Regresi Tersegmen dengan Titik Patahan Diketahui

Charles E. Mongi[1]
[1] **PS Matematika FMIPA Universitas Sam Ratulangi Manado, charlesmongi@ymail.com**

Abstrak

Penggunaan analisis regresi diharapkan akan mendapat model regresi yang menjelaskan sebanyak mungkin informasi yang ada pada data. Akan tetapi belum tentu model yang telah dispesifikasikan tersebut cocok dengan data. Dengan demikian satu model kurang sesuai sehingga dibutuhkan beberapa submodel. Model regresi linier tersegmen terdiri dari beberapa submodel linier, bila digambarkan model ini berupa rangkaian garis linier yang patah-patah. Model ini bisa digunakan sebagai pendekatan terhadap bentuk kurva model regresi tidak linier. Hasil analisis regresi linier tersegmen memberikan nilai koefisien determinasi lebih besar dibandingkan model regresi linier sederhana, yang berarti bahwa model regresi linier tersegmen mampu menjelaskan keragaman data lebih besar dari keragaman yang bisa dijelaskan oleh model regresi linier sederhana. Selain itu, untuk kasus ini regresi tersegmen dengan fungsi tidak kontinu lebih baik dari regresi tersegmen dengan fungsi kontinu.

Kata kunci: Regresi Tersegmen, Titik Patahan.

Segmented Regression with a Given Breakpoint

Abstract

Utilization of regression analysis is expected in order to obtain a regression model that explains as much information as possible on the data. However, the specified model fits the data is uncertain. Thus, the use one model will be less appropriate, so that required some submodel. Segmented linear regression model consists of several sub-linear model that described in the form of a series of faltering linear lines. This model can be used as an approach towards shape of not-linear regression curve. Segmented linear regression analysis provide larger coefficient of determination than the simple linear regression model, which means that the segmented linear regression model is able to explain the diversity of data of more than diversity that can be described by a simple linear regression model. In this case, segmented regression with uncontinuous function is better than segmented regression with continuous function.

***Keywords** : Segmented Regression, Breakpoint.*

1. Pendahuluan

1.1 Latar Belakang

Analisis regresi digunakan untuk menggambarkan hubungan antara peubah respon Y dengan peubah penjelas X dalam suatu bentuk model regresi, dengan harapan model tersebut dapat menjelaskan sebanyak mungkin informasi yang ada pada data. Akan tetapi, belum tentu model yang telah dispesifikasikan tersebut cocok dengan data. Hal ini bisa saja dikarenakan setelah nilai X tertentu, titik-titik data pengamatan berada jauh dari garis regresi, atau dengan kata lain peubah respon menunjukkan pola yang berbeda setelah nilai X tertentu. Dengan demikian, satu model kurang sesuai jika digunakan untuk mewakili hubungan antara peubah respon dan peubah penjelas. Sehingga, dibutuhkan suatu model yang terdiri dari beberapa submodel, dimana masing -masing submodel digunakan pada rentang nilai X yang berbeda.

Analisis regresi linier sederhana, model regresi digambarkan sebagai sebuah garis linier. Analisis ini bisa dilakukan terhadap seluruh data atau pun membagi nilai-nilai peubah penjelas menjadi beberapa bagian (segmen) kemudian menerapkan analisis regresi pada setiap segmen, yang dikenal dengan analisis regresi linier tersegmen (*Segmented Linear Regression*) [1]. Model regresi linier tersegmen terdiri dari beberapa submodel linier, bila digambarkan model ini berupa rangkaian garis linier yang patah-patah. Oleh karena itu, model ini bisa digunakan sebagai pendekatan terhadap bentuk kurva model regresi yang tidak linier.

Analisis regresi linier tersegmen, terdapat suatu titik yang disebut *breakpoint*, yaitu titik batas antar tiap segmen (titik patah). Pada titik ini diduga mulai terjadi perubahan

bentuk hubungan matematis antara peubah respon dengan peubah penjelas. Titik ini juga digunakan sebagai indikator banyaknya segmen (s), di mana s = banyaknya *breakpoint* + 1.

