Jurnal Matematika UNAND
Vol. **2** No. **2** Hal. 115 – 122
ISSN : 2303–291X

TRANSFORMASI BOX-COX PADA ANALISIS REGRESI LINIER SEDERHANA

ELVI YATI, DODI DEVIANTO, YUDIANTRI ASDI

Program Studi Matematika,
Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Andalas,
Kampus UNAND Limau Manis Padang, Indonesia,
lviyati@ymail.com

Abstrak. Asumsi dasar regresi merupakan asumsi yang harus dipenuhi dalam memodelkan hubungan antara variabel tak bebas (Y) dengan variabel bebas (X) dalam analisis regresi linier sederhana. Jika asumsi tersebut tidak dipenuhi, maka dapat dilakukan transformasi Box-Cox terhadap variabel tak bebas, dimana Y dipangkatkan dengan λ, sehingga menjadi Y^{λ}. Pendugaan parameter λ dilakukan dengan Metode Kemungkinan Maksimum dimana dipilih λ yang memiliki jumlah kuadrat sisaan paling kecil. Parameter λ tersebut digunakan dalam transformasi sehingga diperoleh data yang memenuhi asumsi normalitas, homogenitas, dan linieritas.

Kata Kunci: Metode kemungkinan maksimum, transformasi Box-Cox.

1. Pendahuluan

Analisis regresi adalah teknik statistika yang digunakan untuk membentuk model hubungan antara variabel bebas dengan variabel tak bebas. Hubungan antara satu variabel bebas (X) dengan satu variabel tak bebas (Y) untuk analisis regresi linier sederhana dinyatakan dalam model berikut

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i. \tag{1.1}$$

Dalam analisis regresi linier sederhana perlu diperhatikan beberapa asumsi yang dikenal dengan asumsi dasar regresi yaitu asumsi kenormalan data (normalitas), kehomogenan ragam (homogenitas), dan kelinieran data (linieritas). Apabila kenormalan data, kehomogenan ragam dan kelinieran tidak dipenuhi, maka dapat dilakukan transformasi terhadap variabel tak bebas. Salah satu transformasi yang dapat dilakukan adalah Transformasi Box-Cox yang diberlakukan terhadap variabel tak bebas Y yang bernilai positif. Transformasi Box-Cox ini berupa transformasi pangkat berparameter tunggal, katakanlah λ, terhadap Y menjadi Y^{λ}. Pendugaan parameter λ dapat dilakukan dengan menggunakan Metode Kemungkinan Maksimum (*Maximum Likelihood Methods*). λ yang diambil adalah λ yang menghasilkan jumlah kuadrat sisaan terkecil [3].

2. Analisis Regresi

2.1. *Uji Asumsi Dasar Regresi*

Dalam analisis regresi linier sederhana terdapat beberapa asumsi yang harus dipenuhi, di mana asumsi ini disebut asumsi dasar regresi. Pengujian asumsi dasar dalam analisis regresi linier sederhana diuraikan sebagai berikut.

(a) Asumsi Normalitas (Kenormalan Data)
Pengujian asumsi normalitas dengan uji Kolmogorov-Smirnov dapat dinyatakan sebagai berikut
H_0 : Data mengikuti sebaran tertentu
H_1 : Data tidak mengikuti sebaran tertentu
Berdasarkan [6], statistik uji yang digunakan adalah

$$D = \max |F_0(x) - S_N(x)|, \tag{2.1}$$

di mana $F_0(x)$ adalah fungsi kumulatif sebaran, $S_N(x)$ adalah peluang kumulatif sampel, dan N adalah banyak pengamatan. Kriteria untuk pengujian ini adalah tolak H_0 jika nilai D_{hitung} lebih besar dari nilai D_{tabel}.

(b) Asumsi Homogenitas (Kehomogenan Ragam)
Pengujian asumsi homogenitas dengan uji Levene dapat dinyatakan sebagai berikut
$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$
$H_1 : \sigma_i^2 \neq \sigma_j^2$ paling tidak untuk satu pasang (i, j)
Misalkan variabel tak bebas Y dengan ukuran sampel N yang dibagi atas k subgrup, dimana N_i menyatakan ukuran sampel dari subgrup ke-i, maka statistik uji Levene [4] dinyatakan sebagai berikut

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i(\bar{Z}_{i.} - \bar{Z}_{..})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - \bar{Z}_{i.})^2}, \tag{2.2}$$

di mana $Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|$, $\bar{Y}_{i.}$ adalah nilai tengah dari subgrup ke-i. $\bar{Z}_{i.}$ menyatakan nilai tengah grup(kelompok) dari Z_{ij} dan $\bar{Z}_{..}$ menyatakan nilai tengah secara keseluruhan dari Z_{ij}. Kriteria untuk pengujian ini adalah tolak H_0 jika nilai $W > F_{\alpha,k-1,N-k}$.

