Initial Multiple Regression Analysis of Color Values in Relation to Mean Area for Cancer Cells

Caleb Munson, Luke Geel, Leah Dion, Everett Henderson, Jiun Tseng

April 2021

Description of initial multiple linear regression model

Our initial model includes the 7 predictors (respectively Mean Area of Red, Green, Blue, HSV, LAB, HE, BR) in relation to the response, mean area across each nuclei. Each predictor is accounted for linearly. b_0 is the expected mean area assuming each predictor is zero.

$$\hat{Y} = 183.3774 - 5.8856X_1 - 4.3324X_2 - 2.7434X_3 + 12.7141X_4 + 4.1442X_5 + 1.1101X_6 - 0.1687X_7$$

Summary of estimated regression coefficients and standard errors

Summary of b_i values:

- $b_0 = 183.3774$
- $b_1 = -5.8856$
- $b_2 = -4.3324$
- $b_3 = -2.7434$
- $b_4 = 12.7141$
- $b_5 = 4.1442$
- $b_6 = 1.1101$
- $b_7 = -0.1687$

Summary of the absolute value of the residuals:

- mean = 90.691
- s.d. = 81.0104
- min = 2.099
- max = 585.026
- median = 72.985

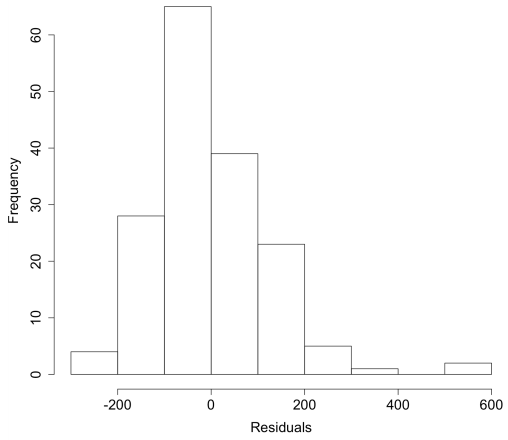


Figure 1: Histogram of the residuals approximately resembles a normal distribution.

Graphical assessment of residuals and their variance

Figure 2: Standard errors against fitted values.

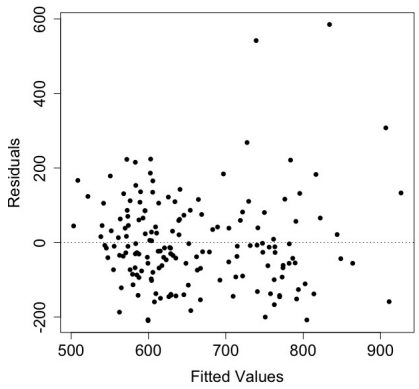
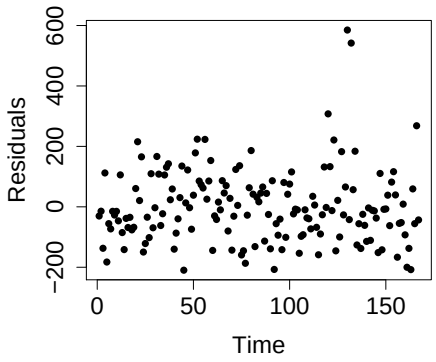


Figure 3: Time plot of the residuals.



Plots of residuals against each predictor

Figure 4: Residuals against mean red intensity.

