Regression Analysis on Cancer

Dan Cahill, Cesar McElaney, Maxwell Owens

Research Question and Variables

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.

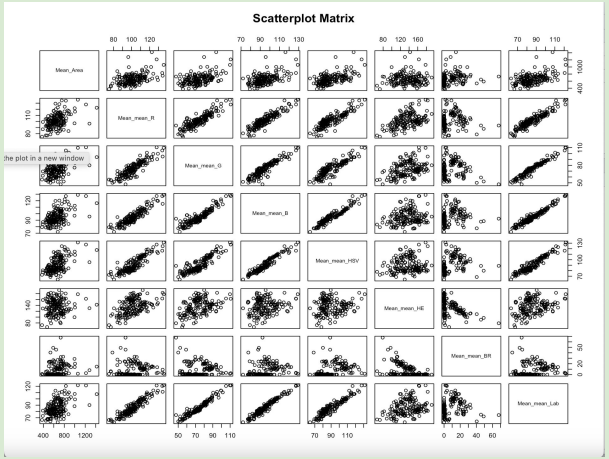
What happens to the mean area when we fit a model with an interaction term between the average lightness across nuclei on the HSV scale and grade?

- Mean_Area: this is the average nuclear area across nuclei
- Mean_mean_R: Average red intensity across nuclei on RGB scale
- Mean_mean_G: Average green intensity across nuclei on RGB scale
- Mean_mean_B: Average blue intensity across nuclei on RGB scale
- Mean_mean_HSV: Average lightness/value across nuclei on HSV scale
- Mean_mean_HE: Average lightness across nuclei from H&E color deconvolution
- Mean_mean_BR: Average lightness across nuclei on BR grey scale
- Mean_mean_Lab: Average lightness across nuclei on Lab and Luv scale
- Grade: from ClassGrade, separated into either UDH or nUDH

Initial Model

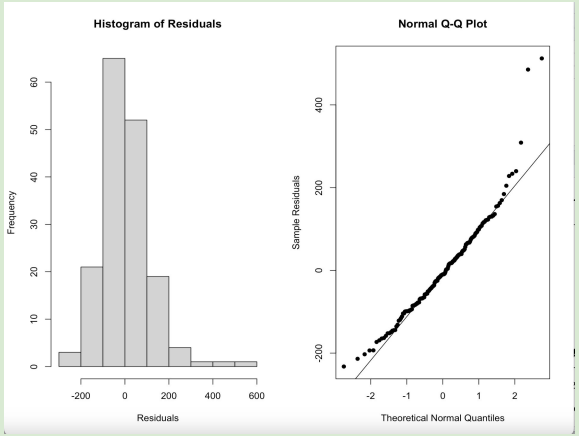
- Original ClassGrade variable came with multiple categories
  - High DCIS, Medium DCIS, Low DCIS, UDH
  - Changed it to UDH and nUDH
- Response variable: Mean_Area
- Predictors variable: Grade, Mean_mean_R, Mean_mean_G, Mean_mean_B, Mean_mean_HSV, Mean_mean_HE, Mean_mean_BR, Mean_mean_Lab
- Interaction term: Mean_mean_HSV*Grade

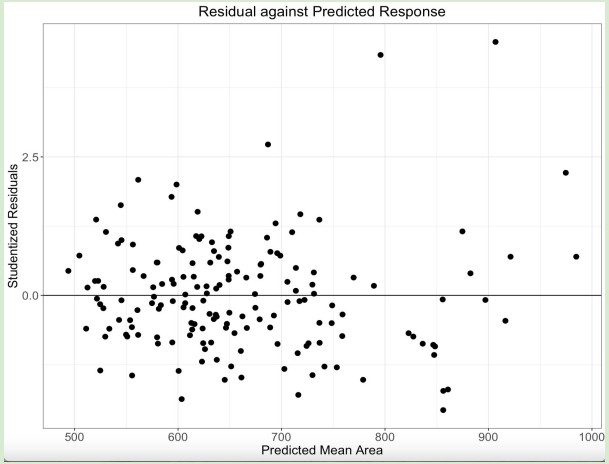
Assumptions



Constant variance

- Variability is different for for larger predicted values
- Levene test
  - P-value = 0.03114
  - P-value < α = 0.05
- Assumption is violated





