Implementasi *Partial Least Square* Dalam Pemodelan Indeks Pembangunan Manusia

Andrew Donda Munthe [1)], Agung Eddy Suryo Saputro [2)]

[1)]*Departemen Statistik, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor, email:dondaandrew@gmail.com*

[2)]*Departemen Statistik, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor, e-mail: agung.eddyss@gmail.com*

Abstrak

Kondisi pembangunan manusia di suatu wilayah dapat digambarkan dengan perhitungan Indeks Pembangunan Manusia (IPM). Komponen yang membentuk IPM cenderung memiliki korelasi yang kuat antara satu dan yang lain, kondisi ini sering disebut multikolinieritas. Penelitian ini bertujuan untuk mengatasi masalah multikolinieritas dalam data IPM Provinsi DKI Jakarta pada periode 2010-2016. Penerapan metode PLS dalam penelitian ini mampu mengatasi masalah multikolinieritas dan membuat asumsi *Gauss Markov* dalam regresi linier berganda terpenuhi.

Kata Kunci. Asumsi *Gauss Markov*, IPM, Multikolinearitas, PLS

Abstract

The condition of human development in a region can be described by calculating the Human Development Index (HDI). The components that make HDI tend to have a strong correlation between one and the other, this condition is often called multicollinearity. This study aims to overcome the problem of multicollinearity in the HDI data of DKI Jakarta Province for the period 2010-2016. The application of the PLS method in this study is able to overcome the problem of multicollinearity and make the Gauss Markov assumption in multiple linear regression fulfilled.

Keywords. *Gauss Markov Assumption, HDI, Multicollinearity, PLS*

1. Pendahuluan

Berbagai persoalan hingga saat ini masih dihadapi oleh pemerintah Indonesia dalam melaksanakan pembangunan manusia. Mulai dari ketersediaan sarana insfrastruktur yang terbatas, kemiskinan, lapangan kerja, hingga permasalahan pelayanan publik kepada masyarakat. Salah satu cara untuk mengukur kualitas pembangunan manusia adalah dengan melakukan perhitungan Indeks Pembangunan Manusia (IPM). Noorbakhsh (1998) menyatakan bahwa indeks ini berfungsi sebagai alat ukur kualitas pembangunan dari sisi fisik (kesehatan dan tingkat penghasilan) serta non fisik (pendidikan).

Tahun 1990, *United Nations Development Programme* (UNDP) memperkenalkan perhitungan IPM dan kemudian melakukan pembaharuan metode perhitungannya di tahun 2010. Badan Pusat Statistik (BPS) mengadopsi perhitungan IPM oleh UNDP untuk diterapkan di Indonesia sejak tahun 2010. Terdapat tiga komponen yang menjadi dasar perhitungan IPM yaitu terkait dengan dimensi kesehatan (*a long and healthy life*), dimensi pendidikan (*knowledge*), dan juga dimensi pengeluaran (*decent standard of living*).

Provinsi DKI Jakarta sebagai Ibukota Indonesia merupakan barometer pembangunan manusia secara nasional. Data BPS menunjukkan IPM DKI Jakarta dari tahun ke tahun selalu merupakan yang tertinggi dibandingkan 33 provinsi lain di Indonesia. Tahun 2016, IPM DKI Jakarta telah mencapai 79.60. Nilai IPM provinsi ini bahkan lebih tinggi dari angka IPM Nasional yang hanya sebesar 70.18. (BPS, 2016)

Komponen-komponen pembentuk IPM cenderung saling berkorelasi tinggi atau disebut dengan multikolinearitas (Putra dan Ratnasari, 2015). Adanya multikolineritas membuat penduga berbias dan hasil regresi menjadi sulit diinterpretasikan. Selain multikolineritas, asumsi *Gauss Markov* harus terpenuhi dalam regresi linear berganda. Jarque dan Bera (1980) menyatakan bahwa asumsi *Gauss Markov* terdiri dari galat yang menyebar normal (normalitas), ragam sisaan homogen (homoskedastisitas), dan sisaan menyebar bebas (non autokorelasi).