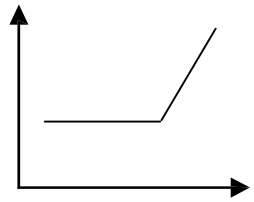
1.2 Tujuan

Tujuan adalah untuk mengkaji analisis regresi linier tersegmen sebagai pendekatan linier untuk pasangan data yang mempunyai kecenderungan bentuk kurva model regresi yang tidak linier.

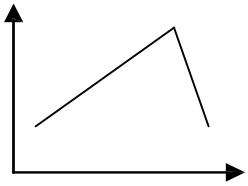
2. Bentuk Model Regresi Linier Tersegmen

Model regresi linier tersegmen adalah model regresi yang terdiri dari dua submodel yang bersifat linier. Model ini banyak digunakan di berbagai bidang, seperti ekonomi, kedokteran dan lain sebagainya dan digunakan apabila terdapat indikasi perubahan parameter setelah nilai tertentu pada peubah penjelas. Dengan demikian, nilai-nilai pada peubah penjelas terbagi menjadi dua bagian (segmen).

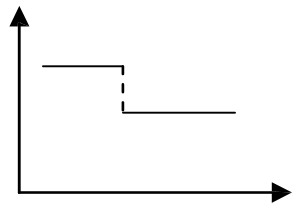
Secara umum, model regresi linier tersegmen dibagi menjadi dua tipe, yaitu model regresi yang kontinu, dimana model regresi linier di segmen pertama bertemu dengan model regresi linier di segmen kedua pada *breakpoint*, seperti tampak pada Gambar 1a dan 1b dan model regresi yang tidak kontinu, dimana model regresi linier di segmen pertama tidak bertemu dengan model regresi linier di segmen kedua pada *breakpoint*, seperti terlihat pada Gambar 1c dan 1d [2].



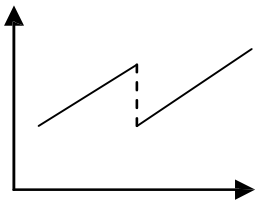
a. Regresi linier tersegmen yang kontinu pada *breakpoint* dengan slope 1 = 0 dan slope 2 > 0



b. Regresi linier tersegmen yang kontinu pada *breakpoint* dengan slope 1 > 0 dan slope 2 < 0



c. Regresi linier tersegmen yang tidak kontinu pada *breakpoint* dengan slope = 0



d. Regresi linier tersegmen yang tidak kontinu pada *breakpoint* dengan slope 1 dan slope 2>0

Gambar 1. Bentuk fungsi regresi linier tersegmen

Secara khusus, ada 7 tipe model pada analisis regresi linier tersegmen, yaitu:

1. Tipe 0 – sebuah garis horizontal tanpa *breakpoint*;
2. Tipe 1 – sebuah garis miring tanpa *breakpoint* (seperti garis regresi linier sederhana);
3. Tipe 2 – garis miring pada segmen pertama dan kedua (seperti Gambar 2.2b);
4. Tipe 3 – garis horizontal pada segmen pertama dan garis miring pada segmen kedua (seperti Gambar 2.2a);
5. Tipe 4 – garis miring pada segmen pertama dan garis horizontal pada segmen kedua;
6. Tipe 5 – garis horizontal pada segmen pertama dan kedua dengan level yang berbeda (seperti Gambar 2.2c);
7. Tipe 6 – dua garis miring yang terpisah (seperti Gambar 2.2d).

Model yang telah diduga digunakan untuk memprediksi perilaku sistem yang diamati, bukan meramalkan. Istilah prediksi mempunyai arti khusus yaitu interpolasi, yakni mencari suatu nilai fungsi yang tidak diketahui di antara beberapa nilai fungsi yang diketahui.

