

Final Project Report

Raynah Cheng

2024-12-07

Abstract:

Introduction:

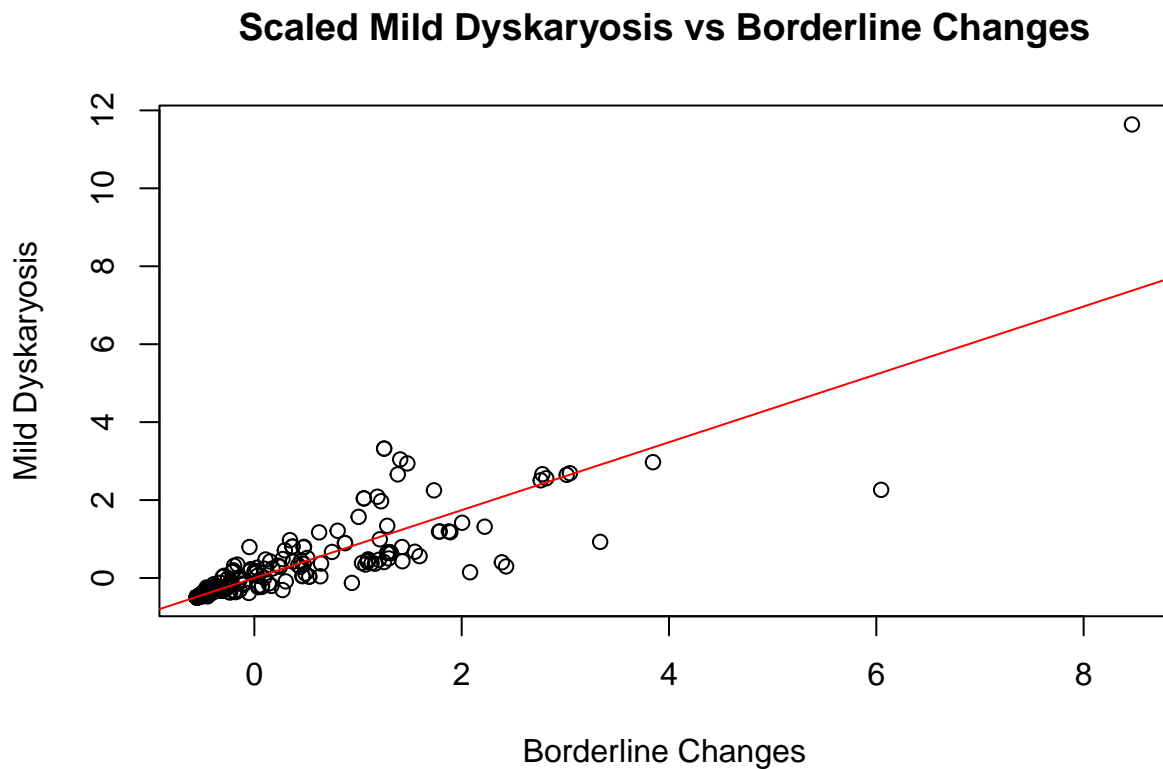
Cervical cancer is a type of cancer that occurs in the cervix, specifically the cells lining the cervix wall. It is the fourth most common cancer in women globally with around 660,000 new cases and 350,000 deaths in 2022, according to the WHO. Our project contains data on patients that have cervical cancer, taken from the NHS England. There were many datasets, but the one that we focused on was titled `k61_sample_result_age_source`. In this dataset, there are 330 observations, with 5 variables of focus. We chose on this one because it contained data on cervical cancer changes based on age and screening. The relevant variables we are looking at are mild, moderate, and severe dyskaryosis, age, and borderline changes. These terms are broad and hard to understand, so defining them is useful. Dyskaryosis is the change in appearance of cells that line the cervix. Mild, moderate, and severe are the classifications that this dataset uses. Mild is little change, moderate is moderate change, and severe is severe change. These changes are defined based on a 3-year screening. Borderline changes are abnormal changes in the cervix that are often pre-cancerous. These could indicate cervical cancer.

Data Cleaning:

Exploratory Data Analysis:

Results:

