|  |
| --- |
| **Southern Methodist University** |
| Group Project 3 |
| **Logistic Regression to model the probability of presence or absence of CRA7152 in apple juice** |

|  |
| --- |
| Andrew Walch, Nithya Doss, Sabitri KC  12-8-2017 |

**Introduction:**

Alicyclobacillus acidoterrestris (CRA7152) is a thermoacidophilic, non-pathogenic, spore forming bacterium which was the main cause of contamination in apple juice. This bacterium is heat resistant and hence it is essential to understand the effects of various other factors to prevent contamination.

The probability of presence or absence of CRA7152 in apple juice can be modeled as a function of multiple predictor variables. This study investigates the impact of pH, Brix, temperature and Nisin concentration on growth of CRA7152 in apple juice.

**Problem Statement:**

Develop a logistic regression model that can easily predict the Presence or Absence of growth of CRA7152 in apple juice based on the explanatory variables in the model. This model can be used in the food quality control and quality management, risk assessment and management and cost reducing by allowing the prediction.

**Data Set Description:**

This data is provided by Journal of Food Processing and Preservation, Vol. 35 and we acquired from University of Florida website <http://www.stat.ufl.edu/~winner/datasets.html>. The dataset contains presence/absence of growth of Alicyclobacillus Acidoterrestris CRA7152 in apple juice as a function of pH (3.5-5.5), Brix (i.e. Sugar content of an aqueous solution, 11-19), temperature (25-50 °C), and Nisin concentration (0-70). The data were downloaded as a .dat file with 74 rows and the following columns:

1.pH

2.Nisin concentration

3.Temperature

4.Brix concentration

5.Growth of CRA7152 (Binomial response with 1=presence and 0=absence)

**Data Dictionary:**

Table below illustrates the exploratory variables available in the data set.

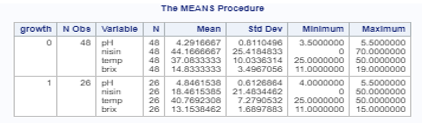
|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name in model** | **Variable Type** | **Data Type** | **Description** |
| pH | Discrete | Numeric (from 3.5 to 5.5) | Potential of Hydrogen is a scale of acidity. |
| Nisin | Discrete | Numeric (from 0 to 70) | Antibacterial peptide |
| Temp | Discrete | Numeric (from 25 to 50°C) | temperature |
| Brix | Discrete | Numeric (from 11 to 19) | Soluble solid concentration |
| Growth | Binomial Response | Binomial response with 1=presence and 0=absence | growth of Alicyclobacillus Acidoterrestris CRA7152 |

**Reasons to Conduct Logistic Regression:**

This data set has a binary response variable called **growth,** which is equal to 1 if there is a presence of CRA7152, and 0 otherwise. We want to check how certain explanatory variables may influence the likelihood of an event occurs and build a model which can predict if an event will occur or not. So, this dataset is a good candidate of logistic regression.

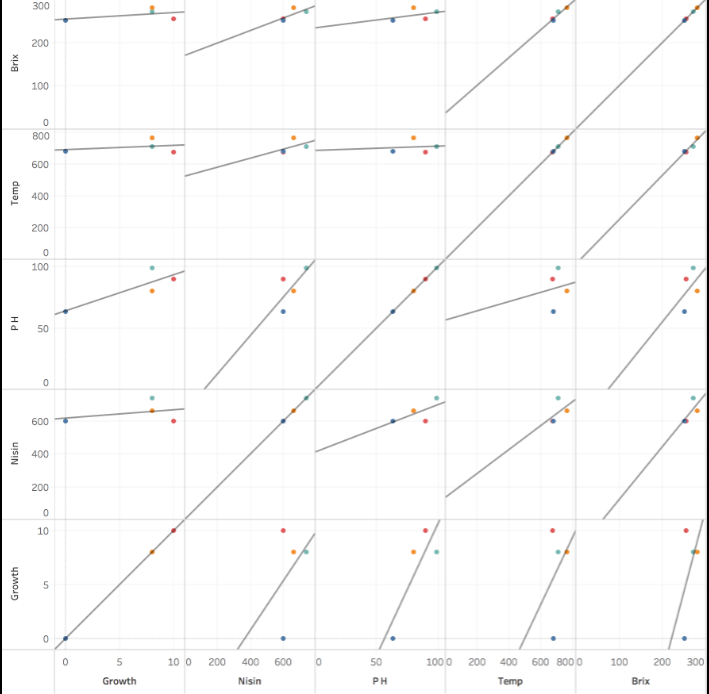
**Summary Statistics of the original variables (Exploratory Data Analysis):**