Salah satu solusi dalam mengatasi multikolinearitas dalam regresi berganda adalah dengan menerapkan *Partial Least Square* (PLS). Penerapan PLS memiliki beberapa keunggulan dibandingkan dengan metode lain. Pertama, PLS dapat diterapkan pada data yang tidak mengikuti sebaran normal (Helland, 1990). Kedua, PLS dapat digunakan ketika hubungan antar peubah prediktor memiliki dasar teori yang lemah. Ketiga, metode PLS juga dapat digunakan meskipun jumlah data yang diteliti berukuran kecil (Kosasih dan Budiani 2007). Oleh karena itu, penelitian ini bertujuan untuk mengimplementasikan metode PLS dalam mengatasi multikolinearitas pada data IPM di Provinsi DKI Jakarta periode 2010-2016.

2. Metodologi

2.1. Data

Data IPM Provinsi DKI Jakarta bersumber dari BPS. Periode data yang digunakan adalah tahun 2010-2016. Pengunaan periode waktu ini karena data IPM tahun 2010-2016 telah mengadopsi metode perhitungan IPM terbaru dengan tahun dasar 2010. Ada satu peubah respon dan empat peubah prediktor yang dimodelkan dalam penelitian. Tabel 1 menunjukkan secara lebih lengkap mengenai peubah-peubah yang digunakan.

Pengolahan dan analisis data menggunakan bantuan *software* statistik R versi 3.5.1, *Minitab,* dan *Microsoft Excel*.

Tabel 1. Peubah-peubah yang digunakan dalam penelitian

Peubah	Nama Peubah	Dimensi Komponen IPM
y	Indeks Pembangunan Manusia	
x_1	Angka Harapan Hidup Saat Lahir (AHH)	Kesehatan
x_2	Harapan Lama Sekolah (HLS)	Pendidikan
x_3	Rata-rata Lama Sekolah (RLS)	Pendidikan
x_4	Pengeluaran per kapita disesuaikan	Pengeluaran

Sumber: BPS

2.2. Multikolinearitas

Keberadaan multikolinearitas dalam data yang diteliti dapat diketahui dengan melakukan perhitungan nilai *Variance Inflation Factor* (VIF). Rumus perhitungan VIF adalah sebagai berikut:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{1}$$

Dengan R_j^2 merupakan koefisien determinasi antara peubah prediktor, dimana j=1,2,…,p. Apabila perhitungan nilai VIF > 10 maka terjadi multikolinearitas.

2.3. *Partial Least Square* (PLS)

Tujuan PLS adalah untuk memprediksi peubah respon dengan komponen yang dibentuk dari peubah prediktor yang signifikan berpengaruh pada komponen tersebut. Model PLS dapat dirumuskan sebagai berikut (Bastien, dkk., 2005):

$$y = \sum_{h-1}^{m} c_h t_h + \varepsilon \tag{2}$$

$$t_h = \sum_{j-1}^{p} w_{(h)j} X_j \tag{3}$$

dengan,

y = peubah respon

c_h = koefisien regresi y terhadap t_h

t_h = komponen utama ke-*h* dan tidak saling berkorelasi (*h=1,2,…,m*)

Kondisi t_h orthogonal merupakan syarat pembentukan komponen PLS sehingga parameter c_h dan w_h dalam persamaan dapat di estimasi. Kombinasi linier antara peubah prediktor X_j dan koefisien pembobot w_1 membentuk komponen PLS yang pertama (t_1) dengan rumus sebagai berikut:

$$t_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \left(cor\left(x_j, y\right)\right)^2}} \sum_{j=1}^{p} \left(cor\left(x_j, y\right)\right) x_j^* \tag{4}$$

dimana,

x_j^* = x_j yang tersandardisasi

$cor\left(x_j, y\right)$ = korelasi variabel x_j dengan y

Langkah selanjutnya, melakukan regresi peubah y terhadap t_1 dan setiap x_j sebelum masuk ke tahapan perhitungan komponen PLS yang kedua (t_2).

Peubah x_j yang digunakan hanya peubah yang berkontribusi secara nyata dalam menjelaskan *y* pada t_1. Persamaan komponen utama yang kedua dapat dituliskan sebagai berikut:

$$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p}\left(cor\left(y, x_{1j}\right)\right)^2}} \sum_{j=1}^{p}\left(cor\left(y, x_{1j}\right)\right)x_{1j}^{*} \qquad (5)$$

dimana,

x_{1j}^{*} = residu tersandardisasi dari regresi x_j terhadap t_1.