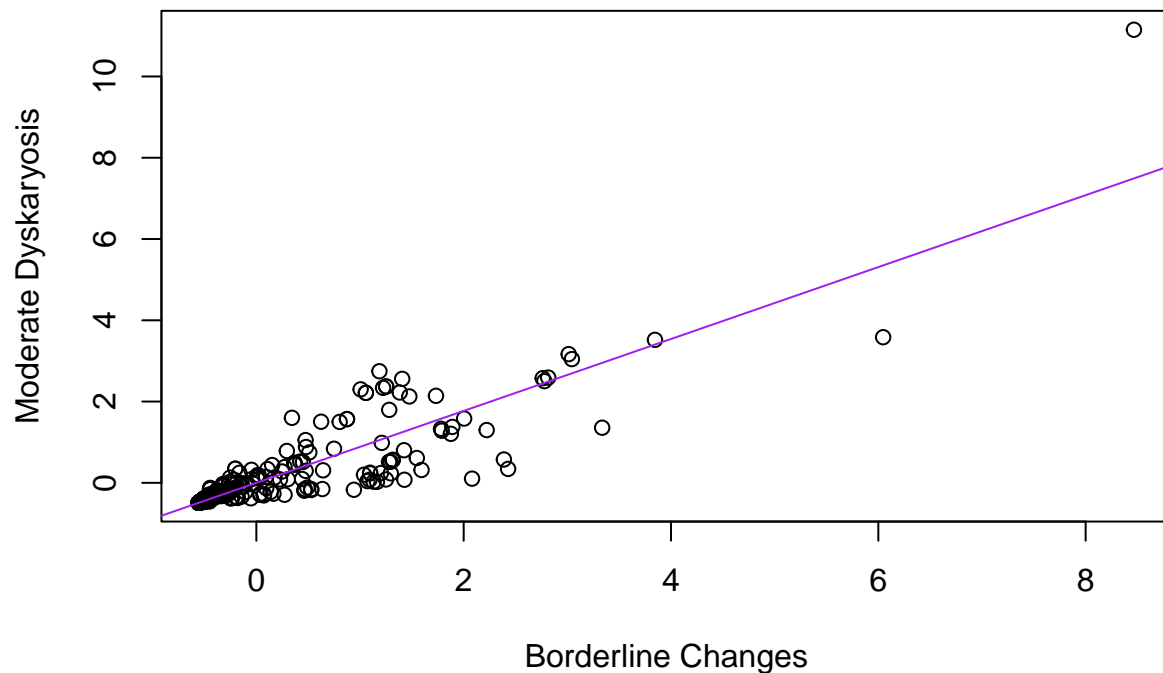
Scaled Linear Regression:



```
##
## Call:
## lm(formula = Mild_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.0040 -0.0170 -0.0125  0.0281  4.2644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.213e-16  2.708e-02   0.00      1
## Borderline_changes 8.711e-01  2.712e-02  32.12 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

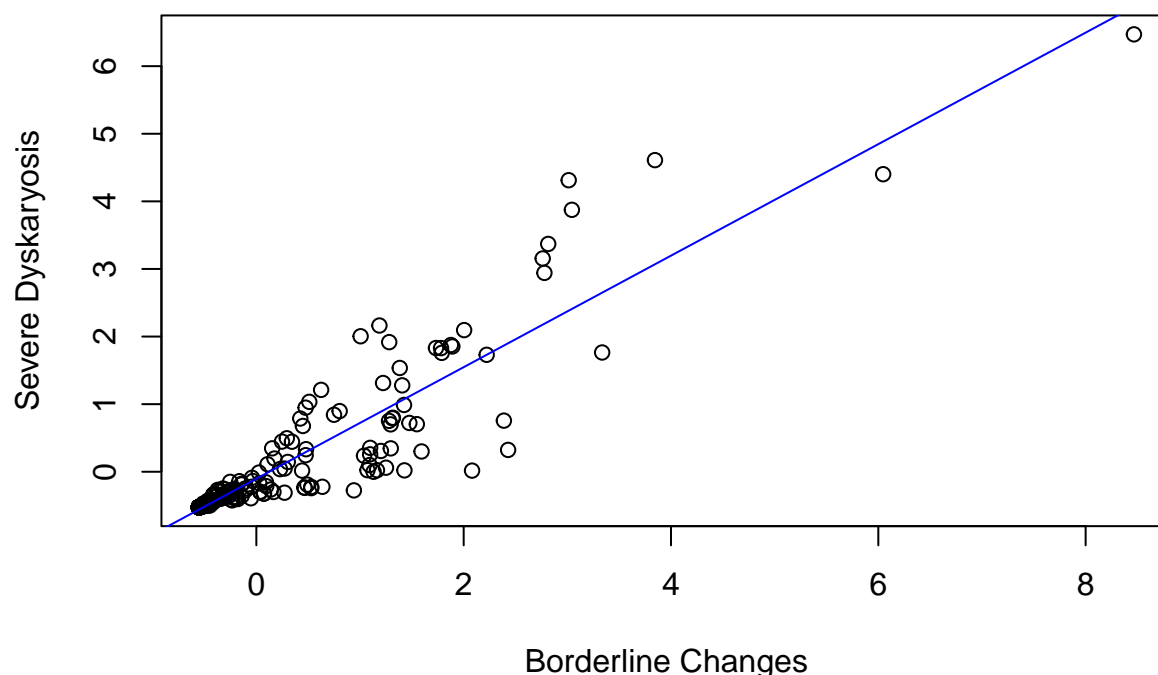
```
## Residual standard error: 0.4919 on 328 degrees of freedom
## Multiple R-squared:  0.7587, Adjusted R-squared:  0.758
## F-statistic: 1031 on 1 and 328 DF,  p-value: < 2.2e-16
```

Scaled Moderate Dyskaryosis vs Borderline Changes