Predictor variables collected for analysis are pH, Nisin, temp and Brix. We will treat them as continuous variables. In SAS, dataset referred to “apple”, the response variable, growth, was coded as 1=presence and 0= absence. We start out by looking at some descriptive statistics. The basic summary is captured in **Table1**. We see that nisin has a minimum of 0 and max of 70 which will amount to higher variation as seen. The relationship of the explanatory variables to response is shown in **Figure 1** below. We don’t see any explicit multicollinearity. We assume that the observations are independent. We can study it further by running some more tests.



**Table 1 (proc means)**

A screenshot of a cell phone

Description generated with very high confidence 

**Figure 1 (corr)**

A screenshot of a cell phone

Description generated with very high confidenceA screenshot of a cell phone

Description generated with very high confidence

A screenshot of a cell phone

Description generated with very high confidence

2.a 2.b 2.c 2.d

**Table 2 (FREQ)**

The **Table 2** displays the frequency plot of the 4 explanatory variables Vs growth . Freq plot of brix and growth in 2.a , nisin and growth in 2.b, pH and growth in 2.c ,temp and growth in 2.d. along with their chi-square statistics. Each cell in the table contains cell frequencies, percentages of total frequency,percentages of row frequencys and percentages of column frequencies. chi-square pval <.05 indicates that there is strong association between each predictor variables and response variable growth except for temp ,since the pval is > 0.05 alpha level.

**Main effects model without Interaction:**

We run the logistic regression model including all 4 explanatory variables with no interactions. Model Fit Statistics in **Table 3** describes and tests the overall fit of the model. The value of AIC, SC and -2 Log L are low. The likelihood ratio chi-square statistic of 43.6147 with a p-value of 0.0001 indicates that our model fits significantly. The score and Wald tests are equivalent test of same hypothesis tested by the likelihood ratio test, so these tests also prove that the model is statistically significant.

The Analysis of Maximum Likelihood Estimates shows the coefficients, their standard errors, the Wald Chi-Square statistic and associated P-values. The coefficients give the change in the log odds of the growth for a one-unit increase in the predictor variables.

**Interpretation:**

* For every one-unit change in pH, the log odds of growth=1 (versus 0) increases by 1.89.
* For every one-unit change in Nisin, the log odds of growth=1 (versus 0) decreases by 0.0663
* For every one-unit increase in Temp, the log odds of growth=1 (versus 0) increase by 0.1104.
* For every one-unit change in Brix, the log odds of growth=1 (versus 0) decreases by 0.3117.

A screenshot of a cell phone

Description generated with very high confidence

**Table3 (MLE)**

A screenshot of a social media post

Description generated with very high confidenceA picture containing text, indoor, computer, map

Description generated with very high confidence

A screenshot of a cell phone

Description generated with very high confidence

**Table 4 – GOF Residuals and HL test (main effects)**

To assess the goodness of fit of the main effects model, we check residual diagnostics plot and Hosmer-Lemeshow test stats from **Table 4**. Since the pval= 0.37 indicating no evidence of poor fit. Looks like all variables have an effect on growth. We may try some interactions to fit the data to a better model.

**With Interaction:**

We ran a logistic regression with all the variables and their interactions.The most straight forward approach would be to use the forward selection model since we treat all our predictor variables as continuous. The Model fit statistics as shown in **Table 5** clearly indicates that AIC=48.86,SC = 67.30 and -2 Log L = 32.867 values are better in this model as compared to main effects model in **Table 3**(Model Fit statistics). The likelihood ratio chi-square of 63.0781 with a p-value of 0.0001 indicates that model is statistically significant.

