Group Project Report

Jonathan Thomas

2024-12-07

Jonathan Thomas

Parametric Linear Model

```
source <- read.csv("cervical-programme-annual-2022-23-csvs/kc61_sample_result_age_source.csv")
# misc. data wrangling
source$Severe_dyskaryosis <- gsub("not included", NA, source$Severe_dyskaryosis)
comparison_data <- source[,c("Borderline_changes", "Mild_dyskaryosis",
                             "Moderate_dyskaryosis","Severe_dyskaryosis")]
comparison_data$Severe_dyskaryosis <- as.numeric(comparison_data$Severe_dyskaryosis)

comparison_data <- comparison_data[order(
  comparison_data$Borderline_changes, decreasing = FALSE),]
```

Model creation and evaluation:

```
x <- comparison_data$Mild_dyskaryosis
y <- comparison_data$Moderate_dyskaryosis
z <- comparison_data$Severe_dyskaryosis

model_all <- lm(comparison_data$Borderline_changes ~
                comparison_data$Mild_dyskaryosis +
                comparison_data$Moderate_dyskaryosis +
                comparison_data$Severe_dyskaryosis)

#the following 2 models are identical, the first is included for code readability
# and the second to make the summary coefficients more intuitive
model_poly <- lm(comparison_data$Borderline_changes ~
                   poly(comparison_data$Mild_dyskaryosis, 2, raw=T) +
                   poly(comparison_data$Moderate_dyskaryosis, 2, raw=T) +
                   poly(comparison_data$Severe_dyskaryosis, 2, raw=T))
model_poly2 <- lm(comparison_data$Borderline_changes ~
                   x + I(x^2) + y + I(y^2) + z + I(z^2))


RSquaredValues <- NULL
RSquaredValues["Model:All"] <- summary(model_all)$r.squared
RSquaredValues["Model:Polynomial Order 2"] <- summary(model_poly2)$r.squared

RSquaredValues
```

```
##               Model:All Model:Polynomial Order 2
##               0.8671448                0.8968722
```

```r
summary(model_poly2)
```

```
##
## Call:
## lm(formula = comparison_data$Borderline_changes ~ x + I(x^2) +
##     y + I(y^2) + z + I(z^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6628.5  -388.7  -305.0   302.7  9810.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.196e+02  1.515e+02   2.110   0.0359 *
## x            1.793e+00  2.098e-01   8.544 1.71e-15 ***
## I(x^2)      -4.809e-05  6.983e-06  -6.887 5.26e-11 ***
## y           -9.649e+00  1.054e+00  -9.152  < 2e-16 ***
## I(y^2)       1.147e-03  1.526e-04   7.518 1.20e-12 ***
## z            5.378e+00  4.967e-01  10.826  < 2e-16 ***
## I(z^2)      -3.314e-04  4.907e-05  -6.755 1.13e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1807 on 233 degrees of freedom
##   (90 observations deleted due to missingness)
## Multiple R-squared:  0.8969, Adjusted R-squared:  0.8942
## F-statistic: 337.7 on 6 and 233 DF,  p-value: < 2.2e-16
```