Linearity

- No curve in scatter plot matrix
- Residual plot shows points scattered above and below 0 residual line with no curves
- Assumption is reasonably met

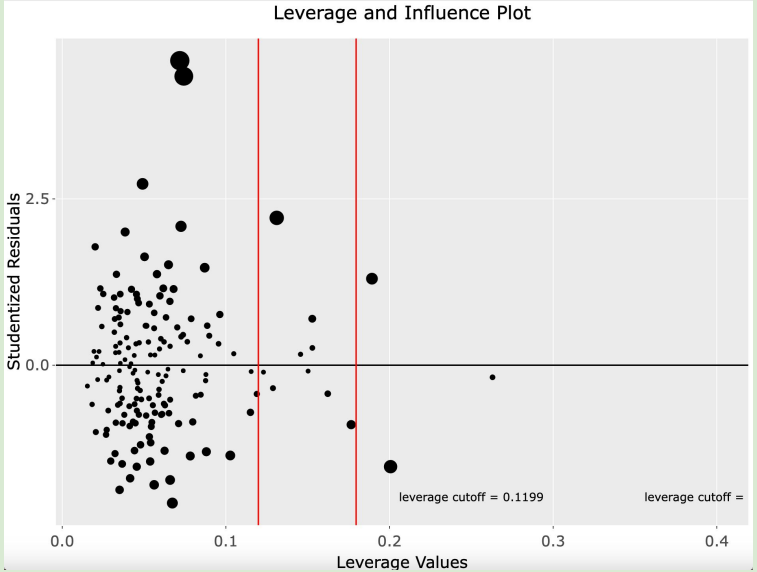
Independence

- Assume random sample for data set
- Assumption is reasonably met

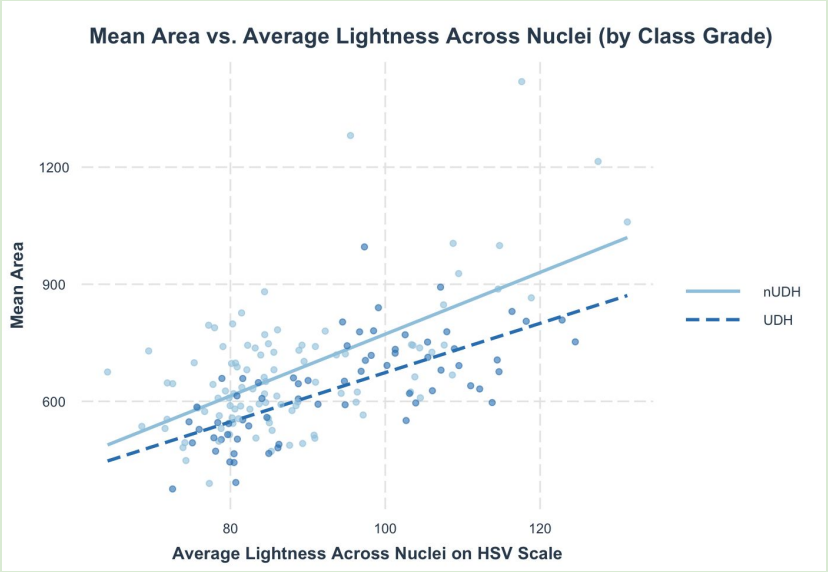
Normality

- Q-Q is close to the line in the middle, tails depart from straight line
- Shapiro-Wilk
  - P-value = $2.408*10^{-6}$
  - P-value < α = 0.05
- Assumption is violated

