Seawater salinity modeling using bivariate probit regression

by Aang Darmawan

Seawater salinity modeling using bivariate probit regression

Seawater salinity modeling using bivariate probit regression

Faisol[1], Tony Yulianto[1], Arsyiah[1], Sugiono[2], Achmad Basuki[3], Muhammad Agus Zainuddin[3]
[1]Mathematics Department, Mathematics and Natural Sciences Faculty, Universitas Islam Madura, Indonesia
[2]Fisheries Agribusiness Department, Faculty of Agriculture, Universitas Islam Madura, Indonesia
[3]Surabaya State Electronics Polytechnic, Indonesia

E-mail: faisol@uim.ac.id, toniyulianto65@gmail.com, arsiarsiyah05@gmail.com, quranyaiman@gmail.com, basuki@pens.ac.id , toniyulianto65@gmail.com

Abstract. Salt is one of the marine resources that is quite a lot needed as a supplementary food for the people of Indonesia. However, efforts to increase salt production have not been in demand, including in efforts to improve its quality, because many factors affect sea salt content or salinity, including the evaporation process, location and size of the sea, wind, air humidity and sea water temperature in this study are expected to produce the best salinity modeling by taking into account the factors that affect salinity. In this study, the method used was probit bivariate. The parameter estimation method used in the bivariate probit is the Maximum Likelihood Estimation (MLE). After the initial bivariate probit regression model is formed, then testing is carried out to determine the significance of each predictor variable to the response variable. After that the model that is formed identifies the criteria of goodness using the smallest Akaike Information Criterion (AIC) value of -9.03 so that the modeling results are good.

1. Introduction

Salt is one of the marine resources that is needed as a complementary food for the people of Indonesia. This is supported by the characteristics of the State of Indonesia which is a maritime country. However, efforts to increase salt production have not been in demand, including efforts to improve its quality. This is because the need for salt with good quality is mostly imported from abroad, especially in this case iodized salt and industrial salt. One of the islands of large salt production in Indonesia is Madura Island [1].

Iodized salt also has advantages, among others, that it can be useful for humans to prevent goiter, avoid miscarriage in pregnancy, increase IQ (Intelligence Quotient), and prevent stunting. Salt farmers in Indonesia do not yet have the ability (skills) to manage salt properly and according to international standards. For example, in Madura, which is known as the island of salt, there are still many farmers who produce krosok salt. This type of salt is of low quality because the production process is done traditionally and done quickly in order to produce more in quantity. The water content contained in krosok salt can reach 15%, whereas the maximum water content should be 2%. As a result, the salt is easy to melt and spoil prematurely. Krosok salt also does not contain iodine and is very dirty because it is not made in a proper place and has not passed laboratory tests so it is dangerous for human consumption [2]. One of the factors that affect the quality of salt in Madura depends on the salt

content. The levels that enter the crystallization table will affect the quality of the results. The quality of salt depends on the NaCl content of the salt, the NaCl content depends on the location where the seawater is taken.

There are many factors that affect seawater salinity or salinity varies in each place including the process of evaporation, rainfall, river water, sea location and size, ocean currents and wind [3]. Therefore, to determine the salinity of seawater, it is necessary to model seawater salinity using the bivariate probit regression method. Bivariate probit regression is a data analysis used to analyze response variables that are qualitative, quantitative or a combination of both [6]. However, in everyday life cases are often encountered, the response variable is a qualitative variable or a dummy variable by taking two or more possible values, such as decisions, such as the decision to choose "yes" or "no" [7].

2. Literature review

2. 1. Salt

Table salt or consumption salt is a food supplement that contains several chemical compounds, including sodium chloride (NaCl), $CaSO_4$, $MgSO_4$, $MgCl_2$. Salt can be obtained in three ways, namely evaporation of seawater by sunlight, mining of salt assistance (rock salt) and from seawater wells (brine). Mined salt has a different composition, it depends on the location of the mine, but generally contains more than 95% NaCl, and this is according to the following table.

Table 1. Composition of seawater at a salinity of 35 ppt.