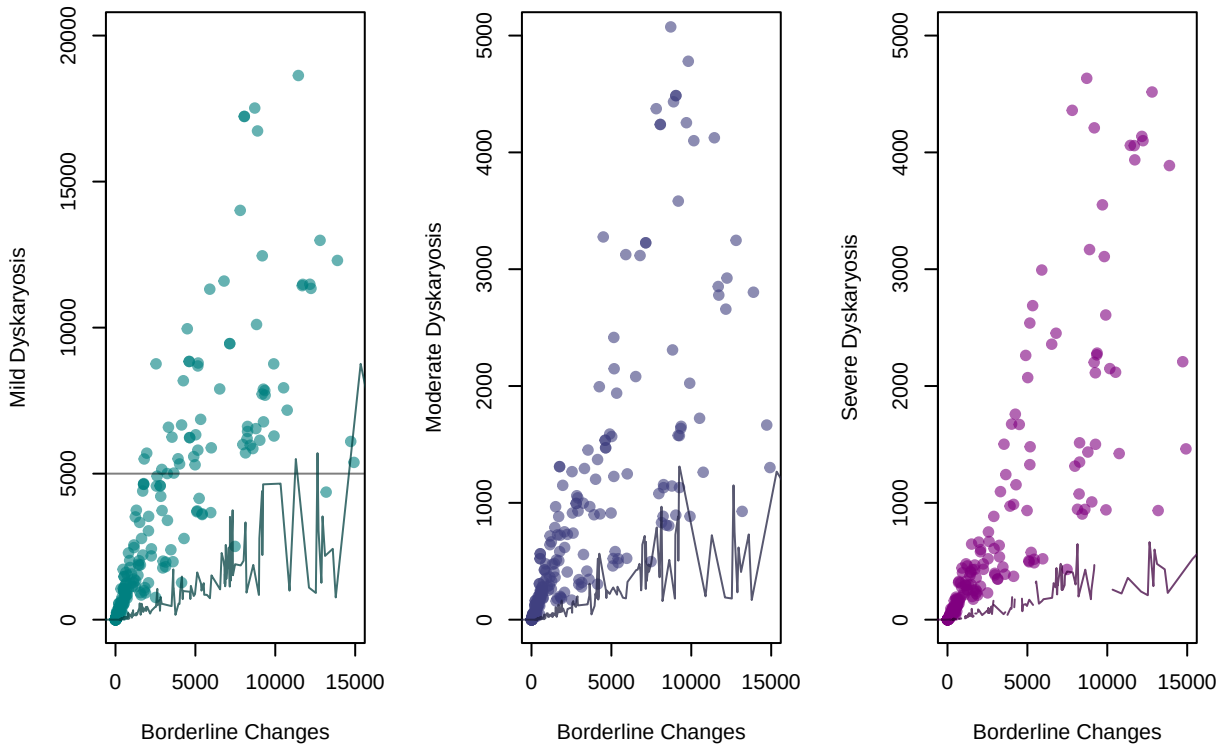
The R^2 values for the two models are 0.7831351 and 0.8968722. The parametric model accounted for 89% of the variation of the data analysed. Plotting the model along with the data is a good sanity check to see if the model is over fitting or properly modeling the data.

The coefficients all are significant, yet the only coefficients that have an absolute value greater than 1 are the non-squared terms. This means the model was mostly only using the value of the `Mild_dyskaryosis` (`x` in the summary), `Moderate_dyskaryosis` (`y`), and `Severe_dyskaryosis` (`z`).

```r
Go <- length(predict(model_poly2))
par(mfrow=c(1,3))
plot(comparison_data$Borderline_changes, comparison_data$Mild_dyskaryosis,
     xlim = c(0,15000), ylim = c(0,20000), col = rgb(0,0.5,0.5,0.6), pch = 19,
     xlab = "Borderline Changes", ylab = "Mild Dyskaryosis")
abline(h = 5000, col =rgb(0,0,0,0.5))
#this line is at the y value that is the top of the other two graphs
# to give a sense of size among the graphs
lines(sort(predict(model_poly2)), col = rgb(0,0.25,0.25,0.75),
      comparison_data$Mild_dyskaryosis[0:Go])

plot(comparison_data$Borderline_changes, comparison_data$Moderate_dyskaryosis,
     xlim = c(0,15000), ylim = c(0,5000), col = rgb(0.25,0.25,0.5,0.6), pch = 19,
     xlab = "Borderline Changes", ylab = "Moderate Dyskaryosis")
lines(sort(predict(model_poly2)), col = rgb(0.125,0.125,0.25,0.75),
      comparison_data$Moderate_dyskaryosis[0:Go])
```

```
plot(comparison_data$Borderline_changes, comparison_data$Severe_dyskaryosis,
     xlim = c(0,15000), ylim = c(0,5000), col = rgb(0.5,0,0.5,0.6), pch = 19,
     xlab = "Borderline Changes", ylab = "Severe Dyskaryosis")
lines(sort(predict(model_poly2)), col = rgb(0.25,0,0.25,0.75),
      comparison_data$Severe_dyskaryosis[0:Go])
```



It is clear from these graphs that the model is over fitting to the data. The model may be fitting accurately to data points with lower amounts of borderline changes while inaccurately fitting to the data points with higher amounts of borderline changes. The high density of data at the lower end would allow the model to keep a high average R^2 value while being inaccurate at the higher end of the graphs.

As the model is clearly inaccurate, by over fitting or by the data's incompatibility with a parametric model, the model should not be used in any case.