

# PENERAPAN METODE TRANSFORMASI LOGARITMA NATURAL DAN PARTIAL LEAST SQUARES UNTUK MEMPEROLEH MODEL BEBAS MULTIKOLINIER DAN OUTLIER

**Margaretha Ohyver**

Mathematics & Statistics Department, School of Computer Science, Binus University  
Jl. K.H. Syahdan No. 9, Palmerah, Jakarta Barat 11480  
mohyver@binus.edu

## ABSTRACT

*Multicollinear and outlier occur when making regression modeling. Multicollinear leads difficulty in separating the influence of each independent variable on the response variable. Outlier causes unmet assumption of normality in the regression. Both cases occur in the number of hotel visitors in Kendari. The purpose of this paper is to find a model that is free from multicollinear and outlier. Using the natural logarithm transformation and partial least squares, obtained model has the value of variance inflation factor less than ten and is able to overcome the outlier.*

**Keywords:** multicollinear, outlier, partial least squares, natural logarithm transformation

## ABSTRAK

*Multikolinier dan outlier merupakan kasus yang sering terjadi ketika membuat pemodelan menggunakan regresi. Multikolinier menyebabkan kesulitan dalam memisahkan pengaruh masing-masing variabel bebas terhadap variabel respon. Sementara outlier menyebabkan tidak terpenuhinya asumsi normalitas dalam regresi. Kedua kasus ini terjadi dalam data tingkat penghunian kamar hotel di Kendari. Oleh karena itu, studi ini bertujuan untuk menemukan model untuk data tersebut yang bebas dari multikolinier dan outlier. Dengan menggunakan transformasi logaritma natural dan partial least squares diperoleh model yang nilai variance inflation factor kurang dari sepuluh dan sudah dapat mengatasi adanya outlier.*

**Kata kunci:** multikolinier, outlier, partial least squares, transformasi logaritama natural

## PENDAHULUAN

Pemodelan dengan regresi telah banyak digunakan mulai dari bidang sosial, ekonomi, kimia, kesehatan, dan sebagainya. Dengan model regresi yang dihasilkan, dapat diketahui variabel-variabel yang secara signifikan mempengaruhi variabel yang lain. Untuk bisa memperoleh variabel-variabel yang berpengaruh tersebut maka model yang diperoleh harus dapat memenuhi asumsi-asumsi yang berlaku di dalam regresi.

Penelitian mengenai pelanggaran asumsi telah banyak dilakukan. Ohlyver (2011) menggunakan metode regresi ridge untuk mengatasi multikolinier yang terjadi pada variabel-variabel yang mempengaruhi produksi kol bulat. Ashok, Mitra, dan Mitra (2005) melakukan pendeteksian *multiple outlier* menggunakan *self-organizing maps* title. Francisco-Fernandez dan Vilai-Fernandez (2008) melakukan pengujian adanya heteroskedastisitas dalam regresi non parametrik.

Makalah ini akan membahas mengenai penerapan dari *partial least squares* (PLS) yang dikombinasikan dengan transformasi logaritma natural. Penerapan ini dimaksudkan untuk mengatasi kasus multikolinier serta memenuhi asumsi normalitas. Kombinasi PLS dan transformasi logaritma natural diterapkan pada data pengunjung hotel di Kendari, Sulawesi Tenggara.

Sulawesi Tenggara (Sultra) merupakan salah satu provinsi di Indonesia. Sultra ditetapkan sebagai daerah otonom berdasarkan Perpu No. 2 tahun 1964 Jungto UU No. 13 Tahun 1964. Pada awalnya terdiri atas empat kabupaten dan kini setelah pemekaran Sultra telah mempunyai sepuluh kabupaten dan dua kota, di mana ibukotanya terletak di kota Kendari. Tanggal 27 April 1964 merupakan hari lahir Provinsi Daerah Tingkat I Sultra (Biro Humas Sultra, 2011).