(c) Asumsi Linieritas (Kelinieran Data)
Pengujian kelinieran dengan uji F dinyatakan sebagai berikut.
H_0 : terdapat hubungan yang linier antara variabel X dan Y
H_1 : tidak terdapat hubungan yang linier antara variabel X dan Y
Dari [1], statistik uji yang digunakan adalah

$$F_{hit} = \frac{\chi_1^2/(k-2)}{\chi_2^2/(n-k)}, \tag{2.3}$$

dengan:

$$\chi_1^2 = \sum \frac{y_i^2}{n_i} - \frac{(\sum y_{ij})^2}{n} - b^2(n-1)s_x^2,$$
$$\chi_2^2 = \sum y_{ij}^2 - \sum \frac{y_i^2}{n_i},$$
$$s_x^2 = \frac{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}{n(n-1)}.$$

Didefinisikan y_{ij} sebagai nilai ke-j bagi peubah acak Y_i, sementara $y_{i.}$ adalah jumlah nilai-nilai Y_i dalam contoh. Kriteria pengujian ini adalah tolak H_0 jika $F_{hit} > F_{tabel}$.

2.2. *Metode Kemungkinan Maksimum*

Metode kemungkinan maksimum (*Maximum Likelihood Methods*) adalah metode yang digunakan untuk menduga parameter-parameter dengan memaksimumkan fungsi kemungkinan yang dibentuk dari fungsi kepekatan peluang bersama beberapa peubah acak.

Fungsi kemungkinan maksimum adalah fungsi dari θ dilambangkan dengan $L(\theta)$. Jika $X_1, X_2, \cdots, X_n$ merupakan peubah acak dari $f(x_i;\theta)$, maka

$$L(\theta) = f(x_1;\theta)f(x_2;\theta)\cdots f(x_n;\theta) = \prod_i^n f(x_i;\theta) \text{ [1]} \tag{2.4}$$

Pandang model regresi dalam notasi matriks

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \;\; ; \;\; \varepsilon \sim N(0,\sigma^2) \tag{2.5}$$

Untuk analisis regresi linier sederhana, persamaan (2.5) dapat ditulis dalam bentuk

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \;\; ; \;\; \varepsilon_i \sim N(0,\sigma^2) \tag{2.6}$$

Dalam regresi linier sederhana, fungsi kemungkinannya dapat dituliskan

$$L = L(\beta_0,\beta_1,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2\right] \tag{2.7}$$

Untuk menentukan penduga kemungkinan maksimum dari parameter-parameter β_0, β_1, dan σ^2 yang dinotasikan dengan b_0, b_1, dan $\hat{\sigma}^2$, maka persamaan (2.7) ekivalen dengan

$$\ln L(\beta,\sigma^2) = -\left(\frac{n}{2}\right)\ln 2\pi - \left(\frac{n}{2}\right)\ln\sigma^2 - \left(\frac{1}{2\sigma^2}\right)\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \tag{2.8}$$

Dengan menurunkan fungsi kemungkinan terhadap setiap parameter β_0, β_1, σ^2,

diperoleh

$$\frac{\partial \ln L}{\partial \beta_0} = 0 \Rightarrow \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0, \tag{2.9}$$

$$\frac{\partial \ln L}{\partial \beta_1} = 0 \Rightarrow \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)X_i = 0, \tag{2.10}$$

$$\frac{\partial \ln L}{\partial \sigma^2} = 0 \Rightarrow -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2 = 0. \tag{2.11}$$

Penyelesaian persamaan (2.9) – (2.11) adalah sebagai berikut.

$$b_0 = \bar{Y} - b_1\bar{X}, \tag{2.12}$$

$$b_1 = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{2.13}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - b_0 - b_1)^2}{n}. \tag{2.14}$$

Sehingga diperoleh penduga model regresi linier sederhana adalah $\hat{Y} = b_0 + b_1 X$.