Summary of Forward selection from **Table 6** showsthat interactions between temp\*temp, pH\*temp are significant and thus added to the final model. The variable brix and its interactions and interaction with nisin wont make it to the final model because they are less significant.

A screenshot of a cell phone

Description generated with high confidence

**A screenshot of a cell phone

Description generated with very high confidence**

**Table 5(MF Stats)**

**Table 6 – MLE (Forward Selection)**

The odds ratio of growth of CRA7152 relative to a 1-unit increase in nisin concentration is 0.88. The odds of a sample of apple juice having growth of CRA7152 is 0.88 times higher for a sample with one unit less nisin concentration. The logistic regression model models the log odds of a positive response (probability modeled is growth=1) as a linear combination the predictor variables.

This is written as

logit(p) = log(p/(1-p)) = β0 + β1\*pH - β2\*nisin + β3\*temp-β4pH\*temp-β5temp\*temp.

where p is the probability that growth=1.

**Our Final Model:**

Log[p/(1-p)] = –166.8 + 24.47 pH – 0.13 nisin + 4.42 temp – 0.3 pH \* temp –0.04 temp \* temp

* For one-unit change in predictor variable, the difference in log-odds for growth=1 is expected to

change by the respective coefficient, given the other variables in the model are held constant.

**Goodness of fit – Forward Selection:**

The hypothesis of the goodness-of-fit test as follows:

H0: No evidence of lack of fit

H1: The model does not fit the data

The insignificant p-value indicates the model fit is adequate and hence we fail to reject the null hypothesis where there is no evidence of lack of fit.

The interaction model with forward selection showed better adjustment to the data than the main effect model. This is evident from the HL test pval = 0.33 in **Table7**

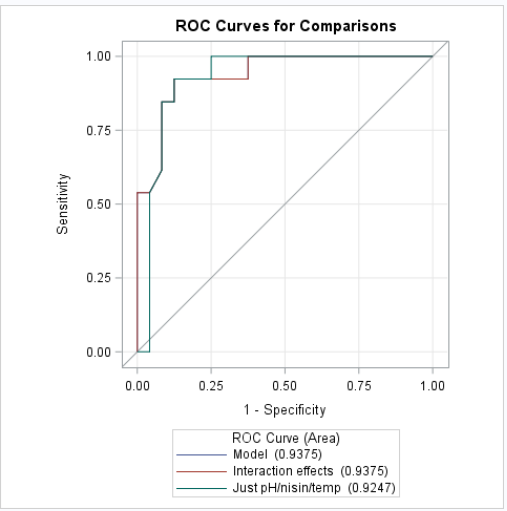
A screenshot of a cell phone

Description generated with very high confidence

**Table 7**

**Table7 – GOF Forward Selection**

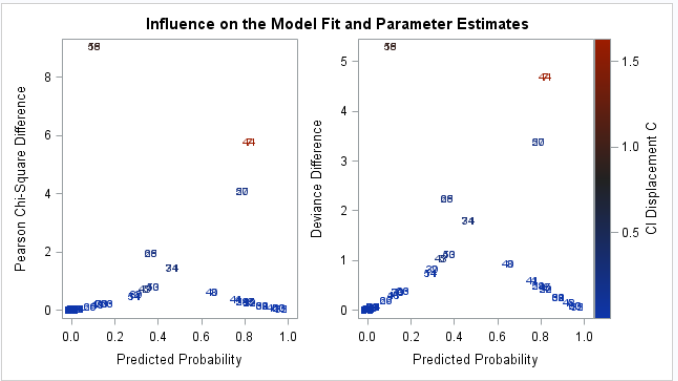
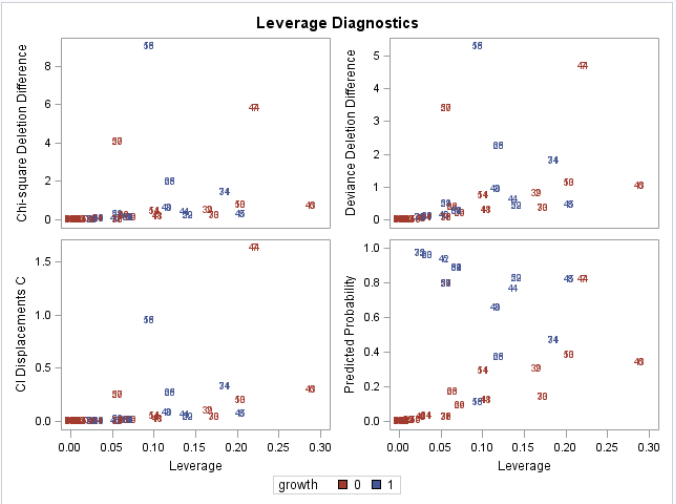
**Visualize Using ROC curves:**



**Figure 2**

From **Figure 2**, we see that the Interaction effects has the area under the ROC curve is 0.94 Vs the main effect with just the 3 explanatory variables pH,nisin,temp has an area of 0.92 which showed an improvement.

**Influence plots:**

**Figure 3**

There is actually one point 58 which has high leverage and influence as seen in the **Figure 3**. When we check the data point it is just a different response for similar concentration of nisin and brix. Since we don’t have enough information on how this data was collected, we will not remove this data point and test it further.

**Conclusion:**

This is an observational study specific to the data set which was provided by Journal of Food Processing and Preservation, Vol. 35. The study’s purpose is to help to develop a statistical model to predict the growth of CRA7152 in apple juice. With the given 4 variables by using data from randomly selected observations, the model is applicable to this particular study, and can’t be applied to other populations because of various situational and potentially confounding factors. We were unable to derive causality because of lack of information on randomization of the treatment.

In this study, we used logistic regression analysis to study the effect of the 4 exploratory variables that were used. Based on the results from the analysis we finalized the model. Other options were tried to consider one or two of the variables as categorical and running the regression model but since the data volume is low, we received a warning “complete separation or quasi-complete separation” which means prediction is infinitely certain with limited number of data. Hence we treated all the variables as continuous/regular variables and studied their interaction.

**References:**

1. <https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#logistic_toc.htm>
2. http://www.stat.ufl.edu/~winner/datasets.html

**Appendix:**

**data** apple;

infile "C:\Nithya\MSDS\Term2\6372\_Stats\_2\Project3\Apple\apple\_juice.csv" delimiter="," firstobs=**2**;

input pH nisin temp brix growth;

**run**;

**proc** **print** data = apple;

**run**;

**proc** **freq** data=apple;

tables pH\*growth nisin\*growth temp\*growth brix\*growth / chisq relrisk;

**run**;**quit**;

**proc** **means** data=apple;

class growth;

var pH nisin temp brix;

**run**;

\*plots(only)=roc(id=obs);

/\* main effects model\*/

**proc** **logistic** data=apple ;

/\*class pH nisin / param=ref;\*/

model growth(event='1')= pH nisin temp brix/ scale=none aggregate influence lackfit;

effectplot;

effectplot slicefit(sliceby=pH plotby=nisin) / noobs;

**run**;

/\* With interactions \*/

**proc** **logistic** data=apple ;

/\*class pH temp / param=ref; \*/

model growth(event='1')= pH nisin temp brix pH\*temp pH\*nisin pH\*brix temp\*temp/ selection=FORWARD start=**4** scale=none details influence lackfit;

effectplot;

effectplot slicefit(sliceby=pH plotby=temp) / noobs;

**run**;

/\* ROC curves \*/

**proc** **logistic** data=apple ;

model growth(event='1')= pH nisin temp pH\*temp temp\*temp / scale=none ctable pprob=**.05** aggregate lackfit;

ROC 'Interaction effects' pH nisin temp pH\*temp temp\*temp;

ROC 'Just pH/nisin/temp' pH nisin temp ;

roccontrast reference('Just pH/nisin/temp') / estimate e;

**run**;

/\* Influence plots \*/

**proc** **logistic** data=apple descending

plots(only label)=(leverage dpc);

model growth = pH nisin temp pH\*temp temp\*temp ;

output out=results p=prob l=lower u = upper

xbeta =logit stdxbeta=selogit /alpha=**.05**;

**run**;