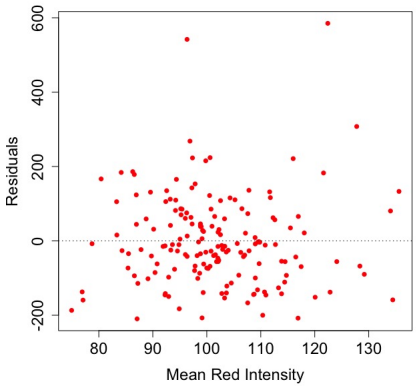


Figure 5: Residuals against mean blue intensity

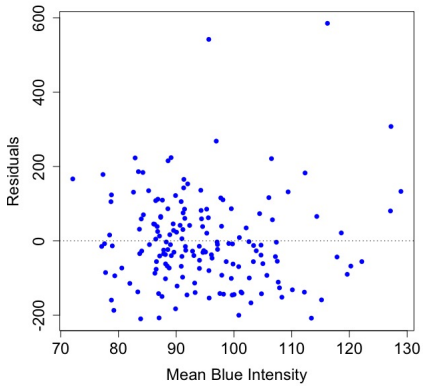


Figure 6: Residuals against mean green intensity.

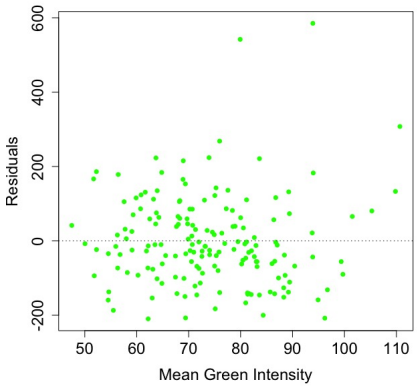


Figure 7: Residuals against mean HSV intensity.

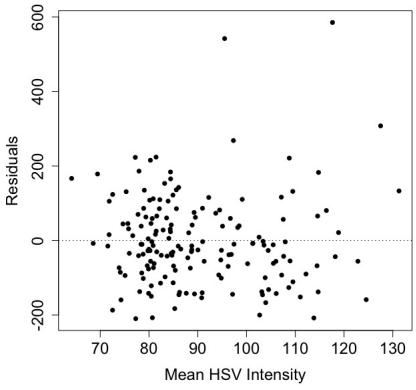


Figure 8: Residuals against mean Lab intensity.

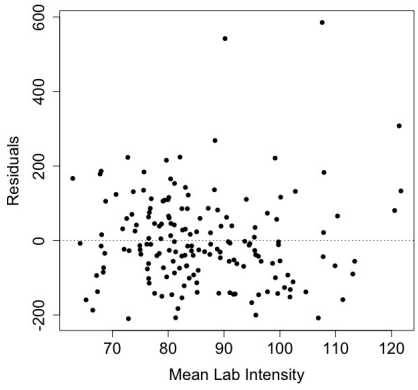


Figure 9: Residuals against mean HE intensity.

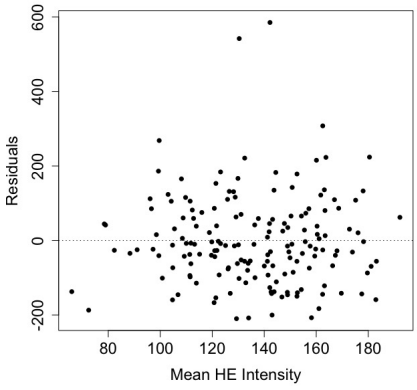
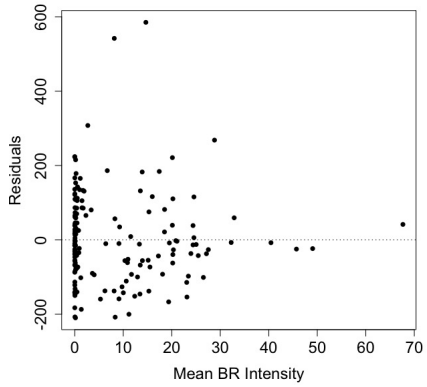


Figure 10: Residuals against mean BR intensity.



Evidence of violations of linear model assumptions

There are 2 extreme values correlating to nuclei with ID 130 and 132. A possible transformation would be to exclude these outlier values. These values severely skewed the residuals in the positive direction. In addition, since the summation of errors equal 0, we predict that the regression function is linear and the error terms appear to be independent. In the BR residual plot, we should note that many of the BR values are 0, and that the residuals are spread less the higher the BR intensity is. This may mean that we want to transform the BR variable in some way.

