Wheat Kernel Type Modeling

By: Ana Boeriu

UC Davis Graduate Group Biostatistics

1. Introduction

Biologically, the wheat kernel, also known as the wheat berry, is the seed from which a wheat plant grows (Simek, 2020). Many of the wheat kernels produced are not planted back into the ground but are ground into flour and made into tasty wheat products for consumers to enjoy around the world (National Festival of Breads, 2015). Morphological variation in seed characteristics included differences in seed shape and size. Seed shape, often related to germination rate, has agronomic importance because it reflects genetic, ecological, and physiological components that can directly affect yield, quantity, and market price (Cervantes, 2016). Larger seeds tend to be more self-sufficient, taking advantage of the rich soils while smaller seeds are more likely to be dispersed over a wide area helping at least some of the seedlings survive (Ray, 2018). Both the size and shape of a seed are important factors in consumer preference and post harvesting processing. In horticulture, the assessment of seed quality is an important factor of progress which must be given with the highest precision. Soft X-ray radiography is a fast and commonly used method of seed observation that prints the inside structure of one or several seeds without damaging its germinative properties (Chavagnat, n.d.-b). While this X-Ray technique cannot solve every horticulture problem, it gives horticulturists the ability to thoroughly examine the quality of the seeds as well as obtain new information that be the foundation for future scientific work.

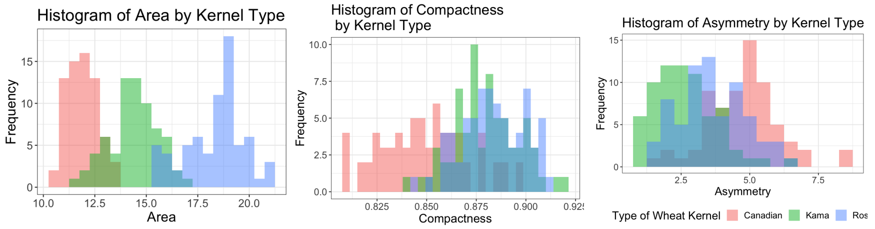
The main goal of this project is to build a model and find the most significant variables when determining wheat kernel type. We will then try and interpret our predictors to see how size and shape affect the type of wheat kernel.

The dataset we will use is the Wheat Seed dataset from Kaggle. This dataset consists of the type of wheat kernel and the seven geometric characteristics of wheat kernels computed from 210 digitized images of a soft X-ray technique. The size of the wheat kernel is described by the numeric predictors length, width, area, perimeter, and grove length while the shape of the wheat kernel is described by the numeric predictors compactness, and asymmetry. Small values for compactness correspond to an elongated kernel shape while large values for asymmetry correspond to irregular kernel shape (Charytanowicz, 2018). Our response variable is a trinary categorical variable that describes the type of wheat kernel as Kama, Canadian or Rosa.

2. Methodology

2.1 Preprocessing

Before we begin modeling our data, we will perform some exploratory data analysis, identify any problems in the data and explain how we addressed these problems. There were no missing values, no negative values and no unusual low or high values thus, there was no need to address these issues. We removed three observations that had identical measurements for all the predictors but were of different kernel types. This resulted in removing 1.4% of our data. There are 68 Canadian kernels, 69 Kama kernels and 70 Rosa wheat kernels where the categories are balanced. We then addressed the issue of collinearity between our predictor variables. This is an important step that prevents the reduced performance of a model by eliminating redundant predictors. The predictor "grove length" is highly correlated with all the other predictors having correlation values greater than 0.89. Similarly, "kernel length" and "kernel width" are highly correlated with "area" and "perimeter" having values greater than 0.95. This is not surprising since both are a measure of kernel size. Thus, we removed grove length, kernel length, kernel width and perimeter from the data since generally area is a better representation of size. To summarize we have kept the following predictors: area, compactness, and asymmetry. We also verified each predictor's distribution for each type of kernel to make sure that we can use a baseline odds multinomial regression model.



