log(Mean_Area), predictors Mean_mean_B, Mean_mean_BR, Mean_mean_HSV, and Grade, and the interaction term between Mean_mean_HSV and Grade.

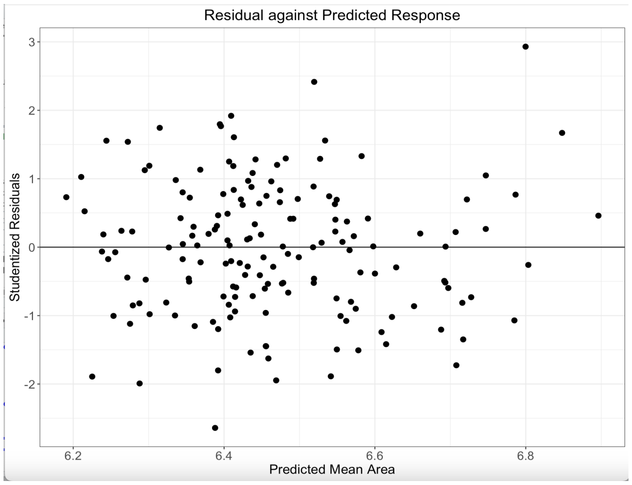
2.6 New Model

After we obtained our model after the sequential selection we removed the outliers in our data, centered the predictor Mean_mean_HSV, and refit the model. The model we now have includes the response variable log(Mean_Area), predictors Mean_mean_B, cMean_mean_BR, Mean_mean_HSV, and Grade, and the interaction term between cMean_mean_HSV and Grade. The least squares regression equation for this model is as follows:

log(Mean_Area) = 7.4405 - 0.0098*Mean_mean_B + 0.0014*Mean_mean_BR + 0.0176*cMean_mean_HSV - 0.1309*Grade - 0.0002*cMean_mean_HSV*Grade.

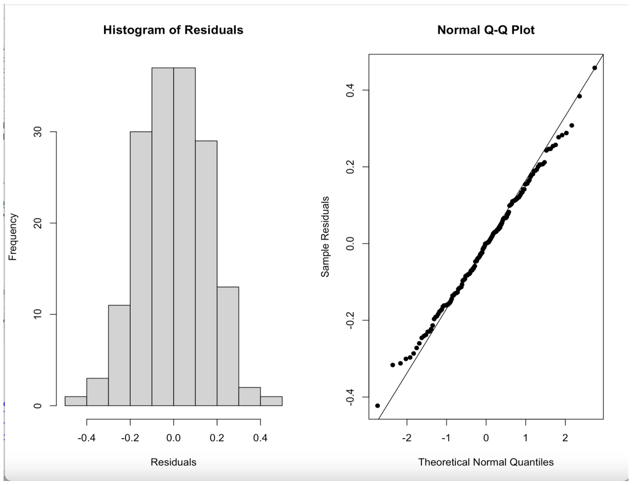
2.6.1 Checking Assumptions





We can see there is no curve in scatter plot matrix. The residual plot shows points scattered above and below 0 residual line with no curves. This means the assumption is reasonably met

Checking the constant variance assumption the residual plot shows similar variability in residuals for various predicted mean areas. We also conducted a Levene test and obtained a p-value of 0.9578. This means $p - value > \alpha = 0.05$ which means the assumption is reasonably met.



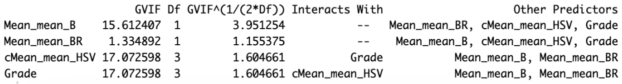
We assume a random sample for this dataset. So the independence assumption is reasonably met.

For the normality assumption, to check if it is met or not, we use a histogram of residuals and a normal Q-Q plot. We can see in the histogram that the data is normally distributed around 0.4 In the normal Q-Q plot, with no outliers, the Q-Q plot is close to the line. These observations lead us to believe that the normality assumption is reasonably met. Furthermore, we ran a Shapiro-Wilk test. We got a calculated p-value of 0.9917. This value is greater than $\alpha = 0.05$ so this means that the normality assumption is reasonably met.

All together, none of our plots indicate any violation of independence, linearity, constant variance, or normality, so all of the MLR model assumptions are reasonably met for our new model

2.6.2 VIF

When checking the VIF values, our original model from sequential selection yielded VIF values greater than 5. This means we have concern for multicollinearity.



```
                    GVIF Df GVIF^(1/(2*Df)) Interacts With                  Other Predictors
Mean_mean_B    15.612407  1        3.951254             --   Mean_mean_BR, cMean_mean_HSV, Grade
Mean_mean_BR    1.334892  1        1.155375             --    Mean_mean_B, cMean_mean_HSV, Grade
cMean_mean_HSV 17.072598  3        1.604661          Grade            Mean_mean_B, Mean_mean_BR
Grade          17.072598  3        1.604661 cMean_mean_HSV            Mean_mean_B, Mean_mean_BR
```

Because of these high VIF values, we removed the Mean_mean_B predictor. We then refit the model with response variable log(Mean_Area), predictors Mean_mean_BR, cMean_mean_HSV, and Grade, and interaction term cMean_mean_HSV*Grade. Using this model yielded VIF values less than 5. This means there is no significant concern for multicollinearity.

2.6.3 Model Significance

In our analysis of the model significance we start by stating the null hypothesis (H_0) which assumes that all values of β_1 through β_5 are zero, showing how none of the predictors have a significant effect on the response variable. The alternative hypothesis (H_a) is the counter stating how one of the coefficients doesnt equal 0, indicating significance.

The statistical test used to evaluate these hypotheses is the F-test, which has yielded an F-statistic of 25.7 with degrees of freedom (5, 161). The result is associated with a p-value of $5.186011 * 10^{-19}$, considering an $\alpha = 0.05$, with the p-value being less than α it leads us to reject the null hypothesis, concluding that there is evidence that the multiple linear regression model with our predictors is significant in predicting the median Mean_Area.

Furthermore, the adjusted R^2 is 0.4266, meaning 42.66% of variability in the transformed log(Mean_Area) is explained by the model with explanatory variables Mean_mean_B, Mean_mean_BR, cMean_mean_HSV, Grade and the centered interaction term after adjusting for complexity of the mode.l

2.6.4 Hypothesis Test

For our hypothesis test based on our research question, we tested whether our interaction term cMean_mean_HSV*Grade is significant or not. For this test we used the null hypothesis $H_0 : \beta_5 = 0$ and the alternative hypothesis $H_a : \beta_5 \neq 0$, where our interaction term is denoted by β_5. Conducting the test, we obtained a p-value of 0.906. This p-value is greater than $\alpha = 0.05$ which means we fail to reject the null H_0 and conclude that the interaction term is not significant. This means that there is sufficient evidence to assume that the variable representing the average lightness across nuclei does not classify the presence of usual ductal hyperplasia (UDH) from any other diagnosis in our dataset (nUDH).

3 Final Model and Conclusion

Because the hypothesis test showed us that the interaction is not significant, so we must remove it from our model and refit the model without it. Doing this we obtain a model with the response variable log(Mean_Area) and predictors Mean_mean_BR, cMean_mean_HSV, and Grade

In conclusion, our regression analysis has provided significant insights into the relationships between nuclear area, color intensity, and their interactions with tissue grades. Although the interaction between mean lightness on the HSV scale and Grade was not significant, the variables Mean_mean_BR, cMean_mean_HSV, and Grade still play a crucial role in predicting the log-transformed Mean_Area. Overall, this study shows the importance of color metrics and casual of distinguishing between UDH and non UDH, directly addressing our initial research question on predictiveness of color intensity in the biopsied tissue.