STA 223 Project 2: Wheat Kernel Type Modeling

Ana Boeriu

MS Biostatistics

UC DAVIS UNIVERSITY OF CALIFORNIA

Graduate Group in BIOSTATISTICS

Introduction

Biologically, the wheat kernel, also known as the wheat berry, is the seed from which a wheat plant grows

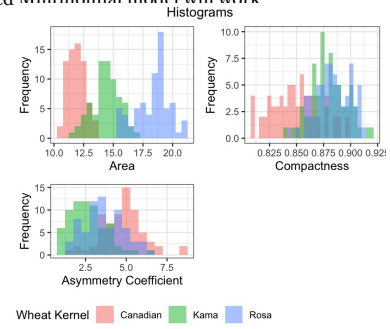
- The main goal of this project is to build a model and find the most significant variables when determining wheat kernel type.

Data Description and Preprocessing

Data: Seed Dataset

This dataset consists of the wheat seed type and the 7 geometric characteristics of the 210 digitized images of soft X-ray technique

- No missing values
- Removed 3 observations
- Response variable: wheat seed type (Kama, Rosa, Canadian)
  - 68 Canadian , 69 Kama , 70 Rosa
- 7 numeric predictors:
  - Length, width, area, perimeter, grove length, compactness,
  - Asymmetry coefficient
- Issue of collinearity between predictors.
  - Kept only area, compactness and asymmetry coefficient.
- Verified Multinomial model will work



- Conditioned on area value of 14.3 and created 2 separate datasets
- Resulted in having two independent logistic regression models.

Methods

Since tumor diagnosis is categorical, we will use logistic regression.

- Linear predictor: $\eta = X\beta$
- Random component: $nY \sim Bin(n, \pi)$
- Link function: $\eta = log\left(\frac{\mu}{1-\mu}\right)$

Before proceeding to model selection, we performed an overall regression test where we test the following:

- $H_o: \beta_p = 0$ versus $H_A: \beta_p \neq 0$ where p denotes the number of coefficients in the model

We proceed to model selection where we choose BIC as our criterion and forwards backwards as our stepwise selection

- BIC(p) = D(p) + p*log(n)

Next, we check for leverage and influential points.

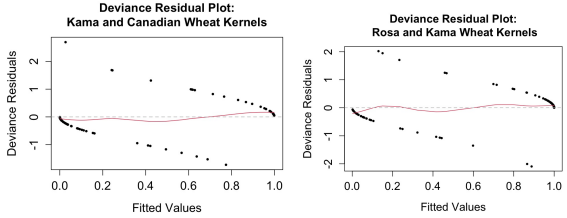
- If $h_{ii} > \frac{2p}{n}$, where n is the sample size, then the observation is a suspected leverage point.
- observations with high Cook's distance are influential points and may need to be deleted.

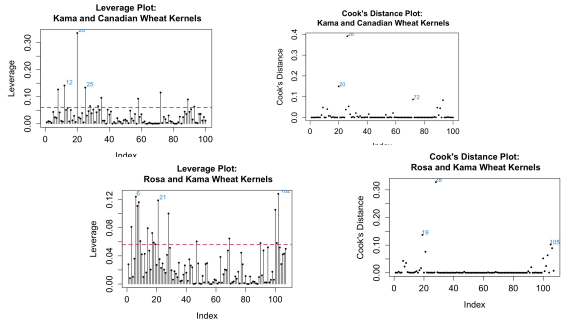
Overall Regression Test & Model Diagnostics

Overall Regression Test:

- Model with only Rosa and Kama wheat seeds
  - $H_o: \beta_{area \leq 14.3} = \beta_{asymmetry\ coefficient} = 0$
  - $H_a: at\ least\ one\ \beta_i \neq 0$
  - we conclude that some predictors have an overall significant effect with type of wheat
- Model with only Kama and Canadian wheat seeds
  - $H_o: \beta_{area > 14.3} = \beta_{asymmetry\ coefficient} = 0$
  - $H_a: at\ least\ one\ \beta_i \neq 0$
  - we conclude that some predictors have an overall significant effect with type of wheat

Model fits the data well and there is no obvious sign of lack of fit





- We found that even after removing these points and performing the same type of analysis as in part A there was no significant difference in our p-values. Thus, there is no need to remove these points

Results

$$ln\left(\frac{P(Y = Kama)}{P(Y = Canadian)}\right)$$

β_i	Estimate	P-value
Intercept	−38.724	1.79×10^{-4}
$Area_{\leq 14.3}$	3.409	9.05×10^{-5}
Asymmetry Coefficient	−1.4124	2.47×10^{-4}

$$ln\left(\frac{P(Y = Rosa)}{P(Y = Kama)}\right)$$

β_i	Estimate	P-value
Intercept	−40.018	7.29×10^{-6}
$Area_{>14.3}$	2.2479	1.94×10^{-5}
Asymmetry Coefficient	1.186	0.00447

- area is the most significant predictor and has a positive effect for both models.
- For small wheat kernels, the more irregular shape it has the more likely it will be Canadian

Acknowledgements

I would like to thank Professor Müller, Poorbita Kundu, and Han Chen for their support and guidance throughout this project

References

Data:
Lustosa, A. (2018, November 19). *Seed_from_UCI*. Kaggle. https://www.kaggle.com/dongeorge/seed-from-uci

Charytanowicz, M. (2018, January 1). *An evaluation of utilizing geometric features for wheat grain classification using X-ray images*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0168169917301990?via%3Dihub