Sultra termasuk dalam provinsi berkembang. Hal ini terlihat dari pertumbuhan ekonomi daerah yang cukup tinggi, yaitu di atas 8%. Angka kemiskinan di Sultra turun dari 21% menjadi 14,6% pada tahun 2011 (Investor Daily Indonesia, 2012). Salah satu faktor yang dapat dianggap berperan adalah adanya perkembangan dalam sektor jasa perhotelan. Berdasarkan data resmi Dinas Tenaga Kerja dan Transmigrasi Sultra per Desember 2011 menunjukkan bahwa sektor perhotelan dan rumah makan di Sultra telah menyerap 8791 tenaga kerja lokal (Media Sultra, 2011).

Jasa perhotelan merupakan bagian dari industri pariwisata yang sangat penting dalam pembangunan ekonomi nasional maupun regional (BPS Provinsi Sultra, 2011). Hal ini tidak terbantahkan mengingat peranannya yang besar terutama dalam hal ketenagakerjaan dan sumber pendapatan daerah. Oleh karena itu, sektor ini perlu mendapat perhatian khusus dari pemerintah. Berbagai upaya yang terencana, terpadu, dan efektif perlu dilakukan untuk menunjang kebijakan pemerintah dalam pembangunan di sektor ini. Akan tetapi selain pemerintah, pihak penyedia jasa perhotelan pun harus melakukan pengembangan. Sebab segala upaya yang dilakukan oleh pemerintah tidak akan ada artinya jika pihak perhotelan tidak melakukan pengembangan atau perbaikan dari dalam.

Pengembangan yang dilakukan oleh penyedia jasa perhotelan sangat penting. Sebab tenaga kerja tidak akan terserap dan investor pun tidak akan mau menanamkan modal jika penyedia jasa perhotelan memberikan pelayanan yang buruk. Pelayanan tersebut dapat berupa tersedianya fasilitas-fasilitas yang terbaik, tarif yang memadai, ketersediaan kamar, dan sebagainya. Hal ini perlu diperhatikan, terutama jika di daerah tersebut tersedia banyak hotel atau penginapan. Semakin banyaknya hotel, tingkat persaingan antar hotel pun tinggi.

Jumlah hotel/akomodasi di Sultra menunjukkan peningkatan yang cukup baik. Hal ini terlihat dari jumlahnya yang terus meningkat sejak tahun 2006 hingga tahun 2010 dengan peningkatan sebesar

4,65% per tahun. Demikian pula untuk jumlah kamar dalam rentang waktu yang sama dengan rata-rata peningkatan sebesar 9,76%. Di samping itu, untuk jumlah tempat tidur meningkat rata-rata sebesar 4,45%. Dari hasil inventarisasi perusahaan akomodasi di Sultra tahun 2010, terdapat 265 buah perusahaan/usaha akomodasi dengan 3389 kamar dan 4918 tempat tidur. Dari 265 hotel hanya ada satu hotel berbintang (bintang satu) dan 264 hotel non bintang. Meskipun termasuk kecil jika dibandingkan dengan provinsi-provinsi lain yang ada di Indonesia, jumlah tersebut akan cukup memberikan banyak pilihan kepada pengunjung. Perusahaan akomodasi tentu tidak ingin kehilangan pengunjung. Oleh karena itu perusahaan perlu mengetahui faktor-faktor yang mempengaruhi jumlah pengunjung hotel. Dengan mengetahui faktor-faktor tersebut maka perusahaan dapat memberikan perhatian khusus sehingga ke depannya jumlah pengunjung dapat bertambah.

Berdasarkan latar belakang yang telah dijelaskan, tujuan dari makalah ini adalah memperoleh model yang bebas dari multikolinier dan memenuhi asumsi normalitas untuk data pengunjung hotel di Kendari.

## METODE

Data yang akan digunakan adalah data sekunder. Data ini diperoleh dari katalog Badan Pusat Statistik (BPS) Sultra tahun 2011. Ada 85 (delapan puluh lima) hotel/penginapan yang menjadi sampelnya. Hotel/penginapan tersebut berada di wilayah Kendari. Variabel-variabel yang digunakan terdapat pada Tabel 1. Terdapat tujuh variabel bebas ( $X$ ) dan satu variabel tak bebas atau variabel respon ( $Y$ ).