```
##
## Call:
## lm(formula = Moderate_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8067 -0.0007  0.0038  0.0512  3.6586
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.621e-16  2.568e-02    0.0      1
## Borderline_changes 8.849e-01  2.572e-02   34.4 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4665 on 328 degrees of freedom
## Multiple R-squared:  0.783, Adjusted R-squared:  0.7823
## F-statistic: 1184 on 1 and 328 DF,  p-value: < 2.2e-16
```

Scaled Severe Dyskaryosis vs Borderline Changes



```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59706 -0.02801  0.03012  0.03551  1.93013
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.10225    0.02608   -3.92 0.000116 ***
## Borderline_changes  0.82478    0.02338   35.27 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4016 on 238 degrees of freedom
## (90 observations deleted due to missingness)
## Multiple R-squared:  0.8394, Adjusted R-squared:  0.8387
## F-statistic: 1244 on 1 and 238 DF, p-value: < 2.2e-16
```

For the first linear regression model of Mild Dyskaryosis vs Borderline Changes, the coefficient of the slope of the line was significant with an alpha level of <0.001 . The intercept of the line was not significant at any level. The adjusted R-squared value is 0.758, indicating that 75.8% of the variance in the data can be explained by the model. The F statistic is large with a p-value < 0.05 , indicating that this regression is significant.

For the second linear regression model of Moderate Dyskaryosis vs Borderline Changes, the coefficient of the slope of the line was significant with an alpha level of <0.001 . The intercept of the line was not significant at any level. The adjusted R-squared value is 0.7823, indicating that 78.23% of the variance in the data can be explained by the model. The F statistic is large with a p-value < 0.05 , indicating that this regression is significant.

For the third linear regression model of Severe Dyskaryosis vs Borderline Changes, the coefficient of the slope of the line was significant with an alpha level of <0.001 . The intercept of the line was not significant at any level. The adjusted R-squared value is 0.8387, indicating that 83.87% of the variance in the data can be explained by the model. The F statistic is large with a p-value of < 0.05 , indicating that this regression is significant.

The regression model of Severe Dyskaryosis vs Borderline Changes appears to be the best at predicting incidence of severe dyskaryosis from borderline changes due to it having the largest F-statistic value (1244) as well as the largest adjusted R-squared value (0.8387). This could be because borderline changes in a neoplasma will typically be larger and more easily detectable in later stage cancers, meaning that large borderline changes would have a higher correlation with severe dyskaryosis.

BIC MIC Model Comparison:

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ ., data = train_frame)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-1583.04	-69.40	66.49	94.30	1515.34

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-92.17422	41.51175	-2.220	0.0278 *
Borderline_changes	0.09507	0.01087	8.748	2.77e-15 ***
Mild_dyskaryosis	-0.14662	0.02321	-6.317	2.49e-09 ***
Moderate_dyskaryosis	1.45536	0.09598	15.163	< 2e-16 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic: 1055 on 3 and 161 DF, p-value: < 2.2e-16

##
## Call:
## lm(formula = Severe_dyskaryosis ~ Borderline_changes + Mild_dyskaryosis +
##     Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-1583.04	-69.40	66.49	94.30	1515.34

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-92.17422	41.51175	-2.220	0.0278 *
Borderline_changes	0.09507	0.01087	8.748	2.77e-15 ***

```
## Mild_dyskaryosis      -0.14662    0.02321  -6.317 2.49e-09 ***
## Moderate_dyskaryosis  1.45536    0.09598  15.163 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic: 1055 on 3 and 161 DF, p-value: < 2.2e-16

##
## Call:
## lm(formula = Severe_dyskaryosis ~ Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2625.70   -49.93    56.48    68.43   1635.61
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -61.79307   48.85798  -1.265    0.208
## Moderate_dyskaryosis  1.17963    0.02726  43.274 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 508.9 on 163 degrees of freedom
## Multiple R-squared:  0.9199, Adjusted R-squared:  0.9194
## F-statistic: 1873 on 1 and 163 DF, p-value: < 2.2e-16
```

The BIC model chooses all of the dyskaryosis variables, with equal about equal levels of significance. However, we can see that the moderate dyskaryosis t value is higher, at 15.163. For our BIC model, 95.16% of the variance can be explained by our model. The MIC model only chooses the moderate dyskaryosis variable, at a .001 level of significance, with a much higher t value than the BIC model. However, the variance that can be explained goes down a little bit, at 91.99%.

Model	R^2	RMSE	MAE
MIC	.997	1128.59	248.29
BIC	.996	806.2	194.54

The BIC model has a lower R^2 value, but the errors are lower. The MIC model has a higher R^2 value, but the errors are higher. I think that we would pick the BIC model because the R^2 is not that much lower and it predicts our data a lot better.

Lasso Model:

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##              s1
## (Intercept)  -50.717419429
## Inadequate    .
## Negative      -0.003401138
## Borderline_changes  0.056396976
## Mild_dyskaryosis .
```

```
## Moderate_dyskaryosis      0.682117740
## Severe_dyskaryosis_Inv    21.832583199
## Glandular_neoplasia       1.101335422
```

```
##                               [,1]
## lambda.1se 0.9402083
```

```
## [1] 483.9233
```

```
## [1] 189.7857
```