Notice that the predictor area has some slight overlap but not as much as the other two predictors. This is a concern because if we use a baseline odds multinomial regression model, there will be a lack of fit. To remedy this, we use a conditional logistic regression model where we condition on the predictor area. We want to choose a value for area between 14 and 15 that will give us the most balanced binary categories. We found that an area greater than 14.3 will give us a dataset that has 37 Kama wheat kernels and 70 Rosa wheat kernels while an area at most 14.3 gives us a dataset with 68 Canadian wheat kernels and 32 Kama wheat kernels.

2.2 Conditional Logistic Regression

By conditioning on area, we have created two seperate datasets where each dataset has only two wheat kernel types either Kama and Canadian or Kama and Rosa. As a result, since neither dataset depends on the other, we can proceed to create two independent logistic regression model fits. The three important components that allow us to relate the response variable to our predictors are the linear predictor, link function and systematic component. In logistic regression the linear predictor is $\eta = X\beta$, the random component is $nY \sim Bin(n,\pi)$, and the link function is $\eta = log\left(\frac{\mu}{1-\mu}\right)$ where $\mu = \frac{e^{\eta}}{1+e^{\eta}}$. We will also perform diagnostics, hypothesis test, as well as interpret our results.

2.3 Overall Regression effect

Before we begin model selection, we must first conduct the overall regression test for each model. We want to test the null that all slopes are zero against the alternative that at least one of the slopes is not zero. The test statistic is the difference between the null deviance and the residual deviance of the model and approximately follows a χ^2_2 distribution in each model. By rejecting the null, we indicate that the predictors indeed have an overall regression effect in determining wheat type. Once we established an overall regression effect, we may proceed to model selection where we want to obtain the most significant predictors in our model.

2.4 Model Selection

Compared to AIC, BIC tends to keep less predictors due to the larger penalty term related to model complexity. Since we are interested in inference, we will focus on using BIC criterion and forwards backwards stepwise selection. These criteria will give us a smaller model that is more interpretable.
We then perform diagnostics to make sure there are no abnormalities with our data. More specifically, we will be examining the goodness of fit and examining leverage and influential points.

2.5 Diagnostic Measures: Goodness of fit.

The deviance residual plot is used to check if the model fits the data well and the aim is to check if there are any systematic patterns left in the residuals. We plot the deviance residuals against fitted values using smoothing splines as visual aids, and check for systematic patterns which signify a lack of fit. After determining there is no obvious lack of fit, we then check for leverage and influential points.

2.6 Diagnostic Measures: Leverage/Influential points

To identify influential data points, we plot the leverage h_{ii} (diagonal of the hat matrix) against the index of the points. If $h_{ii} > \frac{2p}{n}$,where p is the number of coefficients in the model and n is the sample size, then the observation is a suspected leverage point. To detect outliers or influential observations, we will use Cooks distance plot. Points that have a high Cooks distance are suspiciously influential points and we may need to delete them if they greatly affect our model fit.

3. Results

3.1 Overall Regression Effect

For the model that has an area larger than 14.3 and either Kama or Rosa kernels, we want to test:
$H_o\colon \beta_{\text{area} > 14.3} = \beta_{\text{asymmetry}} = 0$ versus $H_a\colon at\ least\ one\ \beta_i \neq 0$ while for the model that has an area of at most 14.3 and either Canadian or Kama kernels we test: $H_o\colon \beta_{\text{area} \leq 14.3} = \beta_{\text{asymmetry}} = 0$ versus $H_a\colon at\ least\ one\ \beta_i \neq 0$

Each model has a p-value of 2.2×10^{-16}. Thus, we reject the null hypothesis and conclude that at least one β_i will have a significant effect with wheat kernel type in each model. This suggests that the predictors in each model have an overall effect in determining the type of wheat kernel

3.2 Final Models

After performing model selection, the final logistic regression model with main effects for an area at most 14.3, wheat kernel types of either Kama or Canadian, and using the smaller of the two kernels as the baseline is:

$$\boxed{ln\left(\frac{P(Y\ =\ Kama)}{\mathrm{P}(Y = Canadian)}\right) = ln\left(\frac{P(Y\ =\ Kama)}{\text{1- P}(Y = \text{Kama})}\right) = -38.72 + 3.409X_{\text{area}\leq 14.3} - 1.412X_{\text{asymmetry}}}$$

