Finding the best parametrization for ice cream batter

1 Introduction

Manufacturing ice cream cones requires baking of a batter before rolling it into specifical conic shapes. This batter is composed of many ingredients (flour, sugar, water, *etc*), and its quality is a crucial point that was studied in [1]. The stake of this study was to optimize the automatization of the manufacturing. Indeed, different problems may emerge from too thin (sticking problem) or too thick (uniformity problem) batter.
To prevent this situation, they evaluated the quality of the batter used to prepare the cones depending on several factors. The factor we study is the viscosity of the flour, influenced by its content in moisture, protein and ash. We explore the data and develop a model to clarify the relationship between these elements.

2 Multiple regression analysis

2.1 Pairwise simple correlations

We first asses the inter-relation between the different predictor variables and look at their correlation. We also look at all pairwise scatter plots.
Figure 1 and Table 1 show that the strongest linear relation is between the protein and ash contents (|r|=0.86). This is likely due to the fact that proteins can capture inorganic materials (*i.e.* the ashes) easier than the moisture does: moisture is made of molecules of water and have no charges to interact with ashes, which are most of the time charged as are the proteins. Looking at the correlation coefficients of moisture with protein and ash (r=-0.57 and r=-0.6, respectively), we see that they are both negative, close to each other and not so strong.

	moisture	protein	ash
moisture	1.00	-0.57	-0.60
protein	-0.57	1.00	0.86
ash	-0.60	0.86	1.00

Table 1: Correlation values betweene pairs of variables

+VISCOSITY ???

2.2 Model selection

In order to be able to judge the quality of different models, we need to use some criteria of comparison like AIC (Akaike Criterion) or adjusted R-squared. Here we choose to use the AIC because the R squared always improves with model complexity whereas AIC allows us to achieve a good balance between model complexity and model fit. To choose the best model,