For the lasso model, the independent variables acting as name tags for the data were removed so the model could be simplified and fitted to the best predictors. The goal of this model is to see if any variables within this data would be a good predictor of a severe dyskaryosis diagnosis. Looking at the R-squared of the lasso model, we see that the model explains 95.18% of the variability within the data which means the model is fitted very well to the data. However, looking at the RMSE and MAE, we see that the model is very lacking in its predictions. This can be because of the odd way the data is organized and collected.

Some data within this dataset is taken across multiple year frames, while others are taken within the same year. The data also varies widely between the number of observations that are taken at each “location”. The largely variability and inconsistency of the data could contribute to a high RMSE and MAE seen within the lasso model.

Because of the data, a portion of the data is singled out to be evaluated again. The data taken from 2012-2023 in England for all ages were subsetting into a new data frame, and the lasso model was ran again under these restrictions.

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                               s1
## (Intercept)      3.496182e+01
## Inadequate        .
## Negative          4.619480e-05
## Borderline_changes 5.220385e-02
## Mild_dyskaryosis  .
## Moderate_dyskaryosis 3.320382e-01
## Severe_dyskaryosis_Inv .
## Glandular_neoplasia 1.557034e+00
```

```
##                               [,1]
## lambda.1se 0.9867399
```

```
## [1] 148.4693
```

```
## [1] 66.5433
```

The R-squared, RMSE, and MAE have all improved from the previous model. However, since the data does have the quality of time attached to it, it may still be an issue in terms of predictions. This may explain the large RMSE and MAE values.

Chosen Model:

Lasso Model, but linear regression and BIC shows strong correlation between severe dyskaryosis and borderline changes.

Appendix:

Data Cleaning:

```
source <-  
  read.csv("cervical-programme-annual-2022-23-csvs/kc61_sample_result_age_source.csv")  
  
dim(source)
```

```
## [1] 330 17
```

```
#columns of interest: Indicator (age), Borderline changes, mild, moderate,  
#and severe dyskaryosis
```

```
#only column that has NA's is severe dyskaryosis; sub in values for NA to clean  
source$Severe_dyskaryosis <- gsub("not included", NA, source$Severe_dyskaryosis)  
sum(is.na(source$Severe_dyskaryosis) == TRUE)
```

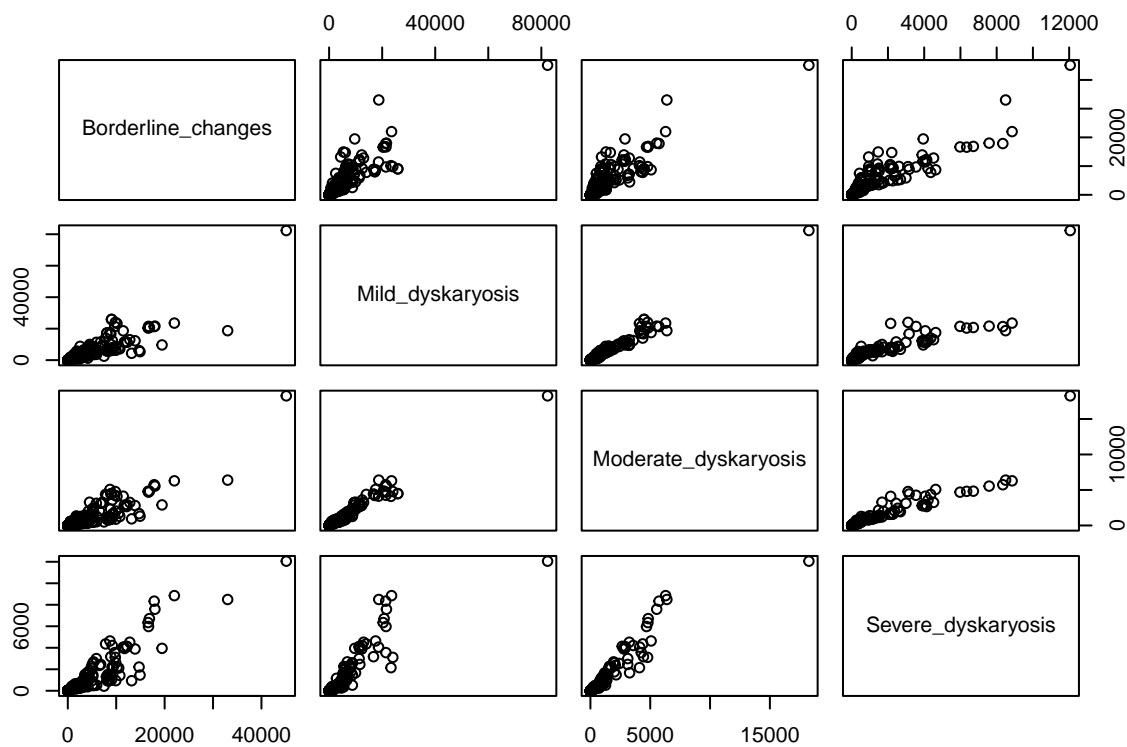
```
## [1] 90
```