3. Analisis Regresi Linier Tersegmen (*Segmented Linear Regression*)

Analisis regresi linier tersegmen adalah suatu metode dalam analisis regresi yang membagi peubah penjelas menjadi beberapa segmen berdasarkan nilai tertentu yang disebut dengan *breakpoint* dimana, pada setiap segmen data terdapat model regresi linier [1].

Analisis regresi linier tersegmen dengan satu *breakpoint* disebut juga dengan analisis regresi linier dua fase. Pada metode ini, garis regresi tidak lagi disajikan dalam satu garis linier, melainkan disajikan dalam dua garis linier yang ber temu pada suatu titik, yaitu titik x = c. Dengan demikian, terdapat dua model regresi berikut:

$Y = a_1 + b_1X \qquad x \leq c$

$Y = a_2 + b_2X \qquad x \geq c$

Pada saat titik x = c,

$a_1 + b_1c = a_2 + b_2c$ (1)

di mana, titik x = c disebut sebagai *breakpoint*.

Persamaan 2.1 juga dapat dituliskan dalam bentuk:

$a_2 = a_1 + c\ (b_1\text{-}b_2)$

Jika a_2 disubstitusikan ke model regresi $Y = a_2 + b_2X$, maka akan diperoleh bentuk lain dari model analisis regresi linier tersegmen yaitu:

$Y = a_1 + b_1X \qquad x \leq c$

$Y = \{a_1 + c\ (b_1\text{-}b_2)\} + b_2X \qquad x \geq c$ [5].

Analisis regresi linier tersegmen, *breakpoint* bisa saja sudah diketahui sebelum analisis. Tapi, pada umumnya titik ini tidak diketahui dan harus diduga terlebih dahulu.

4. Pendugaan Kuadrat Terkecil

Misalkan terdapat suatu model regresi dua fase:

$Y_{si} = \alpha_s + \beta_s x_{si} + \mathcal{E}_{si} \qquad i = 1, 2, ..., n_s\ ; s = 1, 2$

dengan $x_{[11]} < x_{[12]} < ... < x_{[1n_1]} < c < x_{[21]} < x_{[22]} < ... < x_{[2n_2]}$ dan c diketahui, dimana:

$x_{[si]}$ = nilai peubah penjelas yang telah diurutkan dari nilai terkecil ke nilai terbesar

c = *breakpoint*

n_s = banyaknya pasangan data pada segmen ke – s

s = banyaknya segmen

Jika *breakpoint* diketahui, maka pendugaan parameter model regresi linier tersegmen dilakukan dengan metode kuadrat terkecil terkendala (MKTK) dengan kendala pada persamaan (1), yang diselesaikan dengan metode *Lagrange*. Sehingga, jumlah kuadrat sisa pada model regresi linier tersegmen, adalah:

$$JK_{SK} = \sum_{s=1}^{2}\sum_{i=1}^{n_s}(y_{si} - \propto_s - \beta_s x_{si})^2 + 2\,\lambda\big(\alpha_2 - \alpha_1 + c(\beta_2 - \beta_1)\big)$$

-2λ adalah pengganda *Lagrange*. Setelah menurunkan JK_{SK} terhadap α_s, β_s kemudian disamakan dengan nol, diperoleh persamaan:

$$-2\big(\textstyle\sum_i y_{1i} - n_1\tilde{\alpha}_1 - \sum_i x_{1i}\tilde{\beta}_1\big) - 2\lambda = 0 \quad (2)$$

$$-2\big(\textstyle\sum_i y_{2i} - n_2\tilde{\alpha}_2 - \sum_i x_{2i}\tilde{\beta}_2\big) + 2\lambda = 0 \quad (3)$$

$$-2\big(\textstyle\sum_i x_{1i}\big(y_{1i} - \tilde{\alpha}_1 - \sum_i \tilde{\beta}_1 x_{1i}\big)\big) - 2\lambda c = 0 \quad (4)$$

$$-2\big(\textstyle\sum_i x_{2i}\big(y_{2i} - \tilde{\alpha}_2 - \sum_i \tilde{\beta}_2 x_{2i}\big)\big) + 2\lambda c = 0 \quad (5)$$

Dari persamaan (2) dan (3) didapatkan:

$$\tilde{\alpha}_s = \bar{y}_s - \tilde{\beta}_s\bar{x}_s + (-1)^{s-1}\lambda n_s^{-1} \quad (6)$$

$\tilde{\alpha}_s$, $\tilde{\beta}_s$ adalah penduga parameter α_s, β_s dari metode kuadrat terkecil terkendala.