No	Ion	Grams per kilogram of seawater
1	Cl^-	19.354
2	Na^+	10.77
3	K^+	0.399
4	Mg^{2+}	1.290
5	Ca^{2+}	0.4121
6	$SO_{4^{2-}}$	2.712
7	B_{1^-}	0.0673
8	F^-	0.0013
9	B	0.0045
10	Sr^{2+}	0.0079
11	$10_{3^-}, I^-$	$6{,}0 \times 10^{-5}$

NaCl as the main content of salt can be used as a parameter of the quality of a salt. In conventional salt processing, known as the people's salting process, the resulting salt production only uses the total evaporation method. If it is related to the level of NaCl as the main component of the desired salt, the NaCl produced from standard seawater is 27.393 g/kg seawater with a salinity of 35 ppt, or in other words the NaCl produced is only 78.266% (without taking into account the water content), means that it does not meet the desired category, namely quality I and II. Salt quality can be classified based on its water content. Based on this, it can be said that the salt produced only meets the quality of the third class [3].

2. 2. Descriptive Statistics

Statistical methods are procedures used in collecting, presenting, analyzing, and interpreting data. While descriptive statistics is a method related to the collection and presentation of a set of data so that

it can provide useful information. The information provided by descriptive statistics includes measuring data centering, measuring data distribution, and creating and displaying tables, graphs and diagrams. Measures of data concentration (mean) and measures of data spread (variance) are tools that can be used to define numerical measures that describe the characteristics of the data.

If the data size $x_1, x_2, \ldots x_n$ arranges a population so that the size is N, then the mean (μ) d of the population is $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, and variance defined as $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$.

2. 3. *Opportunity Distribution*

In many applications, recognizing certain characteristics will make it possible to find out a distribution that has a special shape, depending on one or more parameters and the numerical value of each parameter applied.

2. 4. *Multinomial Distribution*

A multinomial experiment is a binomial experiment in which each trial yields more than two possible outcomes. For $E_1, E_2, \ldots, E_k$ events and given $P = P(E_i)$ for $i = 1,2 \ldots, k$ mutually independent experiments so that what is given is the number of events that have occurred in a multinomial distribution.

$$f(y_1, \ldots, y_k) = \frac{n!}{y_1! \ldots y_k! \left(n - \sum_{i=1}^{k} y_k\right)!} p_1^{y_1} \ldots p_k^{y_k} \left(1 - \sum_{i=1}^{k} p_i\right)^{n-\sum_{i=1}^{k} y_i}$$

The mean for the multinomial distribution is $\mu = np_i$ and the variance $var(y_i) = np_i(1 - p_i)$

2. 5. *Normal Distribution*

Normal distribution is the most important continuous probability distribution and is often used in statistics. If X is a normal random variable with mean μ and variance σ^2 then probability density function of $X \sim N(\mu, \sigma^2)$ is $f(x, \mu, \sigma) \frac{1}{\sqrt{2\pi}\sigma} exp\left[\frac{1}{2}\left(\frac{x\mu}{\sigma}\right)^2\right]$ for $\infty < x$, with $\infty < x < \infty$ and $\infty < \sigma < \infty$.

2. 6. *Multicollinearity*

Regression modeling requires a requirement that there is no correlation between the predictor variables used. Multicollinearity is a condition where there is a linear relationship or a high correlation between the predictor variables in the regression model. If the magnitude of the correlation exceeds 0.8 or 80%, the pairwise correlation between the two predictor variables is said to be high. The calculation of the correlation coefficient between the two predictor variables is as follows:

$$r_{x_i x_j} = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} x_j\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{j=1}^{n} x_j^2 - \left(\sum_{j=1}^{n} x_j\right)^2}}$$

To find out the correlation between categorical variables, the dependent test was used. The dependent test requirement was that there was no frequency value of a cell that was less than one and if there was a cell with a frequency of less than five then it should not exceed 20% of the total number of cells.

2. 7. *Regression Analysis*

Regression analysis is concerned with the study of the dependence of one response variable on one or more predictors, with the aim of estimating the population mean value of the response variable from the known values of the predictor variables.

Given $z_1, z_2, \ldots z_r$ is a predictor variable as many as r which has a relationship to a response variable y, a linear regression model equation with one response variable is given as

$Y = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r + \varepsilon$, with $\beta_0, \beta_1, \ldots \beta_r$ is the coefficient of the predictor variable, $z_1, z_2, \ldots, z_r$ as the parameters that need to be found. While (is an error whose distribution is known or assumed.

Generally, regression models involve response variables that are quantitative in nature and the predictor variables can be quantitative or qualitative or a combination of both. However, a response variable can also be qualitative, such as choosing a "yes" or "no" decision so that a probit regression model is also needed [4].