```
uniq_ind <- unique(source$Indicator)  
for (i in 1:length(uniq_ind)){  
  source$Indicator <- gsub(uniq_ind[i], i, source$Indicator)  
}  
source$Indicator <- as.numeric(source$Indicator)  
#changing each age group to correspond with a number from 1-19. for example, <20 is 1  
  
#removing NAs and converting to numeric values  
severe_NA <- source[is.na(source$Severe_dyskaryosis) == FALSE, ]  
severe_NA$Severe_dyskaryosis <- as.numeric(severe_NA$Severe_dyskaryosis)  
  
comparison_data <- source[,c("Borderline_changes", "Mild_dyskaryosis", "Moderate_dyskaryosis", "Severe_dyskaryosis")]  
comparison_data$Severe_dyskaryosis <- as.numeric(comparison_data$Severe_dyskaryosis)
```

Subsetting Source Data:

```
## Subset source data for only borderline changes and mild, moderate, severe dyskaryosis
```

```
comparison_data <- source[,c("Borderline_changes", "Mild_dyskaryosis", "Moderate_dyskaryosis", "Severe_dyskaryosis")]  
comparison_data$Severe_dyskaryosis <- as.numeric(comparison_data$Severe_dyskaryosis)  
  
pairs(comparison_data)
```

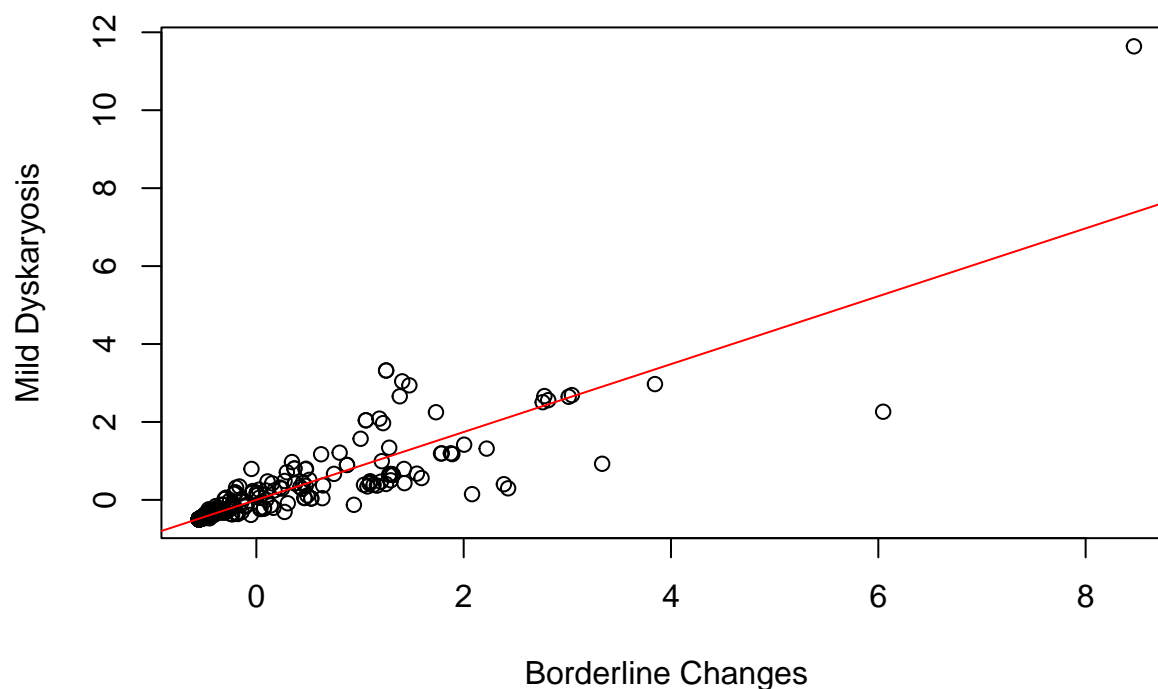



Scaled Linear Regression:

```
scaled_data <- as.data.frame(scale(comparison_data))

# mild dyskaryosis scaled
scaled_mild_model <- lm(Mild_dyskaryosis ~ Borderline_changes, data = scaled_data)
plot(scaled_data$Borderline_changes, scaled_data$Mild_dyskaryosis,
     main = "Scaled Mild Dyskaryosis vs Borderline Changes",
     xlab = "Borderline Changes",
     ylab = "Mild Dyskaryosis")
abline(scaled_mild_model, col = 'red')
```

Scaled Mild Dyskaryosis vs Borderline Changes