Dengan mensubtitusikan persamaan (6) ke dalam persamaan (1) diperoleh:

$$\lambda = w\{\bar{y}_2 - \bar{y}_1 + \tilde{\beta}_1(\bar{x}_1 - c) - \tilde{\beta}_2(\bar{x}_2 - c)\} \quad (7)$$

dimana: $w = \frac{n_1 n_2}{(n_1+n_2)}$

Selanjutnya, substitusi persamaan (6) dan (7) ke persamaan (4) dan (5), menghasilkan sistem persamaan untuk menghitung nilai $\tilde{\beta}_s$:

$$c_{11}\tilde{\beta}_1 + c_{12}\tilde{\beta}_2 = c_{13}$$
$$c_{21}\tilde{\beta}_1 + c_{22}\tilde{\beta}_2 = c_{23}$$

dimana: $c_{ss} = \sum_i (x_{si} - \bar{x}_s)^2 + w(\bar{x}_s - c)^2$

$c_{12} = c_{21} = -w(\bar{x}_i - \gamma)(\bar{x}_2 - c)$ dan

$c_{s3} = \sum_i (y_{si} - \bar{y}_s)(x_{si} - \bar{x}_s) + (-1)^s w(\bar{y}_2 - \bar{y}_1)(\bar{x}_s - c)$

Setelah $\tilde{\beta}_1$ dan $\tilde{\beta}_2$ diketahui, diperoleh nilai λ pada persamaan (7) dan $\tilde{\alpha}_s$ pada persamaan (6). Dengan demikian,nilai minimum Ɛ′ Ɛ adalah:

$$\sum_{s=1}^{2}\sum_{i=1}^{n_s}\left(y_{si} - \tilde{\alpha}_s - \tilde{\beta}_s x_{si}\right)^2$$
$$= \sum\sum\left\{y_{si} - \bar{y}_s - \tilde{\beta}_s(x_{si} - \bar{x}_s) + (-1)^s \lambda n_s^{-1}\right\}^2$$
$$= \sum\sum(y_{si} - \tilde{y}_s)^2 - 2\sum\sum\tilde{\beta}_s(y_{si} - \tilde{y}_s)(x_{si} - \bar{x}_s) + \sum\sum\tilde{\beta}_s^2(x_{si} - \bar{x}_s)^2 + \frac{\lambda^2}{w}$$