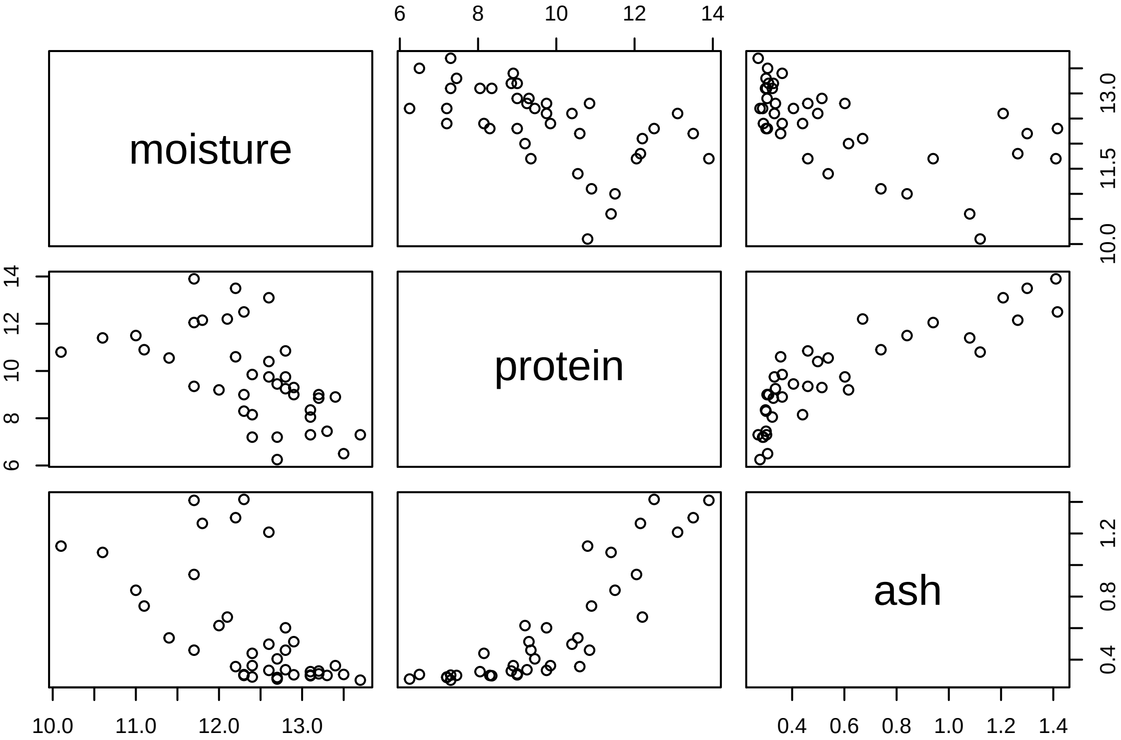

Figure 1: Pairwise scatter plot for all predictive variables

there are different methods that we can use. The most simple approach would be to compute all the different models and choose only that with the best AIC. This could easily be done in our case here, since the model is small, there would only be 9 different models to test, but it would become computationally demanding if the models were to scaled up. For this reason, we used and compared 3 different methods: backward, forward and stepwise selection.

2.2.1 Backward selection

Backward selection consists of computing the AIC for the full model, and remove the feature that has the largest AIC if its *p-value* is higher than a certain threshold (0.05 in our case). As we can see in Table 2b, with this method only 'moisture' is removed.

2.2.2 Forward selection

Forward selection on the other hand starts with the smallest sub-model that we tolerate and adds a parameter with lowest AIC if *p-value* < 0.05.

One of the shortcomings of this algorithm is that can sometimes stop short. Indeed, as we can see from Table 2a, with forward selection, our best model consits of only ash, and the variables protein and moisture are not selected, and so it has one fewer variable than the backward selection.

	Df	RSS	AIC	Pr(>F)
<none>	NA	34259.74	266.35	NA
moisture	1	30932.91	264.36	0.05
ash	1	25390.42	256.66	0.00
protein	1	33952.75	268.00	0.57

(a) Forward model selection

	Df	RSS	AIC	Pr(>F)
<none>	NA	10164.83	224.96	NA
moisture	1	10425.59	223.95	0.35
protein	1	25389.41	258.66	0.00
ash	1	30590.69	265.93	0.00

(b) Backward model selection

Table 2: Forward and backward model selection for the three different parameters (ash, protein and mositure). For both tables, the lower the RSS and the AIC, the better the model. The parameter is added to the model if the *p-valu*e ($Pr(> F)$) is inferior to 0.05.

2.2.3 Stepwise selection

Stepwise selection is a mix of both backward and forward selection: we start with any given model, and do a step of forward selection followed by one of backward selection until no more variable is added / eliminated. Here, as we can see in Table 3, the stepwise selection gives the same result as backward selection (*i.e.* without moisture), which is why we will keep this model.

Step	Resid. Dev	AIC
	10164.83	224.9620
- moisture	10425.59	223.9498

Table 3: Stepwise model selection table

3 Regression diagnostics

We have now selected a plausible best model for batter bakery. However, we have to ensure this model is robust and well-suited. We will compare the predicted value of the model with the real obtained values, as well as comparing some data with other possible models.

3.1 Q-Q plot

We display the Q-Q plot (Figure 2) to see if our data are normally distributed. Indeed, they should be close to the line y=x to infer a normal distribution. Unfortunately, this is not the case here, as one tail is far away from the line. The labeled points are those that deviate the most from the line and represent the extreme cases. Note that we work here with a small amount of data (*n=39*) and QQ-plots tend to stabilize with larger number of values. We shouldn't try to "over-interpret" abnormal phenomena as they may occur frequently with such a small data set.