```
summary(scaled_mild_model)
```

```
##
## Call:
## lm(formula = Mild_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.0040 -0.0170 -0.0125  0.0281  4.2644
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.213e-16  2.708e-02   0.00      1
## Borderline_changes 8.711e-01  2.712e-02  32.12 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4919 on 328 degrees of freedom
## Multiple R-squared:  0.7587, Adjusted R-squared:  0.758
## F-statistic: 1031 on 1 and 328 DF, p-value: < 2.2e-16
```

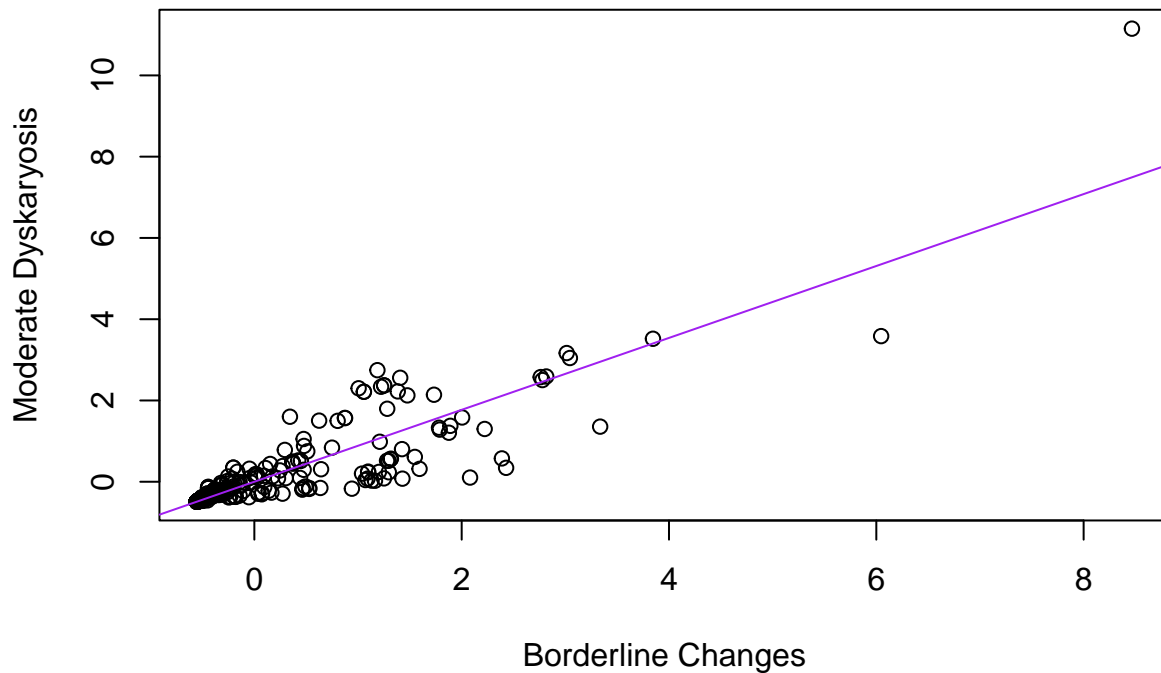
```
# moderate dyskaryosis scaled
scaled_mod_model <- lm(Moderate_dyskaryosis ~ Borderline_changes, data = scaled_data)
plot(scaled_data$Borderline_changes, scaled_data$Moderate_dyskaryosis,
     main = "Scaled Moderate Dyskaryosis vs Borderline Changes",
```

```

xlab = "Borderline Changes",
ylab = "Moderate Dyskaryosis")
abline(scaled_mod_model, col = 'purple')

```

Scaled Moderate Dyskaryosis vs Borderline Changes