Langkah-langkah pengolahan data adalah sebagai berikut. Pertama, membentuk model regresi linier. Kedua, mengecek nilai VIF dengan tujuan mengetahui adanya multikolinier dalam data. Ketiga, membentuk model PLS. Keempat, mentransformasi data dengan transformasi logaritma natural. Kelima, membentuk model PLS baru untuk data yang telah ditransformasi. Keenam, membandingkan model-model yang telah diperoleh.

Tabel 1. Variabel Penelitian

No	Nama variabel	Simbol
1.	Usia hotel/penginapan	( $X_1$ )
2.	Tarif minimal hotel/penginapan	( $X_2$ )
3.	Tarif maksimal hotel/penginapan	( $X_3$ )
4.	Fasilitas yang dimiliki hotel atau penginapan	( $X_4$ )
5.	Jumlah tenaga kerja yang dimiliki hotel/penginapan	( $X_5$ )
6.	Jumlah kamar yang dimiliki hotel/penginapan	( $X_6$ )
7.	Jumlah tempat tidur yang dimiliki hotel/penginapan	( $X_7$ )
8.	Jumlah tamu	( $Y$ )

### Partial Least Squares (PLS)

*Partial least squares* (PLS) merupakan salah satu metode yang dapat digunakan untuk mengatasi masalah multikolinier. PLS merupakan perpaduan antara *principal component analysis* (PCA) dan regresi linier ganda (Abdi, 2003). PCA merupakan suatu metode untuk mereduksi banyaknya peubah bebas menjadi beberapa peubah baru yang dapat menjelaskan dengan baik keragaman data dan peubah baru tersebut tidak berkorelasi. Untuk membentuk hubungan antara peubah respon dan peubah bebas, PLS membentuk peubah bebas yang baru yang disebut faktor, peubah laten, atau komponen, di mana masing-masing komponen yang terbentuk merupakan

kombinasi linier dari peubah-peubah bebas. Tujuan utama dari PLS adalah membentuk komponen yang dapat menangkap informasi dari peubah bebas untuk menduga peubah respon (Hoskuldsson dalam Garthwaite, 1994).

Jika ( $n < p$ ), metode-metode kuadrat terkecil tidak dapat digunakan (Naes, Isaksson, dkk., 2002). Hal ini dikarenakan matriks  $\mathbf{X}^T\mathbf{X}$  singular. Sebaliknya, PLS dapat digunakan untuk kasus ( $n < p$ ). Regresi PLS didasarkan pada dekomposisi komponen:

$$\mathbf{Y} = \mathbf{TQ}' + \mathbf{F} \quad (1)$$

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \quad (2)$$

dengan  $\mathbf{T}$  adalah matriks komponen,  $\mathbf{P}$  dan  $\mathbf{Q}$  adalah matriks *loading*  $X$  dan  $Y$ ,  $\mathbf{E}$  dan  $\mathbf{F}$  adalah vektor *error* (Boulesteix dan Strimmer, 2006).

Metode PLS dapat dipandang sebagai metode yang membentuk matriks komponen  $\mathbf{T}$  sebagai transformasi linier dari  $\mathbf{X}$ .

$$\mathbf{T} = \mathbf{XW} \quad (3)$$

dengan  $\mathbf{W}$  adalah matriks *weights* (bobot). Persamaan (3) dapat dituliskan sebagai berikut.

$$T_1 = w_{11}x_1 + w_{21}x_2 + \dots + w_{p1}x_p$$

$$T_2 = w_{12}x_1 + w_{22}x_2 + \dots + w_{p2}x_p$$

⋮

$$T_c = w_{1c}x_1 + w_{2c}x_2 + \dots + w_{pc}x_p$$

Komponen-komponen kemudian digunakan untuk pendugaan, dengan menggantikan  $\mathbf{X}$ . Sehingga diperoleh penduga kuadrat terkecil:

$$\hat{\mathbf{Q}}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y}. \quad (4)$$

## HASIL DAN PEMBAHASAN

Data yang digunakan adalah data perhotelan di Kendari provinsi Sultra. Akan digunakan 85 pengamatan untuk pembentukan model. Pemodelan akan dilakukan dengan metode regresi, *partial least squares* (PLS), dan gabungan antara transformasi logaritma natural dan PLS.