β_i	Estimate	P-value
Intercept	-38.72	1.79×10^{-4}
$Area_{\leq 14.3}$	3.409	9.05×10^{-5}
Asymmetry	-1.412	2.47×10^{-4}

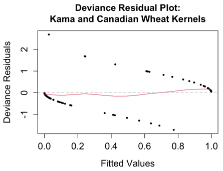
Notice that area is the most significant predictor followed by asymmetry when determining if the wheat kernel is either Kama or Canadian. Furthermore, area has a positive effect on determining kernel type while asymmetry has a negative effect in determining kernel type. As the area increases (ie gets close to but does not exceed 14.3) the probability of the wheat kernel being Kama increases while as asymmetry increases (kernel has an irregular shape), it is more likely that the wheat kernel is Canadian. The final logistic regression model with main effects for kernel types of either Rosa or Kama, area greater than 14.3, and using the smaller of the two kernels as the baseline is:

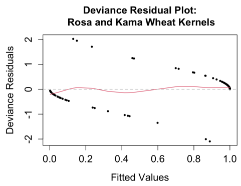
$$\boxed{ln\left(\frac{P(Y\ =\ \text{Rosa})}{\mathrm{P}(Y = \text{Kama})}\right) = ln\left(\frac{\mathrm{P}(Y\ =\ \text{Rosa})}{\text{1- P}(Y = \text{Rosa})}\right) = -40.018 + 2.248X_{\text{area}>14.3} + 1.186X_{\text{asymmetry}}}$$

β_i	Estimate	P-value
Intercept	-40.02	7.29×10^{-6}
$Area_{>14.3}$	2.248	1.94×10^{-5}
Asymmetry	1.186	0.00447

Notice that area is the most significant predictor followed by the asymmetry coefficient when determining the type of wheat kernel. Furthermore, these predictors all have a positive effect on determining if the wheat kernel is Rosa or Kama. In other words, as the area and asymmetry of a kernel increase, the probability of a Rosa kernel increases. This tells us that Rosa kernels tend to have irregular shapes as well as large areas, while Kama kernels tend to have smaller areas (larger than 14.3) and less irregular shape.

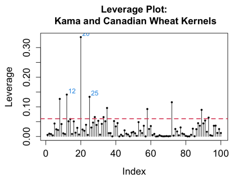
3.3 Goodness of fit

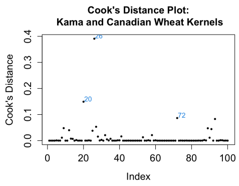


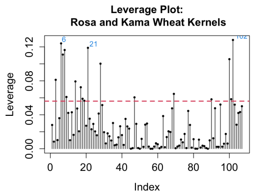


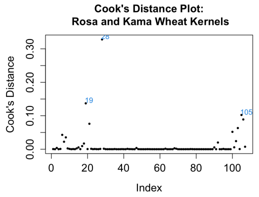
There is no systematic pattern thus we conclude that there is no obvious lack of fit for either model.

3.4 Outliers and influential Points









From the leverage plot, we observe several influential points greatly exceeding the threshold $\frac{2p}{n}$ denoted by the dotted red line. The first two columns show observations 12,20,25, are highly influential, while observations 20, 26,72, have a large cook's distance. Similarly, in columns three and four, observations 6,21,102 are highly influential while observations 28,19, 105 have large cook's distance. We verified if our two regression lines are affected when removing these points. We found that even after removing these points and performing the same type of analysis as in part 3.1 there was no significant difference in our p-values. Thus, we will not remove any more points. Generally, the next step would be to test for interaction terms, however there is no biological reason indicating that the effect of one of these predictors depends on another predictor. Recall area describes the size of a kernel while asymmetry describes the shape of a kernel. Wheat kernels with a small size does not imply that the kernel will also have irregular shape and vice versa. Thus, there is no need for an interaction term and our final model is shown in 3.2