Key takeaways

Examining the residual plots, it looks like the multiple variable regression fits the model fairly well overall. The graphs of mean red, green, and blue intensity vs residuals are very similar to each other, suggesting that each one does equally well at predicting the response mean area. Additionally, we believe that each predictor works equally well at any frequency, so how well the model works doesn't change with frequency. The mean BR data was noticeably different than the other variables as the mean BR data was skewed more towards 0 than any other variable, which we will pay attention to upon further statistical analysis.

```r
cancer <- read.csv("/Users/leahleahleah/Downloads/cancer.csv")

Y <- cancer$Mean_Area
X1 <- cancer$Mean_mean_R
X2 <- cancer$Mean_mean_G
X3 <- cancer$Mean_mean_B
X4 <- cancer$Mean_mean_HSV
X5 <- cancer$Mean_mean_Lab
X6 <- cancer$Mean_mean_HE
X7 <- cancer$Mean_mean_BR

lm.canc <- lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7)

#extract b-values
b0 <- summary(lm.canc)$coef[1,1]
b1 <- summary(lm.canc)$coef[2,1]
b2 <- summary(lm.canc)$coef[3,1]
b3 <- summary(lm.canc)$coef[4,1]
b4 <- summary(lm.canc)$coef[5,1]
b5 <- summary(lm.canc)$coef[6,1]
b6 <- summary(lm.canc)$coef[7,1]
b7 <- summary(lm.canc)$coef[8,1]
```

```r
#extract residuals
e <- lm.canc$residuals
sum(abs(e))
sd(abs(e))
#extract fitted values
Y.hat <- lm.canc$fitted.values
#histogram of residuals
hist(e, xlab = "Residuals", ylab = "Frequency", cex.lab = 1.5,
     cex.axis = 1.5)

#plot of residuals vs fitted values
plot(Y.hat, e, pch = 16,
     xlab = "Fitted Values", ylab = "Residuals", cex.lab = 1.5,
     cex.axis = 1.5)
abline(h = 0, lty = 3)

#plot of residuals vs mean red intensity
plot(X1, e, pch = 16, col = 'red',
     xlab = "Mean Red Intensity", ylab = "Residuals", cex.lab =
     cex.axis = 1.5)
abline(h = 0, lty = 3)

#plot of residuals vs mean green intensity
plot(X2, e, pch = 16, col = 'green',
     xlab = "Mean Green Intensity", ylab = "Residuals", cex.lab
     cex.axis = 1.5)
abline(h = 0, lty = 3)
```

```r
#plot of residuals vs mean blue intensity
plot(X3, e, pch = 16, col = 'blue',
     xlab = "Mean Blue Intensity", ylab = "Residuals", cex.lab = 1.5,
     cex.axis = 1.5)
abline(h = 0, lty = 3)

#plot of residuals vs mean hsv intensity
plot(X4, e, pch = 16,
     xlab = "Mean HSV Intensity", ylab = "Residuals", cex.lab = 1.5,
     cex.axis = 1.5)
abline(h = 0, lty = 3)

##plot of residuals vs mean lab intensity
plot(X5, e, pch = 16,
     xlab = "Mean Lab Intensity", ylab = "Residuals", cex.lab = 1.5,
     cex.axis = 1.5)
abline(h = 0, lty = 3)

#plot of residuals vs mean he intensity
plot(X6, e, pch = 16,
     xlab = "Mean HE Intensity", ylab = "Residuals", cex.lab = 1.5,
     cex.axis = 1.5)
abline(h = 0, lty = 3)

#plot of residuals vs mean br intensity
plot(X7, e, pch = 16,
     xlab = "Mean BR Intensity", ylab = "Residuals", cex.lab = 1.5,
     cex.axis = 1.5)
abline(h = 0, lty = 3)
```