ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 1, Issue 2, October 2021, pp. 68-75

Comparison of Simple and Segmented Linear Regression Models on the Effect of Sea Depth toward the Sea Temperature

Muhammad Bayu Nirwana [a,1,*], Dewi Wulandari [b,2]

[a] Universitas Sebelas Maret, Jl. Ir. Sutami No. 36, Surakarta 57126, Indonesia
[b] Universitas PGRI Semarang, Jl. Dr. Cipto No 24, Semarang 50232, Indonesia
[1] mbnirwana@staff.uns.ac.id*; [2] dewiwulandari@upgris.ac.id
* Corresponding author

ARTICLE INFO

Article history
Received September 1, 2021
Revised November 2, 2021
Accepted November 11, 2021

Keywords
Linear Regression
Maximum likelihood
Breakpoint
Piecewise model
Segmented regression

ABSTRACT

The linear regression model is employed when it is identified a linear relationship between the dependent and independent variables. In some cases, the relationship between the two variables does not generate a linear line, that is, there is a change point at a certain point. Therefore, the maximum likelihood estimator for the linear regression does not produce an accurate model. The objective of this study is to presents the performance of simple linear and segmented linear regression models in which there are breakpoints in the data. The modeling is performed on the data of depth and sea temperature. The model results display that the segmented linear regression is better in modeling data which contain changing points than the classical one.

1. Introduction

Linear regression is one of the statistical methods administered to scrutinize the relationship between the dependent variable affected by the dependent variable [1]. Although it is a classic method, linear regression is significantly powerful to employ even for very large data [2,3]. One of the characteristics of linear regression is that the resulting model provides a straight line which can be identified on the scatter plot. Furthermore, linear regression is a statistical model that is easy to apply and can be interpreted easily. It makes linear regression extensively employed in data analysis.

In its development, the linear relationship between the independent and the dependent variable is constantly not immediately fulfilled. It is possible that there is a change point causing the linear line in the regression model to change direction. It makes the regression model own two or more different linear lines at a change point, in which at the change point, the gradient of the regression line changes in a different direction even though it is still a linear line. Such regression models are well-known as segmented linear regression models or piecewise regression models.

Some literature administering segmented regression models as proposed by Robinson et al [4] discuss the determination of threshold points in biologically suitable hydraulic systems to protect fish in the river infrastructure. The segmented regression is also employed in research to evaluate the relationship between the distribution of energy input and shared particle size [5].

p-ISSN 2798-253X
e-ISSN 2798-3153

Nirwana and Wulandari [6] conducted a literature and simulation study of segmented simple linear regression with one change point based on the model designed by Muggeo [7]. The R program package implementing the segmented library was developed by Muggeo [8] to help perform computations in segmented linear regression modeling. Some of the segmented regression method developments for the segmented linear regression model encompass Muggeo [9] who performed a score-based approach to test model with a nuisance parameter which presents only under the alternative hypothesis, and development of interval estimation for the breakpoint in segmented regression [10].

The sea is an essential component of the Earth which is as home to various creatures. The deeper the ocean, the more the sunlight fades, the temperature decreases, and the pressure increases at a tremendous rate. On the other hand, sea temperature is affected by several factors such as location, ocean currents, local weather, depth, wind and many other factors. In this paper, sea temperature modeling is limited to one factor only, that is the depth of the sea. It aims to show the performance of the proposed method on the data pattern formed from the effect of sea depth toward the sea temperature.

The effect of sea depth toward sea temperature is modeled by employing the simple linear regression models and the segmented linear regression models. These two models are employed if there is a linear effect of the independent variable toward the dependent variable. However, if there is a breakpoint in the data, a segmented linear regression model is better to implement. Therefore, this study compared and analyzed the effect of sea depth on sea temperature using simple linear regression and segmented linear regression.

2. Methods

2.1. Linear Regression

Linear regression is one of the most extensively employed statistical methods. It is because linear regression is easy to implement, and the resulting model is easy to interpret. Linear regression models the linear relationship between the dependent variable influenced by one or more independent variables. Formula for linear regression model with one independent variable is presented in Eq. (1).

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i. \quad (1)$$

The coefficient of linear regression was attained by estimating the linear regression parameters. The estimation can be conducted by employing the least squares method or the maximum likelihood method. In linear regression, it was foreseen by utilizing the least squares method and the maximum likelihood method which produce the same estimator. Those estimators are displayed in Eqs. (2) and (3).

$$\hat{\beta}_1=b_1=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad (2)$$

$$\hat{\beta}_0=b_0=\bar{y}-b_0\bar{x} \quad (3)$$

2.2. Segmented Linear Regression with One Breakpoint

Segmented linear regression is a development of a linear regression model which provides two or more linear lines in a model. Segmented linear regression is administered when a change in the direction of the regression line at an independent variable point occurs [7]. Due to a shape of the linear line which intersects and changes direction at a certain point, segmented linear regression is also known as piecewise regression. The illustration of segmented linear regression is presented in Fig. 1.

p-ISSN 2798-253X
e-ISSN 2798-3153

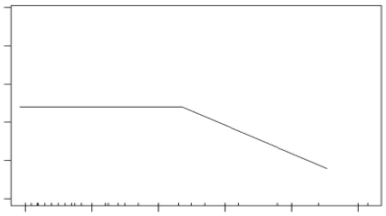

Fig. 1. Segmented linear regression with single change point.

The general model of segmented linear regression with one independent variable follows a method that proposed by Muggeo [8], displayed by the Eq. (4).

$$y_i=\beta_0+\beta_1 x_i+\beta_2(x_i-\psi)_+ +\varepsilon_i \quad (4)$$

In which y_i is the dependent variable, x_i is the independent variable, β_0 is the intercept, β_1 is the regression coefficient or gradient before the change point, β_2 is the regression or gradient coefficient after the change point, and ψ is the change point. Meanwhile, $(x_i-\psi)_+$ can be explained as in Eq. (5).

$$(x_i-\psi)_+=(x_i-\psi)\times I(x_i>\psi) \quad (5)$$

In which I(.) is an indicator function valued one if $(x_i>\psi)$ is fulfilled. The coefficient estimation of the regression model was completed by estimating the location of the change point first. Nonlinear term in Equation 4 owns intrinsic linear form that allowing to form the problem to linear framework. An initial value for the change point, $\tilde{\psi}$, initiated to estimate Eq. (4) by accommodating iteratively the linear model with linear predictor.

$$y_i=\beta_0+\beta_1 x_i+\beta_2(x_i-\tilde{\psi})_+ +\gamma I(x_i>\tilde{\psi})^- \varepsilon_i \quad (6)$$

Where γ is a reparameterization from ψ and yield change point estimates. The location of the change point is examined by the iteration method. In each iteration, a standard linear model is organized and the change point value is replaced with a new value with $\hat{\psi}=\tilde{\psi}+\hat{\gamma}/\hat{\beta}_2$. Iteration continues until convergence is attained, which is $\gamma\approx 0$. The standard error of $\hat{\psi}$ can be yield employing the Delta method for $\hat{\gamma}/\hat{\beta}_2$ which reduces to $SE(\hat{\gamma})/|\hat{\beta}_2|$ if $\hat{\gamma}=0$. The estimation of the change point location also results in the other parameters estimation of the regression model [7,8].

To assess the existence of change point, if the change point exists, the difference-in slopes parameter is not zero [8]. That is

$$H_0:\beta_2(\psi)=0 \quad (7)$$

With p-value is

$$\text{p-value}\approx\Phi(-M)+V\exp\left(-M^2/2\right)(8\pi)^{-1/2} \quad (8)$$

Where $M=\max\left(S(\psi_k)\right)_k$ is the maximum of the K test statistics, $\Phi(.)$ is the standard Normal Distributon function, and $V=\sum_k\left(\left|S(\psi_k)-S(\psi_{k-1})\right|\right)$ is the total variation of $\left(S(\psi_k)\right)_k$ [8].

2.3. Model Selection

There are several criteria in selecting the best regression model. Three of them that often used are R-Square, Akaike's Information Criterion (AIC), and Bayesian Information Criterion (BIC). The regression model is said to be better in modeling data, than other models, if it has a bigger R-Square value, a smaller AIC value, and a smaller BIC value.

2.4. Data Source and Research Variables

This study employs secondary data obtained from the California Cooperative Oceanic Fisheries Investigations (CalCOFI). The data administered are CalCOFI hydrographic and plankton data which can be utilized by the public without restrictions. The data employed for this study is limited to sampling in November 18, 2019, with latitude coordinate is 34 N and longitude coordinate is 121 W.

p-ISSN 2798-253X
e-ISSN 2798-3153

There are two variables administered in this study, which were sea temperature as the dependent variable and sea depth as the independent variable, sea temperature in Celsius degrees and sea depth in meters.

2.5. Data Analysis Method

This research employed R-Studio software as a tool in data analysis. Segmented linear regression analysis was performed using the segmented package of R [8]. The data analysis method in this study is demonstrated in Fig. 2.

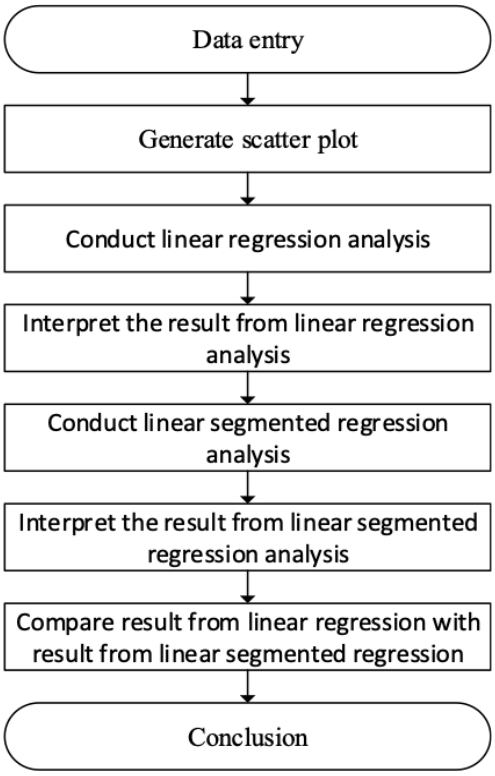

Fig. 2. Flowchart of data analysis method.

4. Results and Discussion

This section elaborates the application of linear regression and segmented linear regression on CalCOFI hydrographic and plankton data, and compares the performance of the two regression models on these data.

4.1. Scatter plot

Before analyzing the data employing a linear regression model, it is necessary to create a scatter plot between the variable of sea temperature and the sea depth. The scatter plot can be perceived in the following figure.

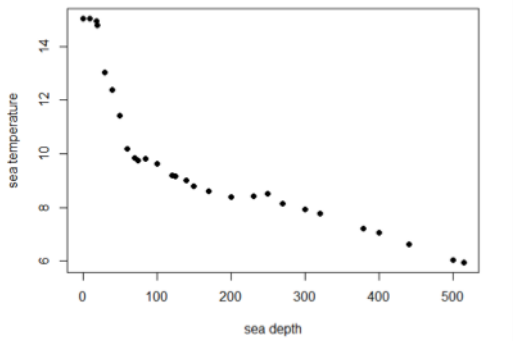

Fig. 3. Scatter plot between sea temperature and sea depth.

Fig. 3 presents that the relationship between temperature and ocean depth is not completely linear. However, there is a change point at a certain point in the sea depth variable which makes the gradient in the scatter plot turn.

4.2. Linear regression model

Data analysis employing linear regression produces a summary model and regression coefficient estimation as in Table 1 and 2.

Table 1. Model Summary of Linear Regression

Summary	value
F-statistic (p-value)	75.4 (1.97e-09)
R-squared	0.7292
AIC	115.7757
BIC	119.9793

Table 2. Estimation of Linear Regression Coefficient

	coefficient	std. error	t-value	p-value
Intercept	12.811168	0.423415	30.257	< 2e-16
depth	-0.016070	0.001851	-8.684	1.97e-09

Table 1 presents that the linear regression model is significant at the 0.05 level of significance. Furthermore, the R-square value of 0.7292 illustrates that the linear regression model is able to accommodate the data quite well. Furthermore, the estimated regression coefficient is also significant at a significance level of 0.05. The following figure presents the accuracy of the regression line against the data:

p-ISSN 2798-253X
e-ISSN 2798-3153

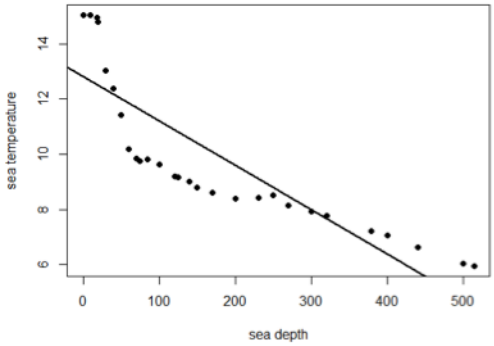

Fig. 4. Fitting plot of regression model to the data.

4.3. Segmented linear regression with single change point

Analyzing data utilizing segmented linear regression with single change point generates a summary model and regression coefficient estimation as in Table 3.

Table 3. Model Summary of Segmented Linear Regression

summary	value
psi1.depth (p.score test)	71.5 (4.138e-16)
R-squared	0.9886
AIC	24.70215
BIC	31.70814

Table 4. Estimation of Segmented Linear Regression Coefficient

	coefficient	std. error	t-value	p-value
Intercept	15.705639	0.158513	99.08	<2e-16
depth	-0.084087	0.004375	-19.22	<2e-16
U1.depth	0.075817	0.004409	17.19	-

Table 3 displayed the results which obtained that the change point is at depth = 71.5. Furthermore, from the results of the p.score test to examine the existence of the change point, the estimation results prove that the change point depth = 71.5 really exists. The location of the change point in the data is presented in Fig. 5.

p-ISSN 2798-253X
e-ISSN 2798-3153

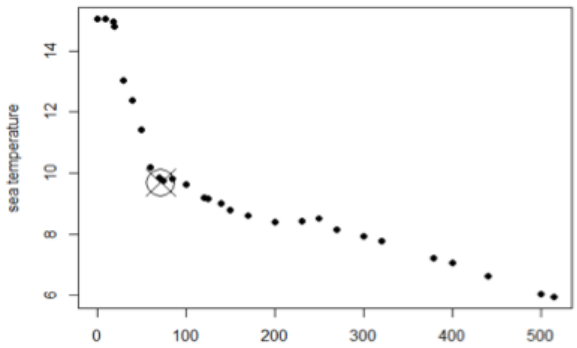

Fig. 5. Change point location.

Table 4 demonstrates that the segmented linear regression model is significant at a significance level of 0.05. In table 5 it can be observed that the R-Square value is 0.9886 indicating that the segmented linear regression model with a single change point can accommodate the data well. Fig. 6 displays the model fitting of the segmented linear regression with a single point change to the data.

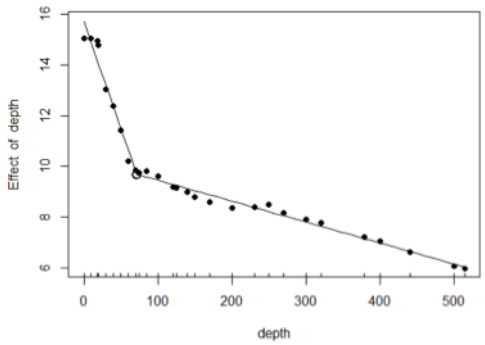

Fig. 6. Segmented linear regression fitting model with a single change point.

4.4. Comparison of the Estimation Results

Based on the estimation results of the two models, model summary comparison of the two models was administered to obtain the best fit model for the data. The summary model is compared based on the R-Squared, AIC, and BIC values.

Table 5. Comparison of Model Summary of Two Models

	Linear regression	**Segmented linear regression**
R-squared	0.7292	0.9886
AIC	115.7757	24.70215
BIC	119.9793	31.70814

Table 5 exhibits R-squared, Akaike's Information Criterion (AIC), and Bayesian Information Criterion (BIC) values form linier regression dan segmented regression models. From table 5, it can be inferred that the segmented linear regression model is better in modeling the data regarding the depth effect on sea temperature than the simple linear regression model. It is because the R-square value of the segmented linear regression model is 0.9886, greater than the R-square of the simple linear regression model, which is 0.7292.

The conclusion is corroborated by the AIC and BIC values in which the AIC value of the segmented linear regression model is 24,70215 which is much smaller than the AIC value of the linear regression model that is 115.7757. Similarly, the BIC value of the segmented linear regression model of 31.70814 is much smaller than the BIC value of the linear regression which is 119,9793.

5. Conclusion

Based on the results and discussion, the segmented linear regression is able to model data properly containing a change point. It is justified analytically by the R-Square value in segmented linear regression which is greater than the R-Square value in linear regression. Furthermore, the values of AIC and BIC are smaller than AIC and BIC in linear regression. Research development can be conducted on a linear regression model with two or more independent variables, or it can also be performed for multiple changing points on one independent variable.

References

[1] D.C. Montgomery, E.A. Peck, G.G. Vining, "Introduction to linear regression analysis", Wiley series in probability and statistics, 5th ed, 2012.

[2] J.D. Kelleher, and B. Tierney, "Data science", Massachusetts Institute of Technology Press, 2018.

[3] J.D. Miller, "Statistics for data science; leverage the power of statistics for data analysis, classification, regression, machine learning, and neural network", Packt Publishing, 2017.

[4] W. Robinson, B. Miller, B. Pflugrath, L. J. Baumgartner, A. Navarro, R. Brown, and Z. Deng, "A piecewise regression approach for determining biologically relevant hydraulic thresholds for the protection of fishes at river infrastructure", Journal of Fish Biology, vol. 88(5), 2016, pp. 1677-1692. doi: https://doi.org/10.1111/jfb.12910

[5] E. Petrakis, E. Stamboliadis, and K. Komnitsas, "Evaluation of the relationship between energy input and particle size distribution in comminution with the use of piecewise regression analysis", Particulate Science and Technology, vol. 35(4), 2017, pp. 479-489. doi: https://doi.org/10.1080/02726351.2016.1168894

[6] M.B. Nirwana and D. Wulandari "Estimasi titik ubah tunggal pada regresi linier dengan satu peubah bebas," Prosiding Seminar Nasional Pendidikan Sains dan Teknologi, Universitas Muhammadiyah Semarang, 2018.

[7] V.M. Muggeo, "Estimating Regression Models with Unknown Break-points", Statistics in Medicine, vol. 22(19), 2003, pp. 3055-3071, doi: https://doi.org/10.1002/sim.1545.

[8] V.M. Muggeo, "Segmented: An R package to Fit Regression Models with Broken-Line Relationships", R NEWS, vol. 8(1), 2008, pp. 20-25, https://cran.r-project.org/doc/Rnews/

[9] V.M. Muggeo, "Testing with a nuisance parameter present only under the alternative: a score-based approach with application to segmented modelling", J of Statistical Computation and Simulation, vol. 86, 2016, pp. 3059-3067. doi: https://doi.org/10.1080/00949655.2016.1149855

[10] V.M. Muggeo, (2017). "Interval estimation for the breakpoint in segmented regression: a smoothed score-based approach", Australian & New Zealand Journal of Statistics, vol. 59, 2017, pp. 311-322. doi: https://doi.org/10.1111/anzs.12200

p-ISSN 2798-253X
e-ISSN 2798-3153