Penentuan penggunaan PLS didasarkan pada anggapan sementara bahwa antar variabel bebas saling berkorelasi. Sebagai contoh, antara jumlah kamar dan jumlah tenaga kerja. Semakin banyak jumlah kamar yang dimiliki oleh sebuah hotel maka semakin banyak jumlah tenaga kerja yang terdapat di hotel tersebut. Atau antara jumlah fasilitas dan tarif maksimal. Semakin lengkapnya fasilitas yang dimiliki oleh sebuah hotel maka semakin tinggi pula tarif maksimal yang diberikan. Berdasarkan contoh yang diberikan maka digunakan PLS di mana berdasarkan beberapa penelitian diketahui metode ini sangat baik dalam mengatasi adanya korelasi antar variabel bebas.

Penggunaan transformasi logaritma natural didasarkan pada identifikasi awal. Berdasarkan identifikasi tersebut diketahui bahwa hanya dengan transformasi logaritma natural, asumsi normalitas untuk data yang digunakan menjadi terpenuhi.

## Pemodelan dengan Partial Least Squares (PLS)

Secara umum metode yang dapat digunakan untuk mengetahui faktor-faktor yang mempengaruhi jumlah tamu pada hotel yang ada di Kendari dapat menggunakan analisis regresi. Akan tetapi hal tersebut akan menghasilkan kesimpulan yang salah jika asumsi-asumsi yang ada dalam regresi tidak dipenuhi. Salah satu asumsinya adalah tidak terjadi multikolinier. Pada data perhotelan Kendari terjadi pelanggaran terhadap asumsi tersebut.

Hasil pemodelan regresi dapat dilihat pada Tabel 2. Pada tabel tersebut diketahui bahwa yang berpengaruh signifikan hanya  $X_3$  dan  $X_5$ , dalam hal ini tarif maksimal dan jumlah tenaga kerja. Selain itu diketahui juga bahwa terjadi multikolinier pada data, hal ini diketahui dari adanya nilai VIF yang lebih dari 10 ( $VIF > 10$ ).

Tabel 2. Hasil Regresi untuk Data Hotel di Kendari

Variabel	Nilai Dugaan Koefisien Regresi	P value	VIF
Konstanta	1107,6	0,003	
$X_1$	-21,18	0,148	1,110
$X_2$	-7,429	0,016	2,996
$X_3$	5,330	0,000	3,983
$X_4$	-10,93	0,886	2,899
$X_5$	39,08	0,218	10,080
$X_6$	36,64	0,207	9,915
$X_7$	-29,02	0,125	12,161

Adanya multikolinier ini memberikan akibat adanya nilai dugaan untuk koefisien-koefisien regresi yang berubah-ubah. Nilai-nilai tersebut berubah tergantung pada jumlah variabel bebas yang bergabung dalam pembentukan model. Misal untuk membentuk model digunakan tiga variabel bebas dan lima variabel bebas. Nilai-nilai dugaan untuk koefisiennya dapat dilihat pada Tabel 3. Jika dilakukan perbandingan antara Tabel 2 dan Tabel 3, akan terlihat adanya perbedaan nilai untuk masing-masing variabel. Dengan demikian tidak dapat diperoleh kesimpulan mengenai besar pengaruh masing-masing variabel bebas terhadap variabel respon.

Tabel 3. Hasil Regresi dengan 3 dan 5 Variabel Bebas

Model dengan 3 (tiga) variabel bebas		Model dengan 5 (tiga) variabel bebas	
Variabel	Nilai Dugaan Koefisien Regresi	Variabel	Nilai Dugaan Koefisien Regresi
Konstanta	1114	Konstanta	1117
$X_1$	-25,9	$X_1$	-25,2
$X_2$	-6,15	$X_2$	-6,65
$X_3$	5,91	$X_3$	5,27
		$X_4$	4,1
		$X_5$	20,4

Pemodelan dengan menggunakan metode PLS dapat dilihat pada Tabel 4. Pada tabel tersebut diperoleh nilai PRESS yang paling kecil adalah pada penggunaan satu komponen. Persamaan regresi dengan menggunakan satu komponen adalah sebagai berikut.