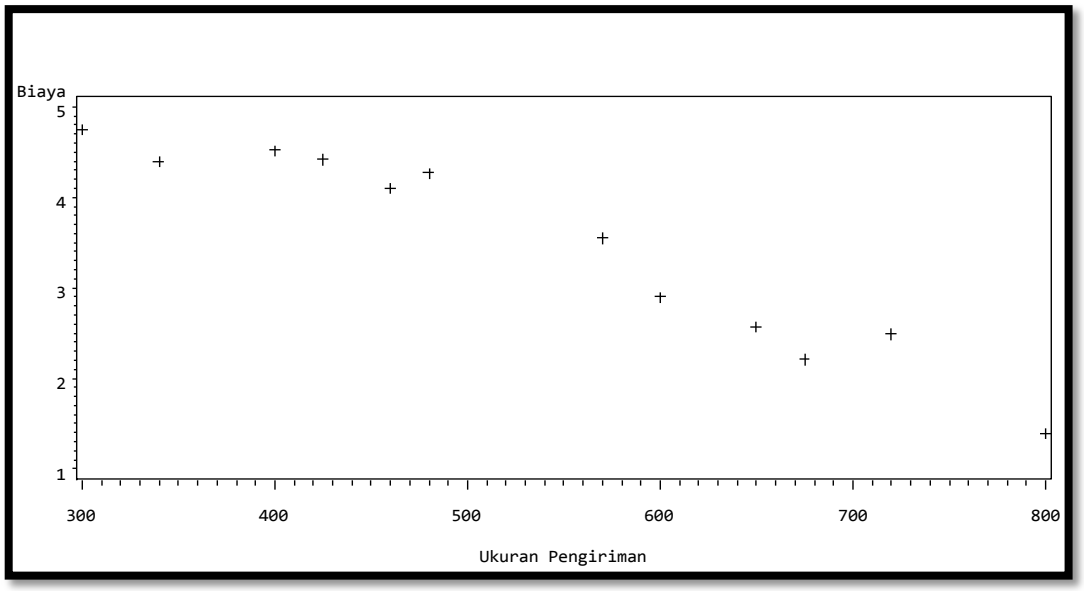
yang selanjutnya disebut JK_{SK} [3][4][6].

5. Hasil dan Pembahasan

Contoh data bahan mentah dengan biaya adalah peubah respon (Y) dan ukuran pengiriman adalah peubah penjelas (X), seperti dalam Tabel 1. Dari data dalam Tabel 1 kemudian dibuat diagram pencar seperti pada Gambar 2.

Tabel 1. Data dengan peubah respon dan peubah penjelas

Y	X
2.57	650
4.40	340
4.52	400
1.39	800
4.75	300
3.55	570
2.49	720
4.27	480
4.42	425
4.10	460
2.21	675
2.90	600



Gambar 2 Diagram Pencar antara peubah respon dan peubah penjelas

Dengan analisis regresi linear sederhana hasilnya seperti dalam Tabel 2.

Tabel 2. Hasil analisis regresi linier sederhana

The REG Procedure					
Model: MODEL1					
Dependent Variable: y Biaya					
Number of Observations Read		12			
Number of Observations Used		12			
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	12.66803	12.66803	141.95	<.0001
Error	10	0.89246	0.08925		
Corrected Total	11	13.56049			
Root MSE		0.29874	R-Square	0.9342	
Dependent Mean		3.46417	Adj R-Sq	0.9276	
Coeff Var		8.62373			
Parameter Estimates					
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	7.09563	0.31677	22.40	<.0001
x1	Uk Peng	1	-0.00679	0.00056973	-11.91	<.0001

5.1. Regresi Tersegmen Fungsi Kontinu

Dengan analisis regresi tersegmen untuk titik patahan pada X=500 hasilnya seperti dalam Tabel 3 (fungsi yang kontinu).

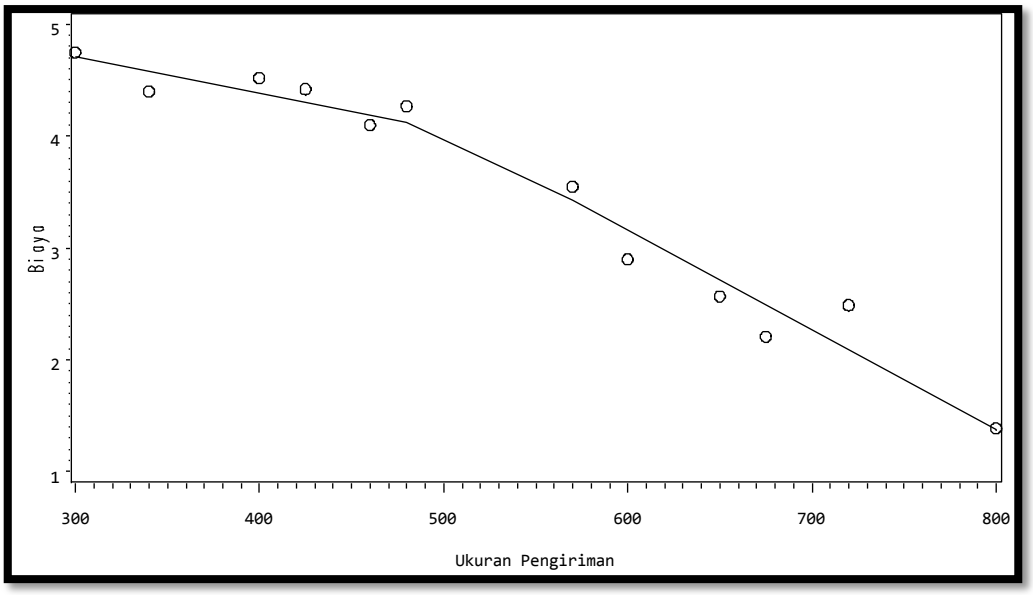
Tabel 3. Hasil analisis regresi linier tersegmen fungsi kontinu