2. 8. Bivariate Probit Regression Model

Bivariate probit regression model is a probit regression model using two response variables in the form of binary data category while the predictor variables are discrete and continuous variables. The assumption that must be fulfilled in the bivariate probit model is that the response variables have a relationship (dependent).

Suppose there are two response variables Y_1 and Y_2 it is assumed to come from the variable y_1^* and y_2^* where [5]

$$\boldsymbol{y_1^* = \beta_1^T x + \varepsilon_1}$$

$$\boldsymbol{y_2^* = \beta_1^T x + \varepsilon_2}$$

with $\beta_1 = [\boldsymbol{\beta_{10}\ \beta_{11}\ \beta_{12} \ldots \beta_{1q}}]^T$, $\beta_2 = [\boldsymbol{\beta_{20}\ \beta_{22}\ \beta_{12} \ldots \beta_{2Q}}]^T$, $x = [\boldsymbol{1\ x_1\ x_2 \ldots x_q}]^T$

Based on the above assumptions, the two response variables are normally distributed so that they can be denoted as $\boldsymbol{Y_1^* \sim N(\beta_1^T x, 1)}$ and $\boldsymbol{Y_2^* \sim N(\beta_1^T x, 1)}$. The formation of response variable categories in the bivariate probit model is the same as the univariate probit model, namely by determining the threshold for each unobserved response variable. The categorization can be as follows:

Assuming the threshold γ for model $\boldsymbol{y_1^* = \beta_1^T x + \varepsilon_1}$, so that the categorization is obtained as follows:

$Y_1 = 0$ if $y_1^* \leq \gamma$ and

$Y_1 = 1$ if $y_1^* > \gamma$

Assuming the threshold δ for model $\boldsymbol{y_2^* = \beta_1^T x + \varepsilon_2}$ so that the categorization is obtained as follows:

$Y_2 = 0$ if $y_2^* \leq \delta$ and

$Y_2 = 1$ if $y_2^* > \delta$

Because there is more than one random variable, namely and each of which is normally distributed, it produces a bivariate normal distribution. Probability density function of the Bivariate Normal distribution is as follows:

$$f(y_1^*, y_2^*) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(\frac{1}{2}\begin{bmatrix} y_1^* & \beta_1^T x \end{bmatrix}^T \Sigma^1 \begin{bmatrix} y_1^* & \beta_1^T x \\ y_2^* & \beta_2^T x \end{bmatrix}\right)$$

Where $\Sigma = \begin{bmatrix} \boldsymbol{var(y_1^*)} & \boldsymbol{cov(y_1^*, y_2^*)} \\ \boldsymbol{cov(y_2^*, y_1^*)} & \boldsymbol{var(y_2^*)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\sigma_1^2} & \boldsymbol{\sigma_{12}} \\ \boldsymbol{\sigma_{12}} & \boldsymbol{\sigma_2^2} \end{bmatrix} \begin{bmatrix} \boldsymbol{1} & \boldsymbol{\sigma_{12}} \\ \boldsymbol{\sigma_{21}} & \boldsymbol{1} \end{bmatrix} \begin{bmatrix} \boldsymbol{1} & \boldsymbol{\rho_{12}} \\ \boldsymbol{\rho_{21}} & \boldsymbol{1} \end{bmatrix}$

The probability density function of the Bivariate Standard Normal is as follows:

$$\phi(z_1, z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left(-\frac{1}{2}\frac{1}{(1-\rho^2)}\left(\boldsymbol{z_1^2 - 2\rho z_1 z_2 + z_2^2}\right)\right)$$

2. 9. Bivariate probit model parameter estimation

To form a bivariate probit model, parameter estimation was carried out using the Maximum Likelihood Estimation (MLE) method.

To estimate the parameters in probit regression, the Maximum Likelihood Estimation (MLE) method is used because the distribution of the probit model is known [8]. For example $Y_1, Y_2, \ldots, Y_n$ is a random variable from the population with probability density function $f(y, \theta)$ where θ is an unknown parameter, then the likelihood function of the random variable is

$$L(\theta) = f(y_1, y_2, \ldots, y_n; \theta) = \prod_{i=1}^{n} f(y_n; \theta)$$

If given $f(y_1, y_2, \ldots, y_n; \theta), \theta \in \Omega$ is a shared probability density function of $y_1, y_2, \ldots, y_n$, then the value $\bar{\theta}$ at Ω is the maximum value at $L(\theta)$ which is called the MLE of θ. So $\bar{\theta}$ is the value of θ defined:

$$f(y_i; \bar{\theta}) = \max_{\theta \in \Omega} f(y_1, y_2, \ldots, y_n; \theta)$$

The likelihood function for the univariate binary probit regression model with Bernouli distribution is:

$$L(\beta) = \prod_{i=1}^{n} \left(p(x_i)\right)^{y_i} \left(q(x_i)\right)^{1-y_i}$$

After obtaining the likelihood function, then the *ln* transformation of the function is obtained so that the following equation is obtained::

$$\begin{aligned} \ln L(\beta) &= \ln\left(\prod_{i=1}^{n} \left(p(x_i)\right)^{y_i} \left(1 - p(x_i)\right)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} (y_i \ln(px_i)) + (1 - y_i) \ln\left(1 - p(x_i)\right) \\ &= \sum_{i=1}^{n} \left(y_i \ln\left(1 - \phi(r - \beta^T x)\right) + (1 - y_i) \ln\left(\phi(r - \beta^T x)\right)\right) \end{aligned}$$

To maximize the likelihood function, it is necessary to reduce the *ln* likelihood function for parameter β and then equate it with zero to get an estimate of the parameter β. The first derivative for the *ln* likelihood function is as follows:

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \frac{\partial}{\partial \beta}\left(\sum_{i=1}^{n} \left(y_i \ln\left(1 - \phi(r - \beta^T x)\right) + (1 - y_i) \ln\left(\phi(r - \beta^T)\right)\right)\right)$$

$$= \sum_{i=1}^{n} x_i \phi(r - \beta^T x) \left(\frac{y_i}{1 - \phi(r - \beta^T x)}\right) - x_i \phi(r - \beta^T x) \left(\frac{1 - y_i}{\phi(r - \beta^T x)}\right)$$

So get

$$\sum_{i=1}^{n} x_i \phi(r - \beta^T) \left(\frac{y_i}{1 - \phi(r - \beta^T x)}\right) - x_i \phi(r - \beta^T x) \left(\frac{1 - y_i}{\phi(r - \beta^T x)}\right) = 0$$

From the MLE method that has been carried out, it is found that the form of the equation is not closed form so it is necessary to use a numerical method using Newton Raphson iterations, namely: $\beta^{(m)} = \beta^{(m-1)} - \left[H\left(\beta^{(m-1)}\right)\right]^{-1} g\left(\beta^{(m-1)}\right)$

2. 10. Significant Testing of Probit Model Parameters

The significant test of the parameters of the bivariate probit model was carried out with two tests, namely the simultaneous test and the partial test. Simultaneous testing was conducted to test whether

the predictor variables had a significant effect on the variables Y_1 and Y_2. Meanwhile, partial testing was carried out to test whether each parameter had a significant effect on the response variable Y_1 and Y_2 [9].

3. Main Results

3.1. Analysis of Seawater Salinity Data and Factors Affecting Salinity

Based on the data that has been obtained for the response variables according to the existing categories, the frequency table for each of the salinity of the south coast seawater and the salinity of the north coast sea water.

Table 2. Descriptive table of seawater salinity.

Category	North coast salinity	South coast salinity
Low salinity	4	7
High salinity	26	23
Total	30	30

Table 3. Table of salinity contingency of south and north coast seawater.

North coast salinity	South coast salinity		Total
	Low	High	
Low	3	1	4
High	4	22	26
Total	7	23	30

In this research, a bivariate probit regression model is applied using case data of seawater salinity and the factors that influence seawater salinity on the south coast and north coast. The research variables used are two response variables, namely seawater salinity on the south coast of Pamekasan Regency (Y_1) and seawater salinity on the north coast of Sampang Regency (Y_2) as well as three predictor variables, namely wind (X_1) humidity (X_2) sea water temperature (X_3)

3.2. Bivariate Probit Regression Modeling

In this step, a bivariate probit regression model will be formed for large case data of seawater salinity values and factors that affect seawater salinity on the south coast of Pamekasan Regency and the north coast of Sampang Regency in July 2019 to May 2020. To model in a regression, it is necessary to test the correlation between variables to see whether there is a correlation between these variables.

3.3. Relationship between Variables

Bivariate probit regression model requires a correlation between two response variables so it is necessary to test the correlation between the variables used. Based on Table 4.3, it is known that there is no frequency value less than one and cells containing frequencies less than five do not exceed 20% of the total cells so that they are eligible for the dependency test.

The test for the two response variables is as follows:

Hypothesis:

H_0: Variable y_i and y_j independent

H_1: Variable y_i and y_j dependent

Test Statistics:

$X^2_{count} = \sum_{i=0}^{I} \sum_{j=0}^{J} \frac{(0_{ij} - E_{ij})^2}{E_{ij}}$

Where 0_{ij}: frequency value for each cell, and E_{ij}: estimated value for each cell

Reject H_0 if $X^2_{count} > X^2_{0.05,1}$

Based on table 3, the values 0_{ij} and E_{ij} respectively are as follows:

$0_{00} = 3$; $E_{00} = \frac{4 \times 7}{30} = 0.9$

$0_{01} = 1$; $E_{01} = \frac{4 \times 23}{30} = 3.1$

$0_{10} = 4$; $E_{10} = \frac{26 \times 7}{30} = 6.1$

$0_{11} = 22$; $E_{11} = \frac{26 \times 23}{30} = 19.9$

So the value $X^2_{count} = \frac{(3-0.9)^2}{0.9} + \frac{(1-3.1)^2}{3.1} + \frac{(4-6.1)^2}{6.1} + \frac{(22-19.9)^2}{19.9} = 7,26714$

Because value $X^2_{count} > X^2_{0.05,1}$=3.84 so reject H_0. It can be concluded that the response variables are not independent or there is a relationship between the response variable salinity of sea water south coast and north.

3.4. Bivariate Probit Parameter Estimation

To estimate bivariate probit parameters, the Maximum Likelihood Estimation (MLE) method can be used. First established contingency tables (2 x 2) involving the response variable Y_1 and Y_2 then determined the probability of to the cell.

Table 4. Contingency table of variable frequency Y_1 and Y_2.

Y_1	Y_2	
	0	1
0	Y_{00}	Y_{01}
1	Y_{10}	Y_{11}

Table 5. Contingency table of variable probability Y_1 and Y_2.

Y_1	Y_2		Total
	0	1	
0	$p_{00}(x)$	$p_{01}(x)$	$p_{0.}(x) = 1 - p_1(x)$
1	$p_{10}(x)$	$p_{11}(x)$	$p_{0.}(x) = p_1(x)$
Total	$p_{.0}(x) = 1 - p_2(x)$	$p_{.1}(x) = p_2(x)$	$p_{..} = 1$

In accordance with the contingency table (2 x 2) formed in table 4.6 and table 4.7, the response variable will have a multinomial distribution with the notation $Y \sim MULTI\,(1; p_{11}, p_{10}, p_{01})$ with $p_{00} = 1 - p_{11} - p_{10} - p_{01}$.

If it is known that the function of the bivariate probit regression is:

$$y_1^* = \beta_1^* x + \varepsilon_1$$
$$y_2^* = \beta_1^* x + \varepsilon_2$$

Where ε_1 and ε_2 assumed standard normal distribution with $\mu = 0$ and $\sigma^2 = 1$ and category formation with thresholds for each response variable, namely:

$Y_1 = 0$ if $y_1^* \leq r$

$Y_1 = 1$ if $y_1^* > r$ and

$Y_2 = 0$ if $y_2^* \leq s$

$Y_2 = 1$ if $y_2^* > s$

Then the probability of forming a joint probability with cells with category 0 salinity of north coast seawater and category 0 of salinity of south coast seawater can be formed as follows:

$p_{00}(x) = p(Y_1 = 0, Y_2 = 0)$

$$= p(y_1^* \le r, y_2^* \le s)$$
$$= p(\beta_1^T x + \varepsilon_1 \le r, \beta_2^T x + \varepsilon_2 \le s)$$
$$= p(\varepsilon_1 \le r - \beta_1^T x, \varepsilon_2 \le s - \beta_2^T x)$$
$$= p(\varepsilon_1 \le z_1, \varepsilon_2 \le z_2)$$
$$= \int_{-\infty}^{z_2} \int_{-\infty}^{z_1} \emptyset\,(t_1, t_2)\,dt_1 dt_2$$

$$= \Phi(z_1, z_2)$$

So that probability equation for category 0 salinity of north shore seawater and category 0 salinity of south coast seawater. Furthermore, the equations for probability are described with category 0 salinity of north shore seawater and category 1 salinity of south coast seawater.

Next, build the likelihood function of bivariate random variables with multinomial distribution:

$$L(\beta, \rho) = \prod_{i=1}^{n} P\left(Y_{11} = y_{11}, Y_{10} = y_{10}, Y_{01} = y_{01}\right)$$

$$= \prod_{i=1}^{n} p_{11}^{y_{11}} p_{10}^{y_{10}}(x_i) p_{01}^{y_{01}}(x_i)[1 - p_{11}(x_i) - p_{10}(x_i) - p_{01}(x_i)]^{1-y_{11}-y_{10}-y_{01}}$$

After obtaining the likelihood function, then the likelihood function is used as ln function, resulting in:

$$\ln L(\beta, \rho) = \prod_{i=1}^{n} p_{11i}^{y_{11}} p_{10i}^{y_{10i}} p_{01i}^{y_{01i}} [1 - p_{11i} - p_{10i} - p_{01i}]^{y_{00i}}$$

$$= \sum_{i=1}^{n} y_{11i} \ln p_{11i} + y_{10i}\, p_{10i} + y_{01i} \ln p_{01i} + y_{00i} ln(1 - p_{11i} - p_{10i} - p_{01i})$$

The above equation contains the parameter β and ρ with $\beta = [\beta_1^T \quad \beta_2^T]^T$ and $\beta_1 = [\beta_{10} \quad \beta_{11} \quad \beta_{12} \ldots \quad \beta_{1q}]^T$. $\beta_1 = [\beta_{20} \quad \beta_{21} \quad \beta_{22} \ldots \quad \beta_{2q}]^T$ and ρ is the correlation coefficient.

3.5. Bivariate Probit Regression Model

The bivariate probit model was built using seawater salinity data on the south coast and north coast in 2019-2020 as two response variables by involving three predictor variables that are thought to have an effect on the response variable. Then the model has been tested simultaneously and partially. By using Matlab, the following bivariate probit model is obtained:

$\hat{y}_1^* = 0{,}4622104 + 0{,}02708829x_1 + 0{,}02708829x_2 - 0{,}06002294x_3$
$\hat{y}_2^* = -0{,}03155939 - 0{,}004274068x_1 - 0{,}007268843x_2 + 0{,}01467278x_3$

Table 6. Test statistics value of each parameter in the initial model.

Parameter	SE	W	Z	Conclusion
$\hat{\beta}_{10} = 0.4622$	3.0924	1.4946	-0.0949	Reject H_0
$\hat{\beta}_{11} = 0.0157$	0.0174	0.8987	-2.1525	Reject H_0
$\hat{\beta}_{12} = 0.0271$	0.0225	1.2031	-1.9254	Reject H_0
$\hat{\beta}_{13} = -0.0600$	0.0131	-4.5906	-1.5546	Reject H_0
$\hat{\beta}_{20} = -0.0316$	3.0924	-1.0205	-1.8584	Reject H_0
$\hat{\beta}_{21} = -0.0043$	0.0048	-0.8987	-2.6296	Reject H_0
$\hat{\beta}_{22} = -0.0073$	0.0060	-1.2031	-2.4437	Reject H_0
$\hat{\beta}_{23} = 0.0147$	0.0032	4.5906	-2.1788	Reject H_0

3.6. Selection of the Best Model

To get the best model to do with how to combine all the possible models is $2^q - 1$ as much as where q is the number of predictor variables and then taken the model with the smallest AIC. To get the best bivariate probit model, the modeling was done 7 times. The best model is selected based on the smallest AIC value for Y_1 and Y_2.

$\hat{y}_1^* = 0.4622 + 0.0271x_2 - 0.0600x_3$
$\hat{y}_2^* = -0.0316 - 0.0043x_1 - 0.0073x_2$

After obtaining the best bivariate probit model for seawater salinity modeling, it is necessary to re-do simultaneous and partial tests to determine the variables that have a significant effect on seawater salinity modeling.

3.7. Simultaneous test

By using matlab, the log likelihood value of each model is obtained so that test statistics are obtained G^2 as follows:

Hypothesis:

H_0: $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = 0$ and $\beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} = \beta_{25} = \beta_{26} = 0$

H_1: at least one $\beta_{kl} \neq 0$; k =1,2 and l = 1,2,3,4

$$G^2 = 2\sum_{i=1}^{n}\left[Y_{11i}\, ln\left(\frac{\hat{p}_{2i}\ \ \hat{p}_{01i}}{\hat{p}_{2i}^* - \hat{p}_{01i}^*}\right) + Y_{10i}\ln\left(\frac{\hat{p}_{1i} - \hat{p}_{2i} - \hat{p}_{01i}}{\hat{p}_{1i}^* - \hat{p}_{2i}^* + \hat{p}_{01i}^*}\right) + Y_{10i}\ln\left(\frac{\hat{p}_{01i}}{\hat{p}_{01i}^*}\right) + \ \ Y_{00i}\ln\left(\frac{1 - \hat{p}_{1i} - \hat{p}_{01i}}{1 - \hat{p}_{1i}^* - \hat{p}_{01i}^*}\right)\right]$$

At a significant level of α the decision reject H_0 if $G^2 > X_g^2, df$ or *p-value* $< \alpha$ with *df* (*degree of freedom*) is the number of parameters under the population minus the number of parameters below H_a.

4. Conclusions

From the discussion above, we got the best bivariate probit model for seawater salinity:

$\hat{y}_1^* = 0.4622 + 0.0271x_2 - 0.0600x_3$
$\hat{y}_2^* = -0.0316 - 0.0043x_1 - 0.0073x_2$

This model is the best model of bivariate probit regression for the case of seawater salinity modeling with an AIC value of -9.03. The significant predictor variables in the first model are variables x_2 (wind speed) and x_3 (sea water temperature). Meanwhile, in the second model, the significant variables are x_1 (air humidity) and x_2 (wind speed).

Acknowledgements

Thank you to the Ministry of Education, Culture, Research, and Technology (Kemendikbudristek) for funding this research, so that this research can be completed. We would also like to thank the leadership of the university, especially the Rector who gave us the opportunity and the means to conduct this research.

References

[1] Hariyanto 2013 *Implementasi Program penyaluran Dana PNPM KP PUGAR (Program Nasional pemberdayaan masyarakat mandiri kelautan dan Perikanan pemberdayaan Usaha Garam Rakyat)* Pati

[2] Ratulangi 2014 *Peran TIK (Teknologi Informasi dan Komunikasi) pada Industri Garam di Indonesia*

[3] Adi T R 2006 *Panduan Pengembangan Usaha Terpadu Garam dan Artemia* Jakarta: Pusat Riset Wilayah Laut dan Sumberdaya Nonhayati Badan Riset Kelautan dan Perikanan Departemen Kelautan dan Perikanan

[4] Sari B Y 2015 *Model Regresi Probit Bivariat Pada Kasus Penderita HIV dan AIDS di Jawa Timur* Surabaya: Fakultas Matematika dan Ilmu Pengetahuan Alam ITS

[5] Li X, Sarkar A, Xia X, Memon W H 2021 Village Environment, Capital Endowment, and Farmers' Participation in E-Commerce Sales Behavior: A Demand Observable Bivariate Probit Model Approach *Agriculture* **11** (9): 868. https://doi.org/10.3390/agriculture11090868
[6] Gilenko E & Chernova A 2021 Saving behavior and financial literacy of Russian high school students: An application of a copula-based bivariate probit-regression approach *Children and Youth Services Review* **127** 106122 doi:10.1016/j.childyouth.2021.106122
[7] Han S & Lee S 2019 Estimation in a generalization of bivariate probit models with dummy endogenous regressors *Journal of Applied Econometrics* doi:10.1002/jae.2727
[8] Schuler M S & Rose S 2016 Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies *American Journal of Epidemiology* **185** (1) p.65–73 doi:10.1093/aje/kww165
[9] Chang L Zhou Z Chen Y Xu X Sun J Liao T & Tan X 2018 Akaike Information Criterion-based conjunctive belief rule base learning for complex system modeling *Knowledge-Based Systems* doi:10.1016/j.knosys.2018.07.029

Seawater salinity modeling using bivariate probit regression

ORIGINALITY REPORT

13%	11%	6%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

4%

★ Submitted to University of Ulster

Student Paper

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		