Leverage and Interaction Plots



- Raises concern for outliers

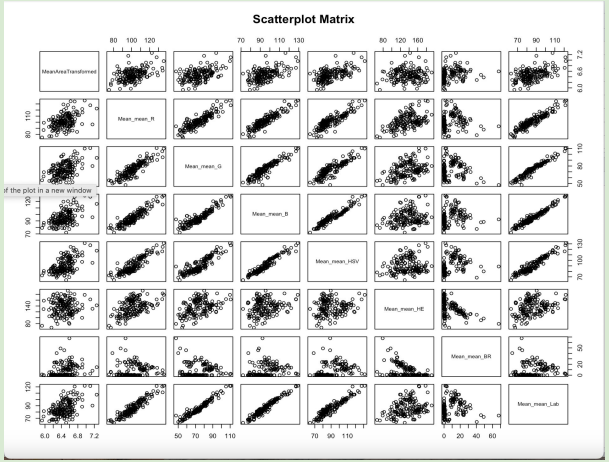


- Interaction term between Grade and Mean_mean_HSV
- Lines are not parallel
  - Consider interaction term

Transformation Model with Interaction Term

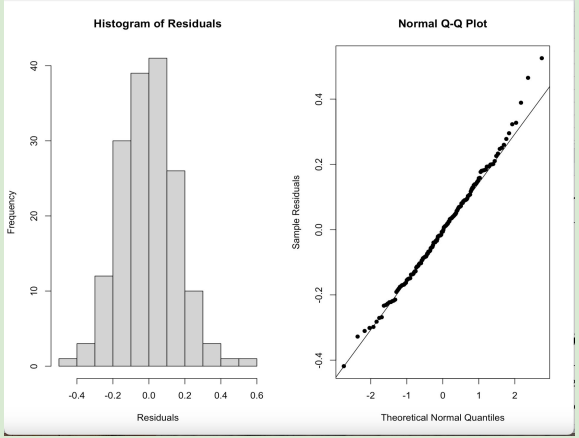
- Saw violations in assumptions, high studentized residuals, and high leverage values
  - Transformed the response variable
- Response variable: log(Mean_Area)
- Predictors variable: Grade, Mean_mean_R, Mean_mean_G, Mean_mean_B, Mean_mean_HSV, Mean_mean_HE, Mean_mean_BR, Mean_mean_Lab
- Interaction term: Mean_mean_HSV*Grade

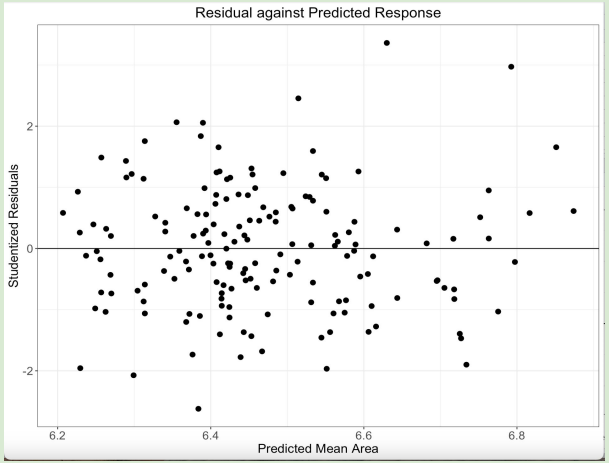
Assumptions



Constant variance

- Residual plot shows similar variability in residuals for various predicted mean areas
- Levene test
  - P-value = 0.9578
  - P-value > α
- Assumption is reasonably met





Linearity

- No curve in scatter plot matrix
- Residual plot shows points scattered above and below 0 residual line with no curves
- Assumption is reasonably met

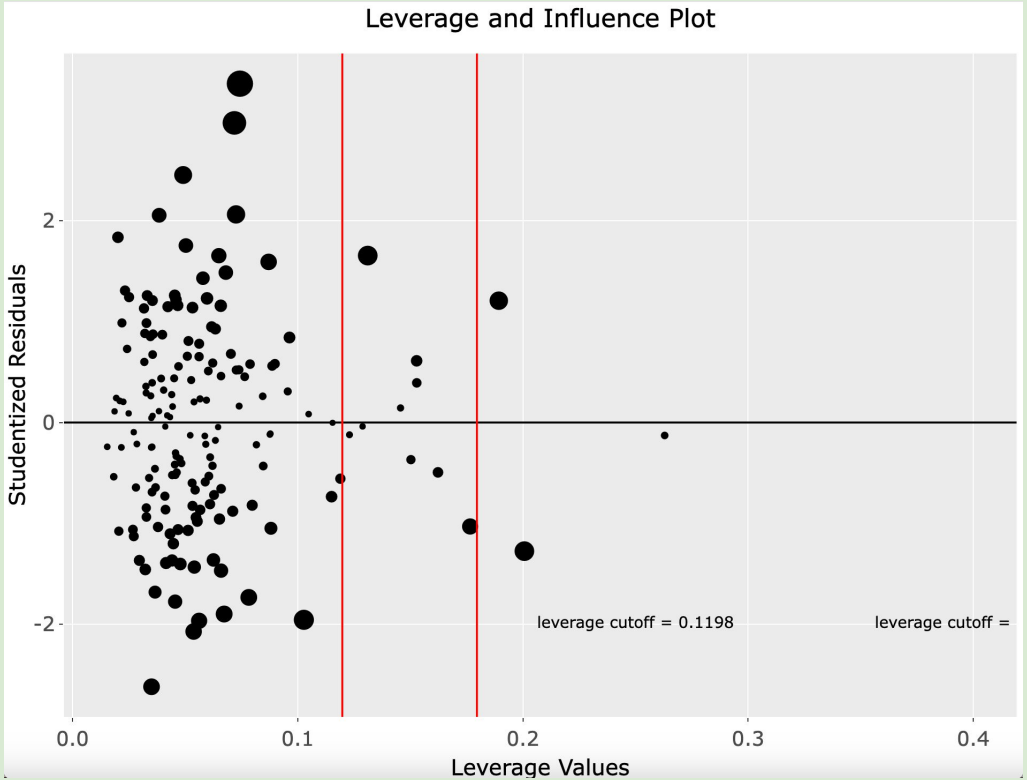
Independence

- Assume random sample for data set
- Assumption is reasonably met

Normality

- Histogram of residuals is normally distributed
- Q-Q is close to the line
- Shapiro-Wilk
  - P-value = 0.6542
  - P-value > α
- Assumption is reasonably met

Leverage Plot After Transformation



- No concern for outliers after transformation
- Many observations can be considered as influential due to their Cook's distance value

Sequential Model Selection

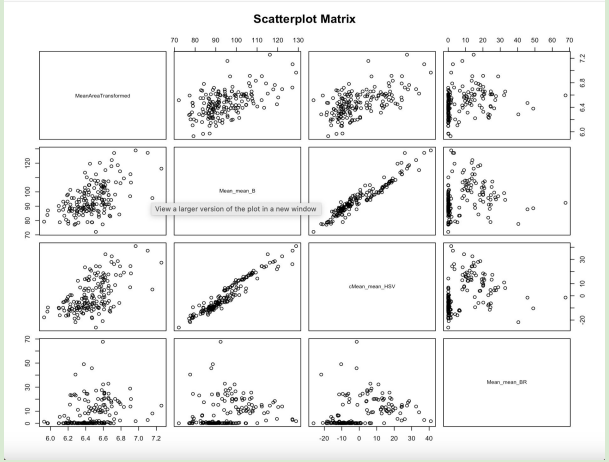
- Full model
  - Response variable: log(Mean_Area)
  - Predictors variable: Mean_mean_R, Mean_mean_G, Mean_mean_B, Mean_mean HE, Mean_mean_BR, and Mean_mean_Lab
- Selection
  - Removed everything but Mean_mean_B and Mean_mean_BR
- After selection
  - Response variable: log(Mean_Area)
  - Mean_mean_B, Mean_mean_BR, Mean_mean_HSV, and Grade
  - Interaction term: Mean_mean_HSV*Grade

Fit the Model

```
Call:
lm(formula = MeanAreaTransformed ~ Mean_mean_B + Mean_mean_BR +
    cMean_mean_HSV * Grade, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42262 -0.11475  0.00124  0.11112  0.55158

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.4404772  0.4379589  16.989  < 2e-16 ***
Mean_mean_B             -0.0098223  0.0045679  -2.150    0.033 *
Mean_mean_BR             0.0014100  0.0012759   1.105    0.271
cMean_mean_HSV           0.0175730  0.0039431   4.457 1.55e-05 ***
GradeUDH                -0.1308888  0.0265830  -4.924 2.09e-06 ***
cMean_mean_HSV:GradeUDH -0.0002287  0.0019288  -0.119    0.906
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1615 on 161 degrees of freedom
Multiple R-squared:  0.4438,    Adjusted R-squared:  0.4266
F-statistic:  25.7 on 5 and 161 DF,  p-value: < 2.2e-16
```

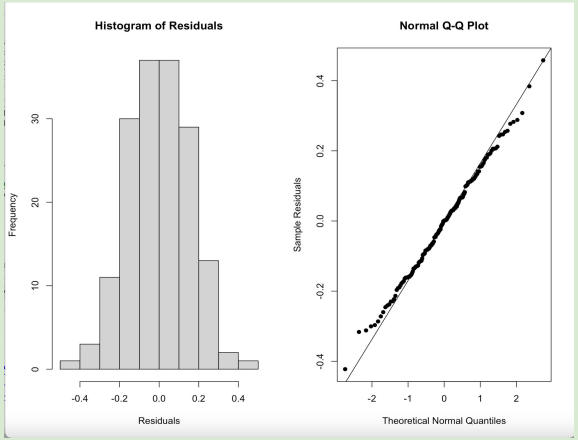
- Remove the outliers
  - Fit the model with a centered predictor
- Response variable: log(Mean_Area)
- Predictor variables: Mean_mean_B, Mean_mean_BR, cMean_mean_HSV, Grade
- Centered interaction term: cMean_mean_HSV*Grade
- log(Mean_Area) = 7.4405 - 0.0098*Mean_mean_B + 0.0014*Mean_mean_BR + 0.0176*cMean_mean_HSV - 0.1309*Grade - 0.0002*cMean_mean_HSV*Grade
  - Least squares regression

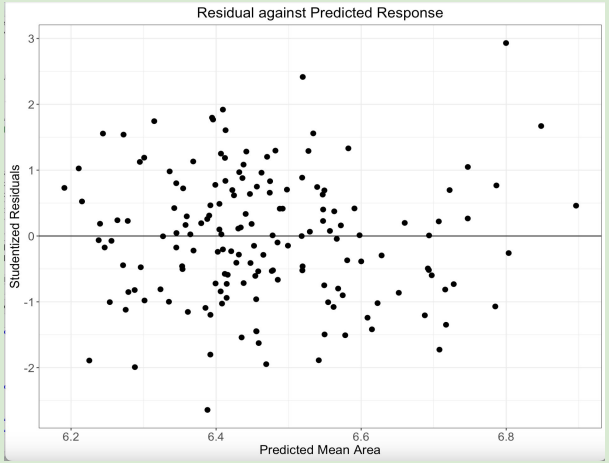
Assumptions



Constant variance

- Residual plot shows similar variability in residuals for various predicted mean areas
- Levene test
  - P-value = 0.9578
  - P-value > α
- Assumption is reasonably met





Linearity

- No curve in scatter plot matrix
- Residual plot shows points scattered above and below 0 residual line with no curves
- Assumption is reasonably met

Independence

- Assume random sample for data set
- Assumption is reasonably met

Normality

- Histogram of residuals is normally distributed
- Q-Q is close to the line
- Shapiro-Wilk
  - P-value = 0.9917
  - P-value > α
- Assumption is reasonably met

VIF

- Original model from sequential selection yielded VIF values > 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