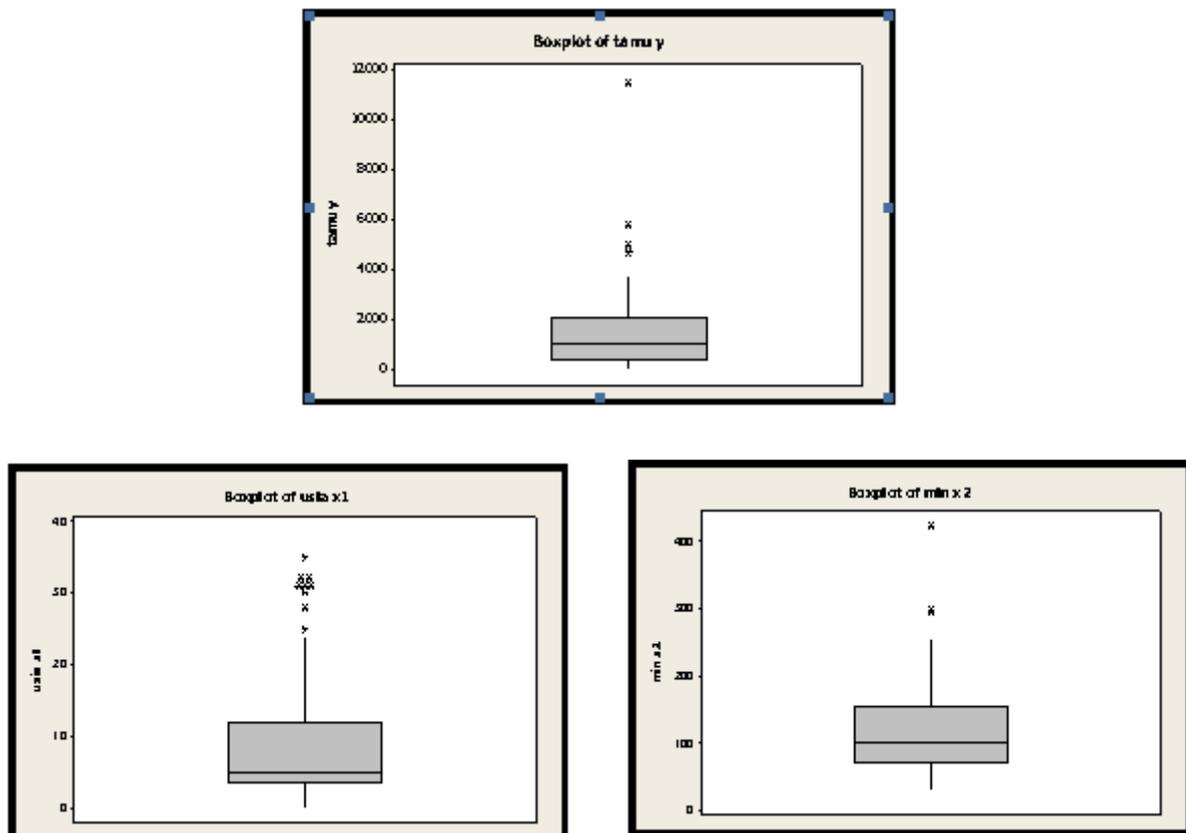
$$y = -0,000 + 0,336T_1 \quad (5)$$

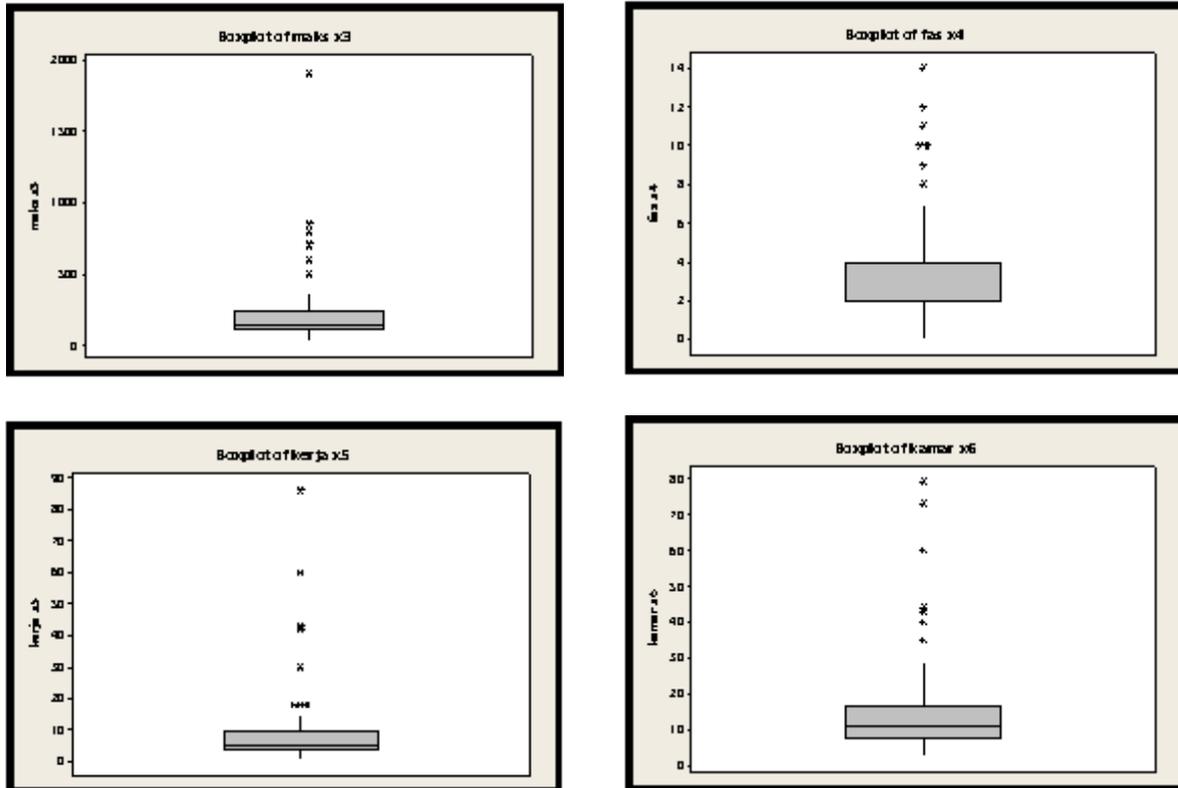
Jika dilakukan perbandingan antara pemodelan dengan menggunakan regresi linier ganda dan PLS, diperoleh hasil sebagai berikut. Nilai VIF yang diperoleh dengan menggunakan PLS sudah lebih kecil jika dibandingkan dengan menggunakan regresi linier ganda. Dengan demikian, masalah multikolinier telah dapat diatasi. Sekarang yang menjadi perhatian adalah nilai  $R^2$ . Nilai  $R^2 = 56\%$  untuk regresi linier ganda sedangkan nilai  $R^2 = 43\%$  untuk PLS. Mengapa nilai  $R^2$  untuk pemodelan dengan regresi linier ganda lebih baik daripada pemodelan dengan PLS? Yang menjadi kecurigaan adalah adanya outlier pada data ini.

Tabel 4. Hasil Partial Least Squares untuk Data Hotel di Kendari

Banyaknya Komponen	% Keragaman	PRESS
1	43,92	58,80
2	53,30	71,97
3	54,32	70,91
4	54,60	69,56
5	54,62	71,71
6	54,65	72,64

Outlier pada data perhotelan dapat dilihat pada Gambar 1. Pada gambar tersebut terlihat ada outlier pada setiap variabel. Adanya outlier ini juga mengakibatkan tidak terpenuhinya asumsi normalitas. Baik menggunakan regresi linier ganda maupun PLS, asumsi tersebut tidak terpenuhi. Gambar 2 dan Gambar 3 menunjukkan hal tersebut.





Gambar 1. Boxplot data perhotelan.

Untuk mengatasi hal tersebut maka nilai variabel respon akan ditransformasi dengan menggunakan transformasi logaritma natural. Hasil pemodelan regresi untuk variabel respon yang telah ditransformasi dapat dilihat pada Tabel 5.