Pada model $Y = X\beta + \varepsilon$, persamaan (2.8) dapat ditulis dalam bentuk

$$\ln L(\beta, \sigma^2) = -\left(\frac{n}{2}\right)\ln 2\pi - \left(\frac{n}{2}\right)\ln \sigma^2 - \left(\frac{1}{2\sigma^2}\right)(Y - X\beta)^t(Y - X\beta) \tag{2.15}$$

di mana $\frac{\partial \ln L}{\partial \beta_i} = 0 \Rightarrow b = (X^tX)^{-1}(X^tY)$.

3. Transformasi Box-Cox

Transformasi Box-Cox adalah transformasi pangkat pada variabel tak bebas di mana variabel tak bebasnya bernilai positif. Box dan Cox mempertimbangkan kelas transformasi berparameter tunggal, yaitu λ yang dipangkatkan pada variabel tak bebas Y, sehingga transformasinya menjadi Y^λ, dimana λ adalah parameter yang perlu diduga.

Prosedur transformasi Box-Cox pada analisis regresi linier sederhana untuk model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ dapat dilakukan dalam dua bentuk transformasi. Menurut [2], transformasi pertama adalah:

$$W_i(\lambda) = \begin{cases} \left(\frac{Y_i^\lambda - 1}{\lambda}\right), \lambda \neq 0 \\ \ln(Y_i), \lambda = 0 \end{cases}, \quad i = 1, 2, \cdots, n \tag{3.1}$$

Dari [5], diperoleh transformasi kedua berdasarkan $W_i(\lambda)$, dengan

$$V_i(\lambda) = \begin{cases} \left(\frac{Y_i^\lambda - 1}{\lambda \hat{Y}^{\lambda-1}}\right), \lambda \neq 0 \\ \hat{Y}\ln(Y_i), \lambda = 0 \end{cases} \tag{3.2}$$

di mana

$$\hat{Y} = \sqrt[n]{Y_1 Y_2 \cdots Y_n} = \left(\prod_{i=1}^{n} Y_i\right)^{\frac{1}{n}}.$$

merupakan rata-rata geometrik dari $Y_1, Y_2, \cdots, Y_n$.

Transformasi Y menjadi W mengakibatkan model persamaan linier dalam notasi matriks menjadi $\mathbf{W} = \mathbf{X}\beta + \varepsilon$. Transformasi Y menjadi V mengakibatkan model persamaan liniernya dalam notasi matriks menjadi $\mathbf{V} = \mathbf{X}\beta + \varepsilon$. Dengan demikian, prosedur utama transformasi Box-Cox adalah menduga parameter transformasinya yaitu λ. Salah satu metode yang dapat digunakan dalam pendugaan parameter λ pada Transformasi Box-Cox adalah Metode Kemungkinan Maksimum.

Dalam model regresi linier $\mathbf{V} = \mathbf{X}\beta + \varepsilon$ diperoleh fungsi kemungkinan sebagai berikut.

$$L(\beta, \lambda, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(V_i - \beta_0 - \beta_1 X_i)^2\right] \tag{3.3}$$

Persamaan (3.3) ekivalen dengan

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \left(\frac{1}{2\sigma^2}\right)(V(\lambda) - X\beta)^t(V(\lambda) - X\beta), \tag{3.4}$$

dengan demikian

$$\frac{\partial \ln L}{\partial \beta} = \frac{\partial[-\frac{n}{2}\ln(2\pi\sigma^2) - (\frac{1}{2\sigma^2})(V(\lambda) - X\beta)^t(V(\lambda) - X\beta)]}{\partial\beta} = 0. \tag{3.5}$$

Sehingga

$$\Leftrightarrow (X^tX)\beta = X^tV(\lambda) \Leftrightarrow b = (X^tX)^{-1}X^tV(\lambda).$$

Selanjutnya,

$$\frac{\partial \ln L}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{(V(\lambda) - \hat{V}(\lambda))^t(V(\lambda) - \hat{V}(\lambda))}{2(\sigma^2)^2} \tag{3.6}$$

di mana

$$\hat{V}(\lambda) = Xb = X(X^tX)^{-1}X^tV(\lambda),$$
$$\Leftrightarrow \hat{\sigma}^2 = \frac{(V(\lambda) - \hat{V}(\lambda))^t(V(\lambda) - \hat{V}(\lambda))}{n} = \frac{RSS(V(\lambda))}{n},$$

di mana $RSS(V(\lambda)) = (V(\lambda) - \hat{V}(\lambda))^t(V(\lambda) - \hat{V}(\lambda))$ merupakan jumlah kuadrat sisaan dari $V(\lambda)$.

Penduga kemungkinan maksimum $\hat{\lambda}$ dari λ merupakan nilai yang memaksimumkan fungsi kemungkinan maksimum. Maka, $\hat{\lambda}$ memaksimumkan

$$\ln L = \frac{-n}{2}\ln(2\pi) - \frac{n}{2}\ln\left[\frac{RSS(V(\lambda))}{n}\right] - \frac{n}{2RSS(V(\lambda))}[V(\lambda) - \hat{V}(\lambda)]^t[V(\lambda) - \hat{V}(\lambda)],$$
$$\propto \frac{n}{2}\ln\left[\frac{RSS(V(\lambda))}{n}\right] \tag{3.7}$$

$\hat{\lambda}$ meminimumkan $\frac{n}{2}\ln[\frac{RSS(V(\lambda))}{n}] \Leftrightarrow \hat{\lambda}$ diperoleh dengan menentukan nilai λ yang meminimumkan

$$RSS(V(\lambda)) = [V(\lambda) - \hat{V}(\lambda)]^t[V(\lambda) - \hat{V}(\lambda)].$$

Penaksiran parameter λ yang biasa dilakukan yaitu menentukan nilai λ pada kisaran nilai tertentu. Biasanya λ yang dipakai yaitu dari kisaran (-2,2) atau (-1,1).

4. Pembahasan

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari website *http://archive.ics.uci.edu.* Data terdiri dari 90 pengamatan, dengan satu variabel bebas dan satu variabel tak bebas. Variabel bebas (X) adalah umur penderita hepatitis yang berusia $30 - 50$ tahun dan variabel tak bebas (Y) adalah level bilirubin seorang penderita hepatitis.

4.1. *Analisis Data Awal*

Dengan menggunakan software SPSS 17 diperoleh model persamaan regresi sebagai berikut

$$Y = 0,454 + 0,024X,$$

dengan X adalah umur dan Y adalah level bilirubin.

Pengujian asumsi normalitas menghasilkan nilai signifikansi $0,000$ lebih kecil dari $0,1$, maka disimpulkan data tidak menyebar normal. Pengujian asumsi homogenitas menghasilkan nilai signifikansi $0,000$ lebih kecil dari $0,1$, maka disimpulkan ragam data tidak homogen. Pengujian asumsi linieritas menghasilkan nilai signifikansi $0,213$ lebih besar dari 0,1, maka disimpulkan bahwa antara variabel bebas (X) dan variabel tak bebas (Y) tidak terdapat hubungan yang linier.

Berdasarkan model regresidi atas, diperoleh selang level bilirubin yaitu $1,174 - -1,654$. Nilai ini tentu tidak sesuai dengan nilai rujukan level bilirubin dewasa. Oleh karena itu, perlu dilakukan transformasi Box-Cox untuk memperoleh model regresi yang sesuai mengenai hubungan antara umur dan bilirubin pada seseorang.

4.2. *Transformasi Terhadap Variabel Tak Bebas (Y)*

Pada penelitian ini dilakukan transformasi Y^{λ} terhadap variabel tak bebas (Y) dengan langkah-langkah sebagai berikut :

(1) Menentukan range λ
Range λ yang diambil adalah $(-2, 2)$ dengan nilai

No.	1	2	3	4	5	6	7	8	9
λ	2	1,5	1	0,5	0	-0,5	-1	-1,5	-2

(2) Menghitung $\hat{Y} = (Y_1 Y_2 ... Y_n)^{\frac{1}{n}}$
(3) Menghitung $\hat{Y}^{\lambda-1}$ untuk tiap harga λ
Nilai $\hat{Y}^{\lambda-1}$ untuk tiap harga λ ditampilkan dalam tabel berikut

λ	2	1,5	1	0,5	0	-0,5	-1	-1,5	-2
$\hat{Y}^{\lambda-1}$	1,1265	1,0614	1,0000	0,9422	0,8877	0,8364	0,7880	0,7425	0,6995

(4) Menghitung $V_i(\lambda)$
Diperoleh nilai V_i untuk tiap harga λ dengan $i = 1, 2, , 90$.

(5) Regresikan antara V dan X, sehingga diperoleh JKS
Diperoleh nilai JKS sebagai berikut

λ	2	1,5	1	0,5	0	-0,5	-1	-1,5	-2
$\hat{Y}^{\lambda-1}$	1603,05	437,862	147,536	65,0742	38,6102	30,1490	29,5105	34,5212	46,5506

(6) Tentukan λ yang mempunyai JKS terkecil
Nilai JKS terkecil adalah $29,5105$ pada $\lambda = -1$.

(7) Melakukan transformasi data menggunakan λ dengan JKS terkecil
Dengan $\lambda = -1$ dilakukan transformasi Y^{-1}, artinya data awal Y dipangkatkan dengan -1 yang diberi simbol dengan $Y^{'}$.

4.3. *Analisis Data Hasil Transformasi*

Pada tahap ini dilakukan regresi terhadap variabel tak bebas hasil transformasi $(Y^{'})$ dengan variabel bebas (X), sehingga diperoleh model persamaan regresi yang baru sebagai berikut

$$Y^{'} = 1,589 - 0,015X,$$

dengan X adalah umur dan $Y^{'}$ adalah level bilirubin.

Pengujian asumsi normalitas menghasikan nilai signifikansi $0,969$ lebih besar dari $0,1$, maka disimpulkan data menyebar normal. Pengujian asumsi homogenitas menghasikan nilai signifikansi $0,785$ lebih besar dari $0,1$, maka disimpulkan ragam data homogen. Pengujian asumsi linieritas menghasilkan nilai signifikansi $0,067$ lebih kecil dari $0,1$, maka disimpulkan terdapat hubungan yang linier antara variabel bebas (X) dan variabel tak bebas (Y).

Dari hasil uji asumsi dasar terhadap data hasil transformasi, diperoleh bahwa data hasil transformasi tersebut telah memenuhi ketiga asumsi. Sehingga, model yang cocok untuk hubungan antara umur dan level bilirubin adalah $Y^{'} = 1,589 - 0,015X$. Model regresi ini memberikan nilai Adjusted R Square sebesar $0,026$. Nilai ini menyatakan bahwa pengaruh variabel bebas (X) terhadap variabel tak bebas (Y) sangat kecil. Untuk model regresi ini diperoleh selang level bilirubin yaitu $0,839 - 1,139$. Nilai ini sesuai dengan nilai rujukan level bilirubin dewasa karena terletak pada selang $0,1 \pm 1,2$ mg/dL.

5. Kesimpulan

Data statistika dengan model regresi yang baik pada analisis regresi linier sederhana adalah data yang memenuhi asumsi-asumsi dasar regresi. Apabila asumsi-asumsi tersebut tidak dipenuhi, maka dapat dilakukan transformasi terhadap data, salah satunya transformasi Box-Cox. Transformasi Box-Cox merupakan transformasi pangkat terhadap variabel tak bebas, yaitu λ yang dipangkatkan terhadap Y dengan bentuk transformasi Y^{λ}. Pendugaan parameter λ dapat dilakukan dengan Metode Kemungkinan Maksimum, dengan tujuan mendapatkan jumlah kuadrat sisaaan yang minimum.

Daftar Pustaka

[1] Bain, L.J dan M. Engelhardt. 1997. *Introduction to Probability and Mathematical Statistics.* Second Edition. PWS-KENT, Boston

[2] Drapper, N.R dan H. Smith. 1992. *Analisis Regresi Terapan.* PT. Gramedia, Jakarta

[3] Ispriyanti, D. 2004. Pemodelan Statistika dengan Transformasi Box-Cox. Jurnal Matematika dan Komputer. Vol.7 No.3

[4] Natrella, M. 2012. *NIST/SEMATECH e-Handbook os Statistical Method.* U.S Commerce Department's Technology Administration, USA

[5] Rawling, J.O, S.G Pantula dan D.A Dickey. 1998. *Applied Regression Analysis : A Research Tool.* Second Edition. Springer-Verlag, New York

[6] Siegel, S.1992. *Statistik Nonparametrik.* PT. Gramedia, Jakarta