```
summary(scaled_mod_model)
```

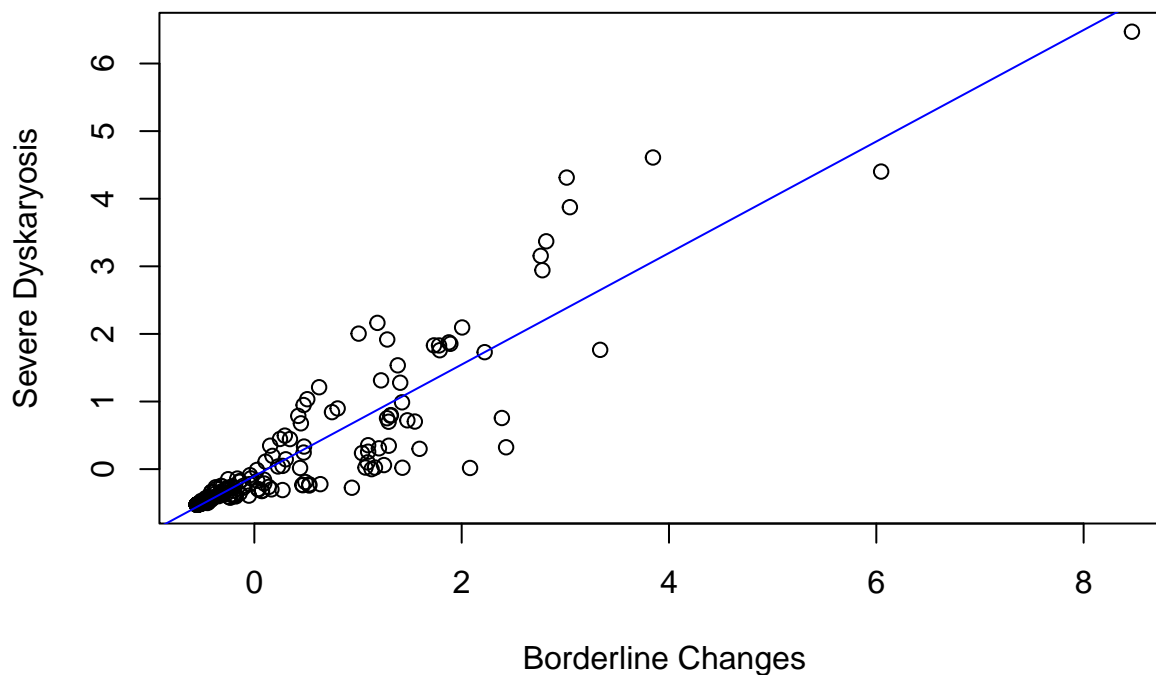
```

##
## Call:
## lm(formula = Moderate_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8067 -0.0007  0.0038  0.0512  3.6586
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.621e-16  2.568e-02     0.0      1
## Borderline_changes 8.849e-01  2.572e-02   34.4 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4665 on 328 degrees of freedom
## Multiple R-squared:  0.783, Adjusted R-squared:  0.7823
## F-statistic: 1184 on 1 and 328 DF, p-value: < 2.2e-16

```

```
# severe dyskaryosis scaled
scaled_severe_model <- lm(Severe_dyskaryosis ~ Borderline_changes, data = scaled_data)
plot(scaled_data$Borderline_changes, scaled_data$Severe_dyskaryosis,
     main = "Scaled Severe Dyskaryosis vs Borderline Changes",
     xlab = "Borderline Changes",
     ylab = "Severe Dyskaryosis")
abline(scaled_severe_model, col = 'blue')
```

Scaled Severe Dyskaryosis vs Borderline Changes



```
summary(scaled_severe_model)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Borderline_changes, data = scaled_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59706 -0.02801  0.03012  0.03551  1.93013
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.10225    0.02608   -3.92 0.000116 ***
## Borderline_changes  0.82478    0.02338   35.27 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4016 on 238 degrees of freedom
## (90 observations deleted due to missingness)
## Multiple R-squared: 0.8394, Adjusted R-squared: 0.8387
## F-statistic: 1244 on 1 and 238 DF, p-value: < 2.2e-16
```

BIC MIC Model Comparison:

```
#splitting data into test and train
train_index <- 1:165
test_index <- 166:nrow(comparison_data)

train_frame <- comparison_data[train_index,]
test_frame <- comparison_data[test_index,]

train_model <- lm(Severe_dyskaryosis ~ ., data = train_frame)
summary(train_model)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ ., data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1583.04   -69.40    66.49    94.30  1515.34
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -92.17422   41.51175  -2.220  0.0278 *
## Borderline_changes    0.09507    0.01087   8.748 2.77e-15 ***
## Mild_dyskaryosis    -0.14662    0.02321  -6.317 2.49e-09 ***
## Moderate_dyskaryosis  1.45536    0.09598  15.163 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared: 0.9516, Adjusted R-squared: 0.9507
## F-statistic: 1055 on 3 and 161 DF, p-value: < 2.2e-16
```

```
back_BIC <- step(train_model, direction = "backward",
                 k = log(nrow(train_frame)), trace = 0)
summary(back_BIC)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Borderline_changes + Mild_dyskaryosis +
##      Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1583.04    -69.40     66.49     94.30  1515.34
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -92.17422   41.51175   -2.220   0.0278 *
## Borderline_changes    0.09507    0.01087    8.748 2.77e-15 ***
## Mild_dyskaryosis    -0.14662    0.02321   -6.317 2.49e-09 ***
## Moderate_dyskaryosis    1.45536    0.09598   15.163 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 398.2 on 161 degrees of freedom
## Multiple R-squared:  0.9516, Adjusted R-squared:  0.9507
## F-statistic: 1055 on 3 and 161 DF,  p-value: < 2.2e-16
```

```
back_MIC <- step(train_model, direction = "backward",
                 k = nrow(train_frame), trace = 0)
summary(back_MIC)
```