- Remove Mean_mean_B predictor
- Refit model with response log(Mean_Area), predictors Mean_mean_BR, cMean_mean_HSV, Grade, and interaction term cMean_mean_HSV*Grade
- Using this model yielded VIF values < 5
  - No significant concern for multicollinearity

	GVIF	Df	GVIF^(1/(2*Df))	Interacts With	Other Predictors
Mean_mean_BR	1.143919	1	1.069541	--	cMean_mean_HSV, Grade
cMean_mean_HSV	1.143919	3	1.022663	Grade	Mean_mean_BR
Grade	1.143919	3	1.022663	cMean_mean_HSV	Mean_mean_BR

Model Significance

```
Residual standard error: 0.1615 on 161 degrees of freedom
Multiple R-squared:  0.4438,    Adjusted R-squared:  0.4266
F-statistic:  25.7 on 5 and 161 DF,  p-value: < 2.2e-16
```

- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
- H_a: at least one $\beta_j \neq 0$, j = 1, … , 5
- F-statistic: 25.7 with (5, 161) df
- P-value = $5.186011 * 10^{-19}$
- $\alpha = 0.05$
- $R^2 = 0.4438$
- Adjusted $R^2 = 0.4266$

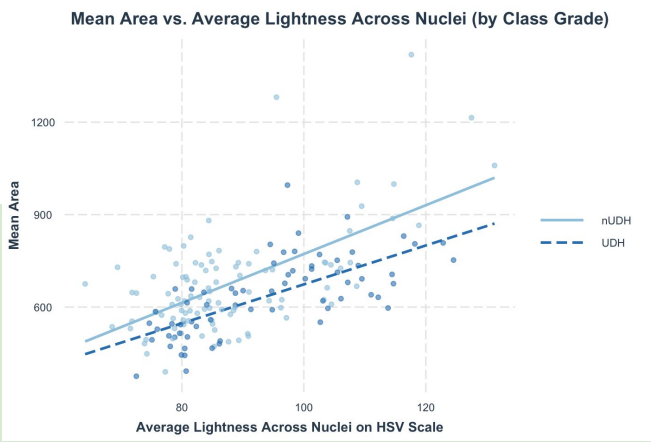
Adj R^2 interpretation - 42.66% of variability in the transformed log(Mean_Area) is explained by the model with explanatory variables Mean_mean_B, Mean_mean_BR, cMean_mean_HSV, Grade and the centered interaction term after adjusting for complexity of the model

- P-value < α
- Reject H_0 and conclude that there is evidence that the multiple linear regression model with our predictors is significant in predicting the median Mean_Area

R^2 interpretation - 44.38% of variability in the transformed log(Mean_Area) can be explained by the linear regression model with the predictors Mean_mean_B, Mean_mean_BR, cMean_mean_HSV, Grade and the centered interaction term

Hypothesis Test

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.4404772  0.4379589  16.989  < 2e-16 ***
Mean_mean_B               -0.0098223  0.0045679  -2.150    0.033 *
Mean_mean_BR               0.0014100  0.0012759   1.105    0.271
cMean_mean_HSV             0.0175730  0.0039431   4.457 1.55e-05 ***
GradeUDH                  -0.1308888  0.0265830  -4.924 2.09e-06 ***
cMean_mean_HSV:GradeUDH   -0.0002287  0.0019288  -0.119    0.906
```



$H_0: \beta_5 = 0$ vs. $H_a: \beta_5 \neq 0$

- Since the lines representing UDH and nUDH are not parallel, we consider this interaction term
- β_5 signifies the coefficient of the interaction term
- p-value = 0.906 and $\alpha = 0.05$
- Since p-value > α, we fail to reject H_0 to conclude that the interaction term is insignificant

- In conclusion, 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).

Final Model

- Interaction term is not significant
  - Remove it from our model
  - Refit to our final model
- Response Variable: log(Mean_Area)
- Predictor Variables: Mean_mean_BR, cMean_mean_HSV, and Grade