The REG Procedure					
Model: MODEL1					
Dependent Variable: y Biaya					
Number of Observations Read		12			
Number of Observations Used		12			
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	13.12025	6.56012	134.11	<.0001
Error	9	0.44024	0.04892		
Corrected Total	11	13.56049			
Root MSE		0.22117	R-Square	0.9675	
Dependent Mean		3.46417	Adj R-Sq	0.9603	
Coeff Var		6.38450			
Parameter Estimates					
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	5.70355	0.51441	11.09	<.0001
x1	Uk Peng	1	-0.00330	0.00122	-2.70	0.0245
x2star		1	-0.00561	0.00185	-3.04	0.0140

Hasil kedua analisis tersebut nilai R^2 seperti dalam Tabel 4. Nilai R^2 untuk regresi tersegmen lebih besar dari dari regresi linear sederhana. Hasil plot regresi tersegmen dengan titik patahan X=500 seperti dalam Gambar 3.

Tabel 4. Nilai R^2 untuk masing-masing analisis regresi

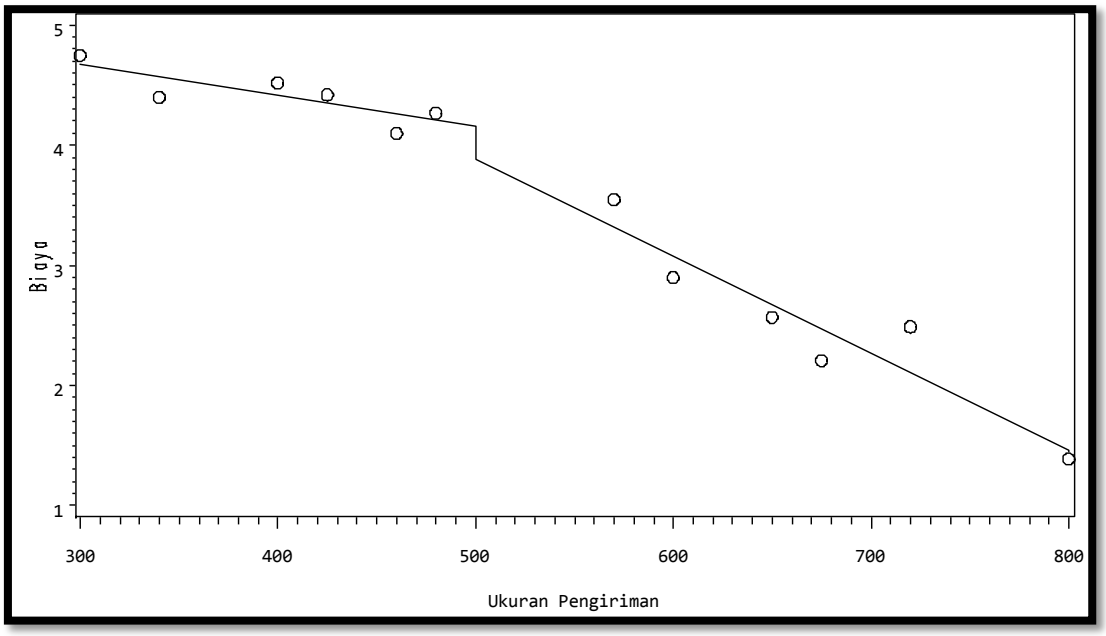
Hasil Analisis	R^2
Regresi Linear Sederhana	0.9342
Regresi Tersegmen	0.9675