4. Discussion

Kernel quality plays an important role in how a crop will perform. Larger kernels tend to be more robust and have a higher yield. Rose or Rosa wheat is one of the higher yielding wheat kernels during the winter season. (*Rose: A Winter Hardy Wheat*, 1993). When comparing Rosa and Kama wheat kernels we found that Rosa wheat kernels have and area greater than 14.3. Thus, it is not surprising since larger seeds increase germination which can increase yield as well (Chacon, 1998). Also, given the area of the kernel is larger than 14.3, kernels that have irregular shapes are more likely to be Rosa kernels. On the other hand, for kernels with areas at most 14.3, increasing the asymmetry coefficient, decreases the probability of the kernel being Kama. In other words, Canadian kernels have unusual shapes when compared to Kama kernels. This data analysis has shown the usefulness of the geometric predictors of area and asymmetry coefficient. For future studies we could expand this concept further by including more information on yield and the germination rate. This could give us an even better understanding in exploring which kernels are more likely to produce the highest yield and which kernels germinate the fastest.

References

Cervantes, E. (2016, April 13). *Updated Methods for Seed Shape Analysis*. Hindawi. https://www.hindawi.com/journals/scientifica/2016/5691825/

Chacon, P. (1998). *The effect of seed size on germination and seedling growth of Lauraceae in Chile*. Research Gate. https://www.researchgate.net/publication/297665318_The_effect_of_seed_size_on_germination_and_seedling_growth_of_Cryptocarya_alba_Lauraceae_in_Chile

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

Chavagnat, A. (n.d.). *USE OF SOFT X-RAY RADIOGRAPHY FOR STUDYING SEED QUALITY IN HORTICULTURE | International Society for Horticultural Science*. International Society for Horticultural Science. https://www.ishs.org/ishs-article/215_20

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

Müller, H. (2022). *GENERALIZED LINEAR MODELS Lecture Notes for BST/STA 223*. UC Davis Canvas Discovery. https://canvas.ucdavis.edu/courses/655751/files/?preview=15312915

National Festival of Breads. (2015, July 28). *Grain's Anatomy: What Makes a Kernel of Wheat*. https://nationalfestivalofbreads.com/hints-and-happenings/grain%E2%80%99s-anatomy-what-makes-a-kernel-of-wheat

Ray, C. C. (2018, January 26). *Seed Size Often Matters*. The New York Times. https://www.nytimes.com/2018/01/26/science/plants-seed-sizes.html

Rose: A Winter Hardy Wheat. (1993). SDSU Extension Fact Sheets. https://openprairie.sdstate.edu/cgi/viewcontent.cgi?article=1969&context=extension_fact

Simek, S. (2020). *Wheat Kernel - NDSU Wheat Quality & Carbohydrate Research, Dr. Senay Simsek*. NDSU. https://www.ndsu.edu/faculty/simsek/wheat/kernel.html