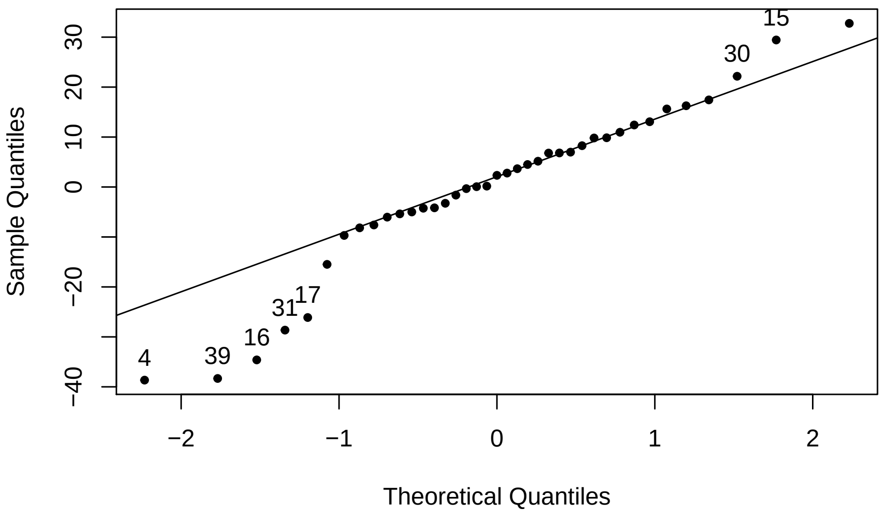

Figure 2: Normal Q-Q plot for the residuals from the lmbest

3.2 Observed vs Fitted

We need to compare the actual outputs of our chosen model with the observed values. This comparison is represented in Figure 3. The ideal data would follow the line y = x, meaning the modeled values are exactly equaled to the observed ones. This is not the case with our model, as many points are quite far from this ideal line. the model is therefore not optimal, as either we miss some parameters or some of them give wrong information. The next step to determine if our model could be improved is to look at the variance inflation factor (VIF).

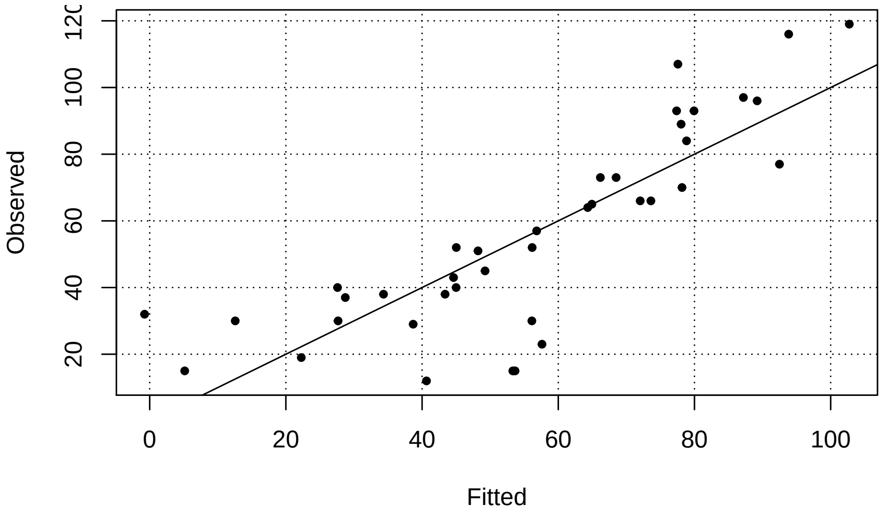

Figure 3: Observed vs fitted ratio

3.3 Variance inflation factor

Variance Inflation factor, or simply put VIF, will allow us to check for dependencies among our factors (multicolinearity). If this is the case, we will remove certain factors that would add redundancies in our model. A VIF above 5 is to be considered a caution and above 10 as a clear sign of multicolinearity. Let's use VIF on our best model:

	VIF
protein	3.75357
ash	3.75357

Table 4: VIF values

We can see that the values are even below 5, indicating no collinearity amoung these two factors. Hence, our model does not contain too many parameters, or at least dependant factors. Therefore, the VIF cannot really explain the difference between our model outputs and the observed data from figure 3. The most plausible answer (to be explored with more data!) could be that our model lacks some relevant variable(s).

4 Conclusion

We have developed a linear model to predict viscosity from ice cream cone batter characteristics. First, we explored the relationships between variables to see whether some are correlated. Next, we selected the best model to describe the batter viscosity as a function of the other variables in the data set. The best estimated model depends on only two variables: the ash and protein content. Finally, once the model is selected, the last step is to check if it is adequate or not. QQ-plot and comparison of fitted vs observed data showed mixed results, underlying that the model isn't perfect but remains solid. Supplementary verification with VIF confirmed that our model does not contain too many parameters, so in order to obtain an optimal model, we would maybe have to use more data to work with.

5 References

[1] V.T.Huanga, J.B.Lindamood1 & P.M.T.Hansen - Ice-cream cone baking: dependence of baking performance on flour and batter viscosity, 1988