Gambar 3 Plot regresi tersegmen fungsi kontinu

5.2. Regresi Tersegmen Fungsi tidak Kontinu

Hasil analisis regresi tersegmen untuk titik patahan pada X=500 hasilnya seperti dalam Tabel 5 (fungsi yang tidak kontinu). Plot regresi tersegmen dengan titik patahan X=500 seperti dalam Gambar 4.



Gambar 4. Plot regresi tersegmen fungsi tidak kontinu

Tabel 5. Hasil analisis regresi linier tersegmen fungsi tidak kontinu

The REG Procedure

Model: MODEL1

Dependent Variable: y Cost

Number of Observations Read	15
Number of Observations Used	12
Number of Observations with Missing Values	3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	13.16584	4.38861	88.96	<.0001
Error	8	0.39465	0.04933		
Corrected Total	11	13.56049			

Root MSE	0.22211	R-Square	0.9709
Dependent Mean	3.46417	Adj R-Sq	0.9600
Coeff Var	6.41155		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	5.45177	0.57919	9.41	<.0001
x1	Uk Peng	1	-0.00260	0.00143	-1.82	0.1061
x2star		1	-0.00548	0.00186	-2.95	0.0185
x3		1	-0.26700	0.27773	-0.96	0.3645

Nilai R^2 untuk fungsi tidak kontinu adalah 0.9709 lebih besar dari R^2 fungsi kontinu. Berarti untuk kasus ini regresi tersegmen dengan fungsi tidak kontinu lebih baik dari regresi tersegmen dengan fungsi kontinu.

6. Kesimpulan

Berdasarkan uraian di atas, dapat disimpulkan bahwa untuk titik patahan diketahui model yang diperoleh dari analisis regresi linier tersegmen fungsi kontinu memberikan nilai koefisien determinasi lebih besar dibandingkan model regresi linier sederhana, yang berarti bahwa model regresi linier tersegmen mampu menjelaskan keragaman data lebih besar dari keragaman yang bisa dijelaskan oleh model regresi linier sederhana. Selain itu, untuk kasus ini regresi tersegmen dengan fungsi tidak kontinu lebih baik dari regresi tersegmen dengan fungsi kontinu.

7. Daftar Pustaka

[1] Anonymous. 2006. *Segmented Regression*. http://en.wikipedia.org/wiki/Segmented_regression

[2] Diniz, C.A.R and L.C. Brochi. 2005. *Robustness of Two-Phase Regression Tests*. REVSTAT-Statistical Journal 3:3 http://www.ine.pt/revstat/pd

[3] Khoirun, I.F. 2011. Pendugaan kecepatan arus sungai dengan menggunakan regresi piecewise (studi kasus sungai soos creek di negara bagian Washington). Skripsi Departemen Statistika FMIPA IPB

[4] Oosterbaan, R.J., D.P. Sharma and K.N. Singh. 1990. *Crop production and soil salinity:Evaluation of field data from India by segmented linear regression*. Symposium on Land Drainage for Salinity Control in Arid and Semi-Arid Regions.Vol.3.Cairo http://waterlog.info/pdf/segmregr.pdf

[5] Ryan, S.E. and L.S.Porth. 2007. *A Tutorial on the Piecewise Regression Approach Applied to Bedload Transport Data*. General Technical Report RMRS-GTR-189 http://www.fs.fed.us/rm/pubs/rmrs_gtr189.pdf

[6] Shofiyati, Arina. 2008. *Kajian Analisis Regresi Linier Tersegmen*. Skripsi Jurusan Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Brawijaya. Malang.