```
##
## Call:
## lm(formula = Severe_dyskaryosis ~ Moderate_dyskaryosis, data = train_frame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2625.70   -49.93    56.48    68.43   1635.61
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -61.79307   48.85798   -1.265   0.208
## Moderate_dyskaryosis    1.17963    0.02726   43.274 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 508.9 on 163 degrees of freedom
## Multiple R-squared:  0.9199, Adjusted R-squared:  0.9194
## F-statistic: 1873 on 1 and 163 DF,  p-value: < 2.2e-16
```

```
test_MIC <- predict(back_MIC, test_frame)
test_BIC <- predict(back_BIC, test_frame)
```

```
cor(test_MIC, test_frame$Severe_dyskaryosis, use = "complete.obs")^2 #r-squared for back_MIC model
```

```
## [1] 0.9966765
```

```
cor(test_BIC, test_frame$Severe_dyskaryosis, use = "complete.obs")^2 #r-squared for back_BIC model
```

```
## [1] 0.9958021
```

```
errors_MIC <- test_frame$Severe_dyskaryosis - test_MIC
errors_BIC <- test_frame$Severe_dyskaryosis - test_BIC
```

```
sqrt(mean(errors_MIC^2, na.rm = TRUE)) #RMSE for back_MIC
```

```
## [1] 1128.586
```

```
sqrt(mean(errors_BIC^2, na.rm = TRUE)) #RMSE for back_BIC
```

```
## [1] 806.2014
```

```
mean(abs(errors_MIC), na.rm = TRUE) #MAE for back_MIC
```

```
## [1] 248.288
```

```
mean(abs(errors_BIC), na.rm = TRUE) #MAE for back_BIC
```

```
## [1] 194.5406
```

Lasso Model:

```
row_select <- sample(1:nrow(severe_NA), nrow(severe_NA) / 2)
col_exclude <- c(1:8, 17)
```

```
source_train <- severe_NA[row_select, -col_exclude]
source_test <- severe_NA[-row_select, -col_exclude]
```

```
suppressMessages(library(glmnet))
```

```
cn <- colnames(source_train)
exclude <- which(cn == "Severe_dyskaryosis" | cn == "CollectionYearRange")
X <- as.matrix(source_train[, -exclude])
```

```
suppressWarnings(source_lasso <- cv.glmnet(X, source_train$Severe_dyskaryosis))
coef(source_lasso)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
```

```
##              s1
## (Intercept)  32.04629768
## Inadequate      .
## Negative        .
## Borderline_changes  0.08762232
## Mild_dyskaryosis  .
## Moderate_dyskaryosis  0.49011428
## Severe_dyskaryosis_Inv  .
## Glandular_neoplasia  1.19278909
```

```
Y <- as.matrix(source_test[, -exclude])
suppressWarnings(source_predict <- predict(source_lasso, newx = Y))

cor(source_predict, source_test$Severe_dyskaryosis) ^ 2
```

```
##              [,1]
## lambda.1se 0.8941753
```

```
source_errors <- source_test$Severe_dyskaryosis - source_predict
sqrt(mean(source_errors ^ 2))
```

```
## [1] 729.4987
```

```
mean(abs(source_errors))
```

```
## [1] 326.4807
```

```
lasso_data <- severe_NA[severe_NA$CollectionYearRange == "2022-23" &
                        suppressWarnings(severe_NA$Indicator == c(1:13)), ]
row_select <- sample(1:nrow(lasso_data), nrow(lasso_data) / 2)
col_exclude <- c(1:8, 17)

source_train <- lasso_data[row_select, -col_exclude]
source_test <- lasso_data[-row_select, -col_exclude]

cn <- colnames(source_train)
exclude <- which(cn == "Severe_dyskaryosis")
X <- as.matrix(source_train[, -exclude])

suppressWarnings(source_lasso <- cv.glmnet(X, source_train$Severe_dyskaryosis))
coef(source_lasso)
```

```
## 8 x 1 sparse Matrix of class "dgCMatrix"
##                               s1
## (Intercept)          13.92074656
## Inadequate              .
## Negative                .
## Borderline_changes      0.08603366
## Mild_dyskaryosis         .
## Moderate_dyskaryosis     0.25211796
## Severe_dyskaryosis_Inv   .
## Glandular_neoplasia      2.41706535
```

```
Y <- as.matrix(source_test[, -exclude])
suppressWarnings(source_predict <- predict(source_lasso, newx = Y))

cor(source_predict, source_test$Severe_dyskaryosis) ^ 2
```

```
##                               [,1]
## lambda.1se 0.9856945
```

```
source_errors <- source_test$Severe_dyskaryosis - source_predict
sqrt(mean(source_errors ^ 2))
```

```
## [1] 125.0619
```



```
mean(abs(source_errors))
```

```
## [1] 53.79081
```

Sources:

Acknowledgements:

Data Cleaning: Paulina Yao

Scaled Linear Regression: Anika Dachiraju

BIC MIC Model Comparison: Raynah Cheng

Lasso Model: Paulina Yao

Bibliography: