# A STUDY ONJOINPOINT REGRESSION MODEL USING LOG-LIKELIHOOD AND INFORMATION CRITERION

**[1]YASUHIKO TAKEMOTO, [2]IKUO ARIZONO**

[12]kindai University, okayama University
takemoto@mech.kindai.ac.jp, arizono@okayama-u.ac.jp

**Abstract** - The traditional research about a joinpoint regression modelhas been established under the concept of minimizing residual sum of squares. Hence, any errors on segmented lines have been assumed to have same variance, namely, homoscedastic errors. Then, the number of lines has been determined using permutation tests in the traditional research. Permutation tests consume a huge of time since many data sets are generated by randomly shuffling all errors. Additionally, it is not necessarily practical that any errors on segmented lines have same variance. Hence, this paper deals with a joinpoint regression model without the assumption that any errors on segmented lines have same variance, namely, heteroscedastic errors. In this case, we construct a joinpoint regression model using likelihood theory. Then, a procedure for determining the number of lines is established using information criterion, which is closely related to likelihood theory.

**Index Terms** - dynamic programming, information criterion, joinpoint regression, maximum log-likelihood.

## I. NTRODUCTION

In recent years, data analysishas an important role in various fieldssuch as business, medicine, production and operation management.We show an example of General Electric (GE) company. GE manufactures and sells wind turbines as a product.GE monitors the operation and condition of the wind turbines using information and communication technology (ICT), and provides advices about an optimal operation of turbines based on data analysis to wind turbine operators.GE realized the increase in generated electricity by 5% or more, and achieved profit up by 20% per turbine[1].Such a project is known as the 4th industrial revolution. In the project, data science including data analysistakes onan important role.

The regression analysis (sometimes, called curvefitting) is very famous method in statistics, data science, and its applied fields.The main purpose of regression analysis is to estimate one of the variables from the others through a mathematical model by determining an equation connecting the variables.The regression analysis is divided into two categories from the view point of the number of variables: single and multiple. On one hand, the regression analysis is divided into two categories in the viewpoint of mathematical models: linear and non-linear. In many cases, linear regression analysis is utilized because of its simplicity. Note thatthe relationships between the variables are various.The relationship between two variables is not always linear. When the relationship between two variablesis considered a quadraticor an exponential function, the linear regression analysis can be applied after transformation of variables.However,the relationship between two variables has not been always understood. Also, the transformation function of variables for applying the linear regression analysisis not always found.In this case, non-linear regression analysis needs to be applied.However, the design and selection of a non-linear regression function is not necessarily easy.For this reason, someresearchers have discussed a method of utilizing linear functions as a simple function in any regression analyses.

A joinpoint regression model is composed of a few continuous linear phases and is often useful to describe changes in trend data [2]. The joinpoint regression function is formulated as follows:

$$y = \alpha + \beta_1 x + \beta_2 (x - \gamma_1)^+ + \cdots + \beta_K (x - \gamma_{K-1})^+, \qquad (1)$$

where $\alpha$ is the intercept, $\beta_k$ is theslope, and $\gamma_k$ is the joinpoint between adjacent segmented lines. Then, $K$ denotes the number of segmented lines. These parameters have to be estimated from observations. Also, $x^+ = \max\{x, 0\}$.

The traditional paper about a joinpoint regression model hasbeen established under the concept of minimizing residual sum of squares (RSS). The parameters of $\alpha$ , $\beta_k$ , and $\gamma_k$ are determined to minimize the sumof squares for difference between observations and estimated values, that is, RSS. Hence, any errors on segmentedlines have been assumed to have same variance, namely, homoscedastic errors.Then, the number of lines has been inferred using permutation tests. Permutation tests is a type of statistical significant tests in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangement of the observed data. In the permutation test on the joinpoint regression model, many data sets are generated by randomly shuffling any errors, and goodness of fit in the number of lines is examined. However, the permutation test consumes a huge of time since many data sets are generatedby randomly shuffling. Further, tests are repeatedly practiced for inference of the number of lines, such as 1 versus 2, 1 versus 3 for the number of lines.Additionally, it is not necessarily practical that any errors on segmented lines have same variance.

This paper deals with a joinpoint regression model under the assumption that any errors on segmented

lines don'talways have same variance, namely, heteroscedastic errors.In this case, we construct a joinpoint regression model using likelihood theory [3]. A likelihood function is constructed using the probability density function (PDF) of errors. The maximum likelihood estimators of $\alpha$ , $\beta_k$ , and $\gamma_k$ are derived from the likelihood function. Then, a procedure for determining the number of lines is established using information criterion. In particular, likelihood theory and information criterion are compatible because the information criterion is an expanded methodology from likelihood theory. Then, we show an example in our proposal and then conclude this paper.

## II. NEW PROCEDURE FOR JOINPOINT REGRESSION MODEL

From eq.(1), a linear regression model on interval $\gamma_{k-1} \le x < \gamma_k$ is given as follows:

$$y = \alpha + \beta_1 x + \beta_2 (x-\gamma_1)^+ + \cdots + \beta_K (x-\gamma_{K-1})^+ = \alpha - \sum_{i=2}^{k} \beta_i \gamma_{i-1} + x \sum_{i=1}^{k} \beta_i \quad (2)$$

$\square\ a_k + b_k x.$

And then,

$$a_k = \alpha - \sum_{i=2}^{k} \beta_i \gamma_{i-1}, \quad (3)$$

$$b_k = \sum_{i=1}^{k} \beta_i, \quad (4)$$

where $a_1 = \alpha$ . From eqs.(3) and (4), we have the following relation:

$$\gamma_{k-1} = -\frac{a_k - a_{k-1}}{b_k - b_{k-1}}. \quad (5)$$

From this notation, we have the following statistical model when the observations $(x_i, y_i), i = 1, \cdots, n, x_i \le x_{i+1},$ are given:

$$y_i = \begin{cases} a_1 + b_1 x_i + \varepsilon_1, & x_i \le \gamma_1. \\ \quad \vdots & \quad \vdots \\ a_k + b_k x_i + \varepsilon_i, & \gamma_{k-1} \le x_i < \gamma_k. \\ \quad \vdots & \quad \vdots \\ a_K + b_K x_i + \varepsilon_n, & \gamma_{K-1} \le x_i. \end{cases} \quad (6)$$

Note that $\varepsilon_i$ in case of $\gamma_{k-1} \le x_i < \gamma_k$ denotes a variable of errors and then follows the normal distribution $N(0, \sigma_k^2)$ . From eq.(6), we can't construct a likelihood function because $\gamma_k$ is unknown in addition to $a_k$ and $b_k$ . Therefore, a good idea is needed.

Now, let continuity requirementstemporarily ignored. Then, we will establish segments based on observations inseted of the point on $x$ axis, i.e., joinpoint. That is, the statistical model in eq.(6) is transformed into the following statistical model:

$$y_i = \begin{cases} a_1 + b_1 x_i + \varepsilon_1, & i = 1, \cdots, \tau_1. \\ \quad \vdots & \quad \vdots \\ a_k + b_k x_i + \varepsilon_i, & i = \tau_{k-1}+1, \cdots, \tau_k. \\ \quad \vdots & \quad \vdots \\ a_K + b_K x_i + \varepsilon_n, & i = \tau_{K-1}+1, \cdots, \tau_K (= n). \end{cases} \quad (7)$$

$\tau_k$ is an observation number of a boundary in $k$ th segment. Based on eq.(7), we can construct a likelihood function.The error terms are expressed by the following probability density function of the normal distribution:

$$f_k(x_i, y_i; \sigma_k^2, a_k, b_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left( -\frac{(y_i - a_k - b_k x_i)^2}{2\sigma_k^2} \right). \quad (8)$$

Then, the log-likelihood function of the statistical model in eq.(7) is given as follows:

$$\log L_{\tau^{(K)}}(\sigma^{2(K)}, a^{(K)}, b^{(K)}) = \sum_{k=1}^{K} \sum_{i=\tau_{k-1}+1}^{\tau_k} \log f_k(x_i, y_i; \sigma_k^2, a_k, b_k) = -\sum_{k=1}^{K} \frac{\tau_k - \tau_{k-1}}{2} \log 2\pi\sigma_k^2 - \sum_{k=1}^{K} \sum_{i=\tau_{k-1}+1}^{\tau_k} \frac{(y_i - a_k - b_k x_i)^2}{2\sigma_k^2}, \quad (9)$$

where $\tau_0 = 0$ .And, note that $\tau^{(K)} = (\tau_1, \cdots, \tau_{K-1})$ , $a^{(K)} = (a_1, a_2, \cdots, a_K), b^{(K)} = (b_1, b_2, \cdots, b_K),$ and $\sigma^{2(K)} = (\sigma_1^2, \sigma_2^2, \cdots, \sigma_K^2)$ . For given $K$ and $\tau^{(K)} = (\tau_1, \cdots, \tau_{K-1})$ , the maximum likelihood estimators of $a^{(K)}, b^{(K)}$, and $\sigma^{2(K)}$ are obtained as follows:

$$\hat{a}_k = \bar{y}_k - \hat{b}_k \bar{x}_k, \quad (10)$$

$$\hat{b}_k = \frac{\sum_{i=\tau_{k-1}+1}^{\tau_k} (y_i - \bar{y}_k)(x_i - \bar{x}_k)}{\sum_{i=\tau_{k-1}+1}^{\tau_k} (x_i - \bar{x}_k)^2}, \quad (11)$$

$$\hat{\sigma}_k^2 = \frac{1}{(\tau_k - \tau_{k-1})} \sum_{i=\tau_{k-1}+1}^{\tau_k} (y_i - \hat{a}_k - \hat{b}_k x_i)^2, \quad (12)$$

where

$$\bar{x}_k = \frac{1}{\tau_k - \tau_{k-1}} \sum_{i=\tau_{k-1}+1}^{\tau_k} x_i, \quad (13)$$

$$\bar{y}_k = \frac{1}{\tau_k - \tau_{k-1}} \sum_{i=\tau_{k-1}+1}^{\tau_k} y_i. \quad (14)$$

We have a maximum log-likelihood for $\log L_{\tau^{(K)}}(\sigma^{2(K)}, a^{(K)}, b^{(K)})$ by substituting $\hat{a}^{(K)}, \hat{b}^{(K)}$, and $\hat{\sigma}^{2(K)}$ in eqs.(10), (11), and (12) for $a^{(K)}, b^{(K)}$, and $\sigma^{2(K)}$ in eq.(9) as follows:

$$\log L_{\tau^{(K)}}(\hat{\sigma}^{2(K)}, \hat{a}^{(K)}, \hat{b}^{(K)}) = -\sum_{k=1}^{K} \frac{(\tau_k - \tau_{k-1})}{2} \log 2\pi\hat{\sigma}_k^2 - \frac{n}{2}. \quad (15)$$

Hence, the maximum log-likelihood for given $K$ and $\tau^{(K)}$ are obtained. Further, the maximum

log-likelihood $\log L_{\tau^{(K)}}(\hat{\sigma}^{2(K)}, \hat{a}^{(K)}, \hat{b}^{(K)})$ is maximized in $\tau^{(K)}$ . The maximum likelihood estimators of $\tau^{(K)}$ is obtained from solving the following problem.

$$\hat{\tau}^{(K)} = \arg\max_{\tau^{(K)}} \log L_{\tau^{(K)}}(\hat{\sigma}_k^{2(K)}, \hat{a}^{(K)}, \hat{b}^{(K)}) \ . \qquad (16)$$

The maximizing problem in eq.(16) is interpreted as a combinatorial optimization problem. We use a dynamic programmingtechnique [4] and formulate the problem as follows, where the second term in eq.(15) is removed because of a constant independent of observations:

$$\begin{aligned} D_K(\tau_K) &= \min_{\tau^{(k)}} \left\{ \sum_{k=1}^{K} \frac{(\tau_k - \tau_{k-1})}{2} \log 2\pi\hat{\sigma}_k^2 \right\} \\ &= \min_{\tau_{K-1}} \left\{ \frac{(\tau_K - \tau_{K-1})}{2} \log 2\pi\hat{\sigma}_K^2 + D_{K-1}(\tau_{K-1}) \right\}. \end{aligned} \qquad (17)$$

From eq.(17), we have the optimal combination of boundaries $\hat{\tau}^{(K)}$ to maximize the log-likelihood in (15). After all, for any $K$ , the most appropriate statistical model in eq.(7) can be obtained.

Recall that continuity requirements are temporarily ignored. In the solution procedure mentioned above, the most appropriate segmented lines can be obtained. However, adjacent segmented lines are not always continuous. We can confirm whether adjacent segmented lines are continuous or not from the following relation:

$$\hat{\tau}_{k-1} \le \hat{\gamma}_{k-1} = -\frac{\hat{a}_k - \hat{a}_{k-1}}{\hat{b}_k - \hat{b}_{k-1}} < \hat{\tau}_k . \qquad (18)$$

We have some statistical models satisfying the constraint that continuity requirements are imposed on adjacent segmented lines by checking eq.(18) for any $K$ . That is, some candidates of the joinpoint regression model are screened carefully by checking eq.(18) for any $K$ . Note that the statistical model of $K=1$ is always a candidate.

At last, a procedure for determining the number of segmented lines is considered. As already mentioned above, we have evaluated the maximum log-likelihood in any joinpoint regression models for given $K$ . Therefore, we can practice model selection based on the information criterion. In particular, we use Akaike information criterion (AIC) [5]for the model selection. On the model selection of the joinpoint regression model, AIC is defined as follows:

$$AIC(K) = -2\log L_{\hat{\tau}^{(k)}}(\hat{\sigma}^{2(K)}, \hat{a}^{(K)}, \hat{b}^{(K)}) + 2B_K . \qquad (19)$$

$B_K$ is commonly called bias, which isgiven as the number of unknown parameters in the statistical model. In this case, every line has three parameters as $a_k, b_k$, and $\sigma_k^2$ . Note that $\tau^{(K)}$ is not consideredas an unknown parameter because $\tau^{(K)}$ is determined by $a^{(K)}$ and $b^{(K)}$ . Therefore, the bias is given as follows:

$$B_k = 3K. \qquad (20)$$

After all, the optimal model, that is, the number of segmented lines, $K$ is given as follows:

$$K^* = \arg\min_K AIC(K). \qquad (21)$$

The solution algorithm proposed in this paper without the assumption that any errors on segmented lines have same variance is bellows:

i) Set $K = 1$ .

ii) Set initial values in $\tau^{(K)}$, where $\tau^{(K)}$ is $\tau_1(=n)$ only in case of $K=1$.

iii) Based on the observations $(x_i, y_i)$ and the boundaries in each segment $\tau^{(K)}$ , the maximum likelihood estimators $\hat{a}^{(K)}, \hat{b}^{(K)}$, and $\hat{\sigma}^{2(K)}$ are obtained from eqs.(10)-(12).Evaluate the maximum log-likelihood in eq.(15).

iv) Search the optimal combination of boundaries $\hat{\tau}^{(K)}$ to maximize the log-likelihood in eq.(15) using eq.(17) based on the dynamic programming technique. In this search, repeat ii) to iv).

v) The model is accepted as a candidate of the optimal regression model if any continuity requirements in eq.(18) are satisfied.

vi) Set $K+1$ to $K$ . When reaches an upper bound of $K$ , go to viii). Otherwise, return step ii).

vii) The optimal joinpoint regression model is picked up based on eq.(21) out of some candidatesof the optimal regression model obtained before this step.

## III. NUMERICAL EXAMPLE

We show an example where the proposed solution procedure of the joinpoint regression model is applied to real data. The data set is quarter sales of department stores in JAPAN from 1st quarter in 2004 to 4th quarter in 2013 and is shown in Fig. 1.The vertical axis denotes the quarter sales in trillion yen, and horizontal axis denotes the quarters during ten years.Note that modification for seasonality has been practiced since seasonality is included in the quarter sales. In this case, we have used the link-relative method[6] (sometimes called Person's method).

Then, the result of applying the proposed solution procedure of the joinpoint regression model to marketing data in Fig. 1. is shown in Fig. 2.In this case, the number of segmented lines is estimated at 4consequently.The details ofsegments as below:

1st segment: 1st to 15th quarters.
2nd segment: 16th to 18th quarters.
3rd segment: 19th to 24th quarters.
4th segment: 25th to 40th quarters.

The 3rd line has negative large slope. The 19th and 20th quarters are corresponding to the 3rd and 4th quarters in 2008. Hence,we can consider thatthis negative large slope depends on the bankruptcy of Lehman Brothers. Also, after 25th quarter, it can be seen thatthe business improved somewhat.

Consider the application of the traditional joinpoint regression analysiswith the assumption that any errors on segmented lines have same variance to the data in Fig.1. In the analysis, we used joinpoint trend analysis software provided broadly from National Cancer

institute in US [7]. Note that the traditional joinpoint regression analysis is designed under the constraint that any errors on segmented lineshave same variance, namely, homoscedastic errors. In the case of this data set, there has been no big difference between the results by the traditional joinpoint regression analysis and the solution procedure proposed in this paper. For 40 observations in Fig.1, the traditional joinpoint regression analysis has consumed about 2 minutes for calculation, while the solution procedure proposed in this paper is instant. In the case of a large number of data, there is remarkable difference in calculation time between both.

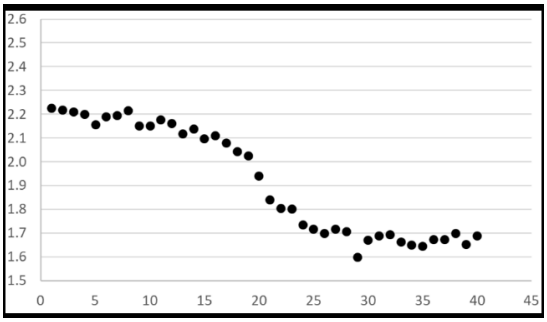



**Fig.1: quarter sales of department stores in JAPAN**

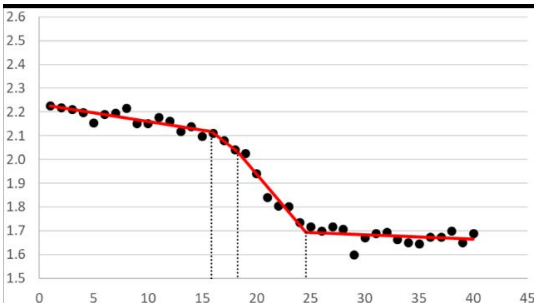



**Fig.2: Application of joinpoint regression analysis proposed in this paper**

## CONCLUDING REMARKS

The traditional researches about a joinpoint regression model have been established under the concept of minimizing residual sum of squares. Hence, any errors on segmented lines have been assumed to have same variance, namely, homoscedastic errors. Then, the number of lines has been determined using permutation tests in the he traditional researches. Permutation tests consume a huge of time since many data sets are generated by randomly shuffling any errors. On one hand, this paper has deals with a joinpoint regression model under the assumption that any errors on segmented lines don't always have same variance, namely, heteroscedastic errors. In this case, we have constructed a joinpoint regression model using likelihood theory. In particular, we have proposed the procedure that some statistical modelsare obtained under the assumption that continuity requirements are ignored and some statistical models with continuity are selected as a candidate of the optimal joinpoint regression model. Then, a procedure for determining the number of lines has been established using AIC. Further, we have shown the example in real data of marketing. Throughout the example, we have explained the difference between the traditional joinpoint regression analysis and the solution procedure proposed in this paper.

## ACKNOWLEDGEMNRT

This work was supported by JSPS KAKENHI Grant Number 17K01266: ``An investigation of data visualization and its application to production and operation management". We would like to appreciate the grant for our research.

## REFERENCES

[1] Fortune, GE uses holograms and the industrial internet to make wind farms 20% more efficient, http://fortune.com/2015/05/19/ge-holograms-wind/ (checked at May 31, 2018).

[2] H. J.Kim, M. P. Fay, E. J. Feuer, and D. N. Midthune, "Permutation Tests for Joinpoint Regression with Applications to Cancer Rates," Statistics in Medicine, vol. 19, pp. 335-351, 2000.

[3] M.G.Kendall, and A.Stuart, The Advanced Theory of Statistics, Vol.II: Inference and Relationship (4th ed.). Charles Griffin, 1979.

[4] R. R. Bellman, and S. E. Dreyfus, Applied Dynamic Programming,Princeton University Press, Princeton,1962.

[5] H.Akaike, "A new look at the statistical model identification," IEEE Transaction on Automation Control, vol.AC-19, pp.716-723, 1974.

[6] J. D.Schwager, and S. C.Turner,Futures: Fundamental Analysis (Wiley Finance), Wiley, 1995.

[7] National Cancer Institute in US: https://surveillance.cancer.gov/joinpoint/ (checked at May 31, 2018).

★★★