R appendix

Preprocessing

```
names(seed) = c("area", "perimeter",
"compactness", "length_kernal", "width_kernal",
              "asymCoef", "grove_length",
"kernalType")
seed = seed[,c(8,1:7)]
unique(seed$kernalType)
table(seed$kernalType)
seed[1:70,"kernalType"] = "Kama"
seed[71:140, "kernalType"] = "Rosa"
seed[141:210, "kernalType"] = "Canadian"
seed$kernalType = as.factor(seed$kernalType)
str(seed)
## check neg vals
checkNegVals = function(Cname){
  freq = length(which(seed$Cname < 0))
  return(freq)}
sapply(seed,checkNegVals)
### check repeats/duplicates rows
library(data.table)
which(duplicated(seed) == TRUE)
range(seed$compactness)
sapply(seed[2:8], range)
library(ggplot2)
library(corrplot)
ggplot(seed)+
  geom_point(aes(x=area, y=length_kernal))
ggplot(seed)+
  geom_point(aes(x=area, y=compactness))
ggplot(seed)+
  geom_point(aes(x=area, y=width_kernal))
ggplot(seed)+
  geom_point(aes(x=area, y=asymCoef))
ggplot(seed)+
  geom_point(aes(x=area, y=grove_length))
ggplot(seed)+
  geom_point(aes(x=asymCoef,
y=grove_length))
seed$eccentricity_index =
seed$length_kernal/seed$width_kernal
#correlationVals = cor(seed[,1:7])
correlationVals = cor(seed[,2:8])
corrplot(correlationVals, method = "number")
correlationVals = cor(seed2[,2:7])
corrplot(correlationVals, method = "number")
#take out length and width
which(names(seed) %in% c("width_kernal",
'length_kernal'))
seed2 = seed[,-which(names(seed) %in%
c("width_kernal", 'length_kernal'))]
seed2 = seed[, c(8,1:7)]
names(seed2)
head(seed2)
seed2[c(62, 200,210),]
#investigate
which(seed2$area == 11.23)
seed2[c(62,189,207),] #rm these pts
which(seed2$area == 12.76 | seed2$area ==
12.30)
seed2[c(200,210),]
#seed3 = seed2[-c(62,189,207,200,210),]
seed3 = seed2[-c(62,189,207),]
```

```
seed3 = seed3[,c(2,1,3:8)]
#reset index
rownames(seed3) = NULL
library(gridExtra)
library(ggpubr)
library(grid)
areaHist <- ggplot(seed3, aes(x=area,
fill=kernalType)) +
  geom_histogram(binwidth=.5, alpha=.5,
position="identity")+
  xlab("Area") + ylab("Frequency")+
  theme_bw()+theme(legend.position='none')
compactHist <- ggplot(seed3,
aes(x=compactness, fill=kernalType)) +
  geom_histogram(position="identity", alpha =
0.5)+
  xlab("Compactness") + ylab("Frequency")+
  theme_bw()+theme(legend.position='none')
asymCoefHist<- ggplot(seed3, aes(x=asymCoef,
fill=kernalType)) +
  geom_histogram(binwidth=.5, alpha=.5,
position="identity")+
  xlab("Asymmetry Coefficient") +
ylab("Frequency")+
  labs(fill = "Type of Wheat Kernel")+
  theme_bw()+theme(legend.position="bottom")
grid.arrange(areaHist,compactHist,
asymCoefHist, ncol = 2,
top=textGrob("Histograms"))
#split data
kamaCanada = subset(seed3, area <= 14.3)
table(kamaCanada$kernalType)
table(kamaCanada$kernalType)[1]/sum(table(kamaCanada$kernalType))
table(kamaCanada$kernalType)[2]/sum(table(kamaCanada$kernalType))
rownames(kamaCanada) = NULL
kamaRosaBG = subset(seed3, area >14.3)
table(kamaRosaBG$kernalType)
kamaRosaBG$area >14.3
kamaRosaBG$kernalType =
droplevels(kamaRosaBG$kernalType)
rownames(kamaRosaBG) = NULL
```

Regression Effect

```
blueGreenArea = glm(kernalType ~area +
compactness + asymCoef ,
                family = binomial(link = "logit"),
                data = kamaRosaBG)
emptyBG = glm(kernalType ~1 , family =
binomial(link = "logit"),
            data = kamaRosaBG)
anova(emptyBG,blueGreenArea, test = "Chisq")
pinkGreenArea = glm(kernalType ~area +
compactness + asymCoef , family =
binomial(link = "logit"), data = kamaCanada)
emptyPG = glm(kernalType ~1 , family =
binomial(link = "logit"), data = kamaCanada)
anova(emptyPG,pinkGreenArea, test = "Chisq")
```

Model Selection& Final Model

```
fbs_pg = step(emptyPG, scope = list(lower =
emptyPG, upper= pinkGreenArea),direction =
"both", k=log(nrow(kamaCanada)), trace =
FALSE)
summary(fbs_pg)
fbs_BG = step(emptyBG, scope = list(lower =
emptyBG, upper= blueGreenArea),
            direction = "both",
k=log(nrow(kamaRosaBG)), trace = FALSE)
```

```
finalGreenBlue = fbs_BG
summary(finalGreenBlue)
```

Goodness of Fit

```
par(mfrow = c(2,2))
res.D = residuals(fbs_pg, type="deviance")
plot(fbs_pg$fitted.values, res.D, pch=16,
cex=0.6, ylab='Deviance Residuals',
    xlab='Fitted Values',
    main = "Deviance Residual Plot:\n Kama and
Canadian Wheat Kernels ")
lines(smooth.spline(fbs_pg$fitted.values, res.D,
spar=1.9), col=2)
abline(h=0, lty=2, col='grey')
res.DBG = residuals(finalGreenBlue,
type="deviance") #or residuals(fit), by default
plot(finalGreenBlue$fitted.values, res.DBG,
pch=16, cex=0.6,
    ylab='Deviance Residuals',
    xlab='Fitted Values',
    main = "Deviance Residual Plot:\n Rosa and
Kama Wheat Kernels")
lines(smooth.spline(finalGreenBlue$fitted.value
s, res.DBG, spar=1.95), col=2)
abline(h=0, lty=2, col='grey')
```

Leverage and Influential Points

```
par(mfrow = c(2,4))
leveragePG = hatvalues(fbs_pg)
plot(names(leveragePG), leveragePG,
xlab="Index", type="h", ylab = "Leverage",
    main = "Leverage Plot:\n Kama and
Canadian Wheat Kernels")
points(names(leveragePG), leveragePG, pch=16,
cex=0.6)
susPtsPG <- as.numeric(names(sort(leveragePG,
decreasing=TRUE)[1:3]))
text(susPtsPG, leveragePG[susPtsPG],
susPtsPG, adj=c(-0.2,-0.3), cex=0.7, col=4)
p <- length(coef(fbs_pg))
n <- nrow(kamaCanada)
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPtsPG <- which(leveragePG>2*p/n)
cooksPG = cooks.distance(fbs_pg)
plot(cooksPG, ylab="Cook's Distance", pch=16,
cex=0.6,
    main = "Cook's Distance Plot:\n Kama and
Canadian Wheat Kernels")
susPtsPG <- as.numeric(names(sort(cooksPG,
decreasing=TRUE)[1:3]))
text(susPtsPG,cooksPG[susPtsPG] ,susPtsPG,
adj=c(-0.1,-0.1), cex=0.7, col=4)
# susPtsPG <-
as.numeric(names(sort(cooksPG[infPtsPG],
decreasing=TRUE)[1:3]))
# text(susPtsPG, cooksPG[susPtsPG], susPtsPG,
adj=c(-0.1,-0.1), cex=0.7, col=4)
###
leverageBG = hatvalues(finalGreenBlue)
plot(names(leverageBG), leverageBG,
xlab="Index", type="h",
    main = "Leverage Plot:\n Rosa and Kama
Wheat Kernels")
points(names(leverageBG), leverageBG,
pch=16, cex=0.6)
susPtsBG <-
as.numeric(names(sort(leverageBG,
decreasing=TRUE)[1:3]))
```

```
text(susPtsBG, leverageBG[susPtsBG],
susPtsBG, adj=c(-0.2,-0.3), cex=0.7, col=4)
p <- length(coef(finalGreenBlue))
n <- nrow(kamaRosaBG)
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPts <- which(leverageBG>2*p/n)
cooksBG = cooks.distance(finalGreenBlue)
plot(cooksBG, ylab="Cook's Distance", pch=16,
cex=0.6,
    main = "Cook's Distance Plot:\n Rosa and
Kama Wheat Kernels")
susPts <- as.numeric(names(sort(cooksBG,
decreasing=TRUE)[1:3]))
text(susPts, cooksBG[susPts], susPts, adj=c(-
0.1,-0.1), cex=0.7, col=4)
seed4 = kamaCanada[-c(12,20,25,26,72),]
a = glm(kernalType ~1 , family = binomial(link
= "logit"),
           data = seed4)
b = glm(formula = kernalType ~ area +
asymCoef, family = binomial(link = "logit"),
      data = seed4)
###reggression effect
anova(a,b, test = "Chisq")
seed5 = kamaRosaBG[-c(6,21,102,28,19,105),]
c=  glm(kernalType ~1 , family = binomial(link
= "logit"),
      data = seed5)
d = glm(formula = kernalType ~ area +
asymCoef, family = binomial(link = "logit"),
      data = seed5)
anova(c,d,test = "Chisq")
```