Langkah selanjutnya adalah membentuk komponen PLS ke-*h* (t_h). Langkah ini dilakukan sama seperti pada perhitungan komponen PLS sebelumnya. Model regresi *y* terhadap t_1, t_2, …, t_{h-1} dan setiap x_j adalah sebagai berikut:

$$y = c_1 t_1 + c_2 t_2 + \ldots + c_{h-1} t_{h-1} + a_{hj} x_j + residu \qquad (6)$$

Untuk memperoleh komponen t_h yang orthogonal terhadap t_{h-1} maka dilakukan regresi x_j terhadap komponen PLS dengan persamaan:

$$x_j = p_{1j} t_1 + p_{2j} t_2 + \ldots + p_{h-1j} t_{h-1} + x_{(h-1)j} \qquad (7)$$

dengan,

$x_{(h-1)j}$ = residu dari regresi setiap x_j pada t_1, t_2, …, t_{h-1}.

Komponen PLS ke-*h* (t_h) dirumuskan sebagai berikut:

$$t_h = \frac{1}{\sqrt{\sum_{j=1}^{p}\left(cor\left(x_{(h-1)j}, y\right)\right)^2}} \sum_{j=1}^{p}\left(cor\left(x_{(h-1)j}, y\right)\right)x_{(h-1)j}^{*} \qquad (8)$$

dengan,

$x_{(h-1)j}^{*}$ = residu standar hasil regresi setiap x_j terhadap t_1, t_2, …, t_{h-1}.

Perhitungan komponen PLS ke-*h* (t_h) tidak dilanjutkan lagi apabila tidak ada lagi prediktor yang signifikan untuk membangun komponen PLS ke-*h* tersebut.

3. Hasil dan Pembahasan

3.1. IPM Provinsi DKI Jakarta

Periode tahun 2010-2016, IPM Provinsi DKI Jakarta dan komponen-komponen penyusunnya terus mengalami peningkatan (Tabel 2). Hasil persamaan regresi yang terbentuk berdasarkan data pada Tabel 2 adalah sebagai berikut:

$$\hat{y} = \mathbf{1.0 + 0.592}\boldsymbol{AHH} + \mathbf{1.004}\boldsymbol{HLS} + \mathbf{1.1825}\boldsymbol{RLS} + \mathbf{0.0006}\boldsymbol{Pengeluaran}$$

Tabel 2. Perkembangan IPM di Provinsi DKI Jakarta, Tahun 2010 - 2016

Tahun	IPM	Dimensi dan Komponen Penyusun IPM			
		AHH (Tahun)	HLS (Tahun)	RLS (Tahun)	Pengeluaran (ribu rupiah)
2010	76.31	71.71	11.86	10.37	15.111
2011	76.98	71.87	11.91	10.40	15.943
2012	77.53	72.03	11.96	10.43	16.613
2013	78.08	72.19	12.24	10.47	16.828
2014	78.39	72.27	12.38	10.54	16.898
2015	78.99	72.43	12.59	10.70	17.075
2016	79.60	72.49	12.73	10.88	17.468

Sumber : BPS

3.2. Penerapan *Partial Least Square* (PLS)

Hasil persamaan awal regresi yang terbentuk dilakukan uji multikolinearitas. Perhitungan matriks korelasi dan VIF pada data IPM Provinsi DKI Jakarta periode tahun 2010 – 2016 ditunjukkan pada Tabel 3 dan Tabel 4. Korelasi antara peubah prediktor dalam penelitian (Tabel 3) mengindikasikan adanya permasalahan multikolinearitas. Hal ini dapat terlihat dari korelasi antara peubah bebas yang nilainya mendekati 1. Perhitungan VIF (Tabel 4), semua peubah prediktor memiliki nilai VIF > 10. Berdasarkan hasil perhitungan korelasi peubah prediktor dan VIF maka dapat dipastikan adanya multikolinearitas pada data IPM Provinsi DKI Jakarta periode 2010-2016.

Tabel 3. Korelasi antar peubah bebas (prediktor)

Nama Peubah	AHH	HLS	RLS	Pengeluaran
AHH	1.000	0.966	0.890	0.958
HLS	0.966	1.000	0.948	0.862
RLS	0.890	0.948	1.000	0.793
Pengeluaran	0.958	0.862	0.793	1.000

Sumber : Hasil pengolahan data

Tabel 4. Nilai *Variance Inflation Factor* (VIF) peubah prediktor

	Peubah Prediktor			
	AHH	HLS	RLS	Pengeluaran
VIF	256.89	136.34	15.25	62.42

Sumber : Hasil pengolahan data

Langkah awal penerapan PLS adalah melakukan pembentukan komponen pertama (t_1). Sebelumnya, dilakukan uji signifikansi setiap peubah x_j. Hasil ujinya seperti ditunjukkan pada Tabel 5.

Tabel 5. Uji Signifikansi masing-masing peubah bebas untuk pembentukan t_1

Peubah Bebas	Peubah bebas (prediktor)			
(Prediktor)	Koefisien	SE	T	P-Value
AHH	3.944	0.201	19.60	0.000
HLS	3.215	0.349	9.22	0.000
RLS	5.700	1.03	5.55	0.003
Pengeluaran	0.001375	0.000194	7.10	0.001

Sumber : Hasil pengolahan data

Hasil perhitungan komponen PLS pertama (t_1) dari persamaan (4) adalah sebagai berikut:

$$t_1 = \frac{0.994AHH^* + 0.972HLS^* + 0.928RLS^* + 0.954Pengeluaran^*}{\sqrt{0.994^2 + 0.972^2 + 0.928^2 + 0.954^2}}$$

$$t_1 = 0,514AHH^* + 0,509HLS^* + 0,489RLS^* + 0,487Pengeluaran^*$$

Substitusi nilai x_j^* ke persamaan di atas maka diperoleh nilai dari komponen PLS pertama (t_1) (Tabel 6).

Tabel 6. Nilai dari Komponen PLS pertama (t_1)

Observasi	1	2	3	4	5	6	7
t_1	-2.684	-1.724	-0.866	0.070	0.645	1.766	2.792

Sumber : Hasil pengolahan data

Setelah didapatkan komponen PLS pertama maka langkah selanjutnya adalah membentuk komponen PLS yang kedua (t_2). Hasil perhitungan komponen PLS kedua (t_2) dari persamaan (5) adalah sebagai berikut:

$$t_2 = 0.215AHH^* - 0.299HLS^* - 0.626RLS^* + 0.690Pengeluaran^*$$

Dari persamaan di atas maka diperoleh nilai dari komponen PLS yang kedua (t_2) seperti tertera pada Tabel 7.

Tabel 7. Nilai dari Komponen PLS kedua (t_2)

Observasi	1	2	3	4	5	6	7
t_2	-0.715	0.009	0.584	0.496	0.262	-0.170	-0.466

Sumber : Hasil pengolahan data

Langkah selanjutnya adalah membentuk komponen PLS ketiga. Hasil perhitungan diketahui bahwa semua peubah berkorelasi tinggi dengan t_2. Dengan demikian maka perhitungan komponen PLS ketiga tidak diperlukan sehingga hanya diperoleh dua komponen baru yaitu t_1 dan t_2. Setelah didapatkan komponen baru, kemudian peubah respon y diregresikan terhadap komponen baru tersebut. Hasil pengolahannya menghasilkan persamaan regresi linear sebagai berikut:

$$\hat{y} = \mathbf{77.9829 + 0.59181}\ t_1 + \mathbf{0.1080}\ t_2$$

Persamaan di atas dalam bentuk peubah prediktor menjadi sebagai berikut:

$$\hat{y} = 77.9829 + 0.59181\,(0.514AHH^* + 0.509HLS^* + 0.489RLS^* + 0.487Pengeluaran^*) + 0.1080\,(0.215AHH^* - 0.299HLS^* - 0.626RLS^* + 0.690Pengeluaran^*)$$

sehingga hasil akhir persamaan regresi yang terbentuk adalah sebagai berikut:

$$\hat{y} = \mathbf{77.9829 + 0.3276AHH^* + 0.2689HLS^* + 0.2217RLS^* + 0.363Pengeluaran^*}$$

3.3. Pengujian Asumsi *Gauss Markov*

Hasil akhir model regresi berganda dengan penerapan PLS kemudian dilakukan pemeriksaan asumsi *Gauss Markov*.

3.3.1. Asumsi Kenormalan

Plot *Quantile-Quantile* dapat terlihat secara jelas bahwa data tersebar mendekati garis normal (Gambar 1.A). Berdasarkan pengujian pada hasil akhir PLS didapatkan nilai *Kolmogorov Smirnov Test* sebesar 0.167 dan p-value sebesar 0.150. Dengan demikian maka keputusannya adalah H_0