Tabel 5. Regresi dengan Variabel Respon yang Ditransformasi

Variabel	Nilai Dugaan Koefisien Regresi	P value	VIF
Konstanta	6,6841	0,000	
$X_1$	-0,02282	0,068	1,110
$X_2$	-0,004028	0,121	2,996
$X_3$	0,0011513	0,184	3,983
$X_4$	0,05266	0,418	2,899
$X_5$	0,04297	0,113	10,080
$X_6$	0,05520	0,027	9,915
$X_7$	-0,04134	0,011	12,161

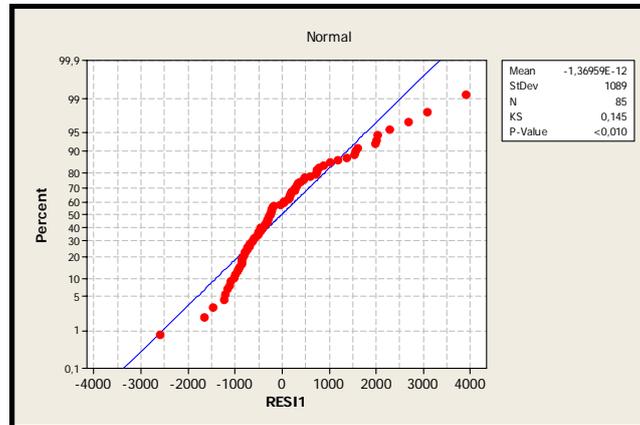
Dengan menggunakan PLS diperoleh lima komponen utama, di mana model regresinya dapat dilihat sebagai berikut:

$$\ln Y = 6,75 + 0,258T_1^* + 0,271T_2^* + 0,401T_3^* + 0,325T_4^* + 0,356T_5^* \quad (6)$$

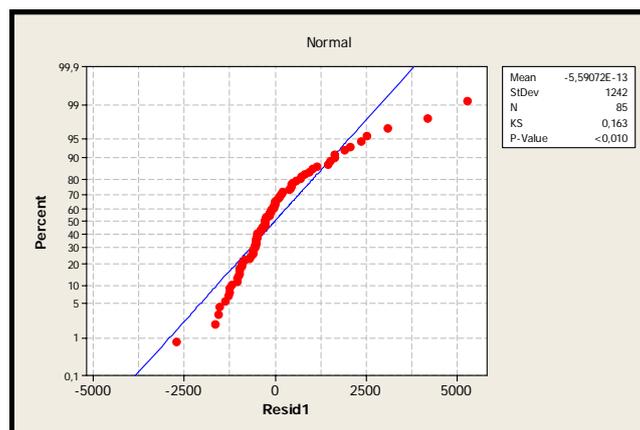
Jika dilihat pada nilai  $R^2$ , tidak ada perubahan yang cukup signifikan. Hal ini terlihat pada nilai  $R^2$  untuk regresi linier ganda adalah 34% dan  $R^2$  untuk PLS adalah 34%. Nilai  $R^2$  tidak menunjukkan bahwa setelah dimodelkan dengan PLS meningkat. Bahkan hasil yang diperoleh nilai  $R^2$

sama untuk regresi linier ganda dan PLS. Akan tetapi asumsi kenormalannya menjadi terpenuhi. Hal ini dapat dilihat pada Gambar 2 – 5.

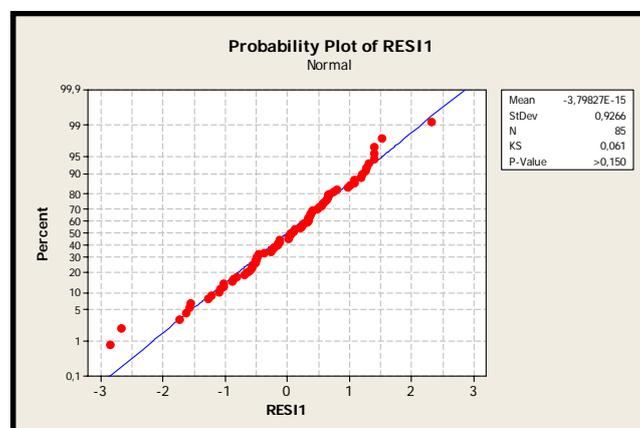
Berdasarkan hasil yang diperoleh di atas, dapat dikatakan bahwa data ini masih perlu untuk dikaji lebih lanjut. Tidak tertutup kemungkinan diaplikasikan metode-metode lain yang dapat meningkatkan nilai  $R^2$ .



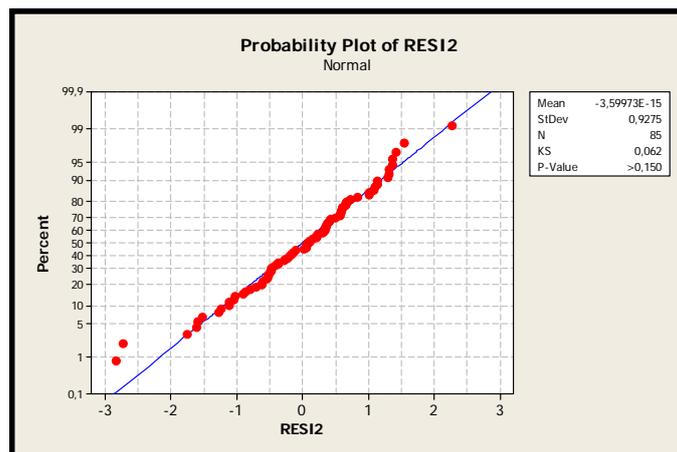
Gambar 2. Plot untuk uji normalitas pada regresi linier ganda.



Gambar 3. Plot untuk uji normalitas pada *partial least squares*.



Gambar 4. Plot untuk uji normalitas pada regresi linear ganda dengan  $Y = \ln Y$ .



Gambar 5. Plot untuk uji normalitas pada *partial least squares* dengan  $Y = \ln Y$

## SIMPULAN

Berdasarkan penelitian yang telah dilakukan maka disimpulkan sebagai berikut. Pertama, dengan menggunakan gabungan antara transformasi logaritma natural dan *partial least squares* diperoleh model yang bebas multikolinier dan outlier. Kedua, model yang diperoleh masih mempunyai nilai  $R^2$  yang kecil.

## DAFTAR PUSTAKA

- Abdi, H. (2003). Partial Least Squares (KTP) Regression. *Encyclopedia of Social Sciences Research Methods* (online), 1-7. Diakses dari [www.utdallas.edu/~herve](http://www.utdallas.edu/~herve).
- Ashok, K. N., Mitra, A., & Mitra, S. (2005). Multiple outlier detection in multivariate data using self-organizing maps title. *Computational Statistics*, 20(2): 245-264. doi: <http://dx.doi.org/10.1007/BF02789702>.
- Biro Humas Pemerintahan Provinsi Sulawesi Tenggara. (2012). Profil Sultra.
- Boulesteix, A. and Strimmer, K. (2006). *Partial Least Squares: A Versatile Tool For The Analysis Of High-Dimensional Genomic Data*. Diakses dari <http://www.slcmr.net/boulesteix/papers/review>
- Garthwaite, P. H. (1994). An Interpretation of Partial Least Squares. *Journal of the American Statistical Association*, 89: 122-127.
- Investor Daily Indonesia. (2012). *Pertumbuhan Ekonomi Sultra Lampaui Nasional*. Diakses dari <http://www.investor.co.id/home/pertumbuhan-ekonomi-sultra-lampaui-nasional/28618>.
- Katalog BPS. (2011). *Direktori dan Tingkat Penghunian Kamar Hotel Provinsi Sulawesi Tenggara Tahun 2011*. Kendari: Badan Pusat Statistik Provinsi Sulawesi Tenggara.

- Mario Francisco-Fernández, & Juan, M. V. (2009). Two tests for heteroscedasticity in nonparametric regression. *Computational Statistics*, 24(1), 145-163. doi: <http://dx.doi.org/10.1007/s00180-008-0110-3>
- Media Sultra. (2011). *Sektor Perhotelan Sultra Serap 8791 Tenaga Kerja*. Diakses dari <http://sindikasi.inilah.com/read/detail/1812732/sektor-perhotelan-sultra-serap-8791-tenaga-kerja>.
- Naes, T., Isaksson, T., Fearn, T., dan Davies, T. (2002). *Multivariate Calibration and Classification*. Chichester: NIR Publications.
- Ohyver, M. (2011). Metode Ridge untuk Mengatasi Kasus Multikolinear. *Comtech*, 2(1): 451-457.