On Rapid assessment methods using Statistical Modeling: Multiple Least Squares Regression vs. Logistic Regression

Fleisher JM[1,2]

[1] NOVA Southeastern University [2] NSF NIEHS Oceans & Human Health Center, University of Miami

ABSTRACT

TITLE:

On Rapid assessment methods using Statistical Modeling: Multiple Least Squares Regression vs. Logistic Regression

BACKGROUND:

There is a need to develop rapid assessment of bacterial water quality. To this end many statistical models have been published mostly using-environmental variables to predict concentrations of a particular FIO. The majority of these statistical models have used Multiple Least squares regression in which the major indicator of the goodness of fit of these models have largely depended on the R^2 value, which to date have been quite low. Since Beach management decisions have to be dichotomous in nature (Open/Close Beach) we explored the use of the Multiple logistic model in relation to the Multiple Lease Squares approach.

METHODS:

668 samples were utilized in this analysis. 10 major environmental variables and several FIO's were collected on each sample date. Both types of models were run on these data.

RESULTS:

Our Best Multiple Least Squares Regression was computed with a R Square value of 0.26, while the Multiple Logistic Regression Model yielded a maximum Sensitivity of 72.9% and a maximum Specificity of 65.9% at a cut point = 0.1. A backward selection routine was used in both the Logistic and Least Squares Model.

CONCLUSIONS:

Since the Logistic regression yields a much less nebulous goodness of fit statistic coupled with the fact that the Beach Managers decision is a dichotomous one, more attention should be paid to research using the Multiple Logistic Model.

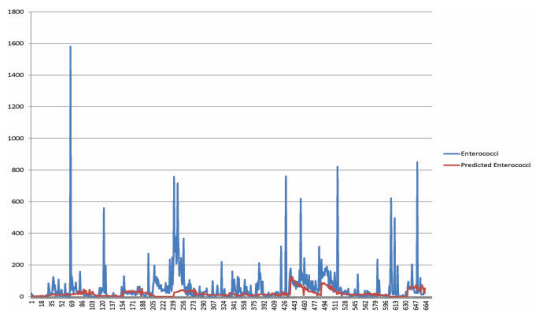
ENVIRONMENTAL VARIABLES

- pH
- Salinity Water
- Water Temperature
- Tidal Stage
- Turbidity
- Amount of Rainfall in the preceding 6 hours prior to sampling
- Amount of Rainfall in the proceeding 24 hours of sampling
- Wind Direction
- Wind Speed
- Solar Radiation

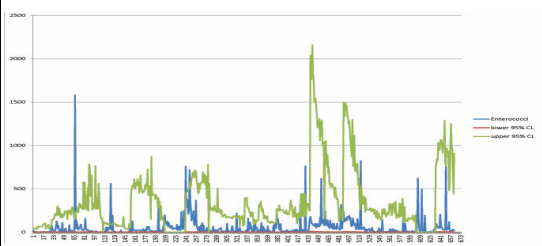
Results Least Squares Model

Variable	Estimate	Error	SS	F Value	Pr > F
•Intercept	-0.28651	0.36244	0.24457	0.62	0.4296
•Temperature	0.07583	0.01320	12.91371	32.99	<.0001
•Tide	1.05945	0.12993	26.02290	66.49	<.0001
•Rain 24 Hr Prior	-0.02254	0.00428	10.85001	27.72	<.0001
•Wind direction	0.00133	0.00035045	5.61027	14.33	0.0002
•Wind Speed	-0.11523	0.01280	31.70014	80.99	<.0001
•Solar radiation	-0.00108	0.00013634	24.74859	63.23	<.0001

Model R Square = 0.26



Actual Vs. Predicated Enterococci Densities



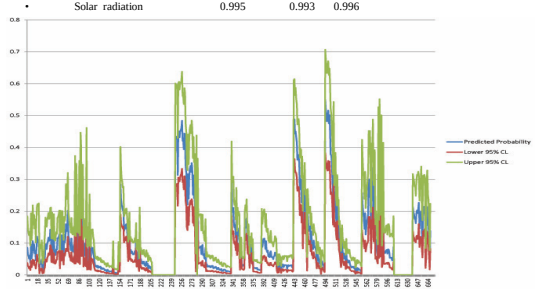
Actual Enterococci Densities Vs. Predicted 95% Confidence Limits

Results Logistic Regression Above or Below Single Sample Criteria

Parameter	DF	Standard Estimate	Wald Error	Chi-Square	Pr > ChiSq
Intercept	1	-19.1244	4.9439	14.9637	0.0001
Salinity	1	0.3526	0.1170	9.0784	0.0026
Temperature	1	0.1886	0.0663	8.1041	0.0044
Tide	1	2.8824	0.7645	14.2163	0.0002
Solar radiation	1	-0.00530	0.000860	38.0301	<.0001

Odds Ratio Estimates

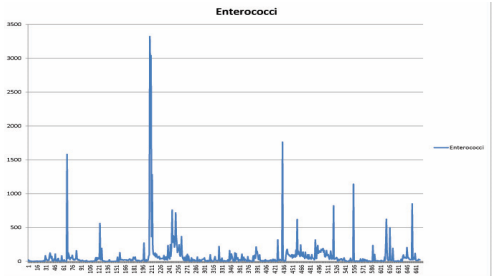
Effect	Point Estimate	95% Wald Confidence Limits	
Salinity	1.423	1.131	1.790
Temperature	1.208	1.061	1.375
Tide	17.858	3.991	79.903
Solar radiation	0.995	0.993	0.996



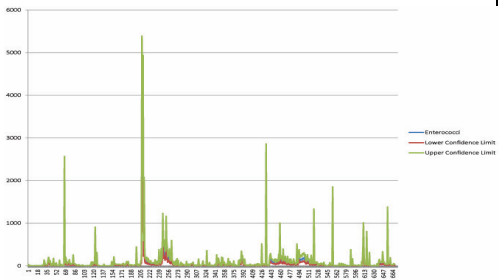
Predicated Probabilities with 95% Confidence Limits

Sensitivity of 72.9% and a maximum Specificity of 65.9% at a cut point = 0.1

Enterococci by Culture Only



Enterococci/100ml



Enterococci by Culture with 95% Confidence Levels

CONCLUSIONS

Since the Logistic regression yields a much less nebulous goodness of fit statistic coupled with the fact that the Beach Managers decision is a dichotomous one, more attention should be paid to research using the Multiple Logistic Model. It should also be noted that the precision of the Logistic Model seems better than both the Least Squares Model approach and the actual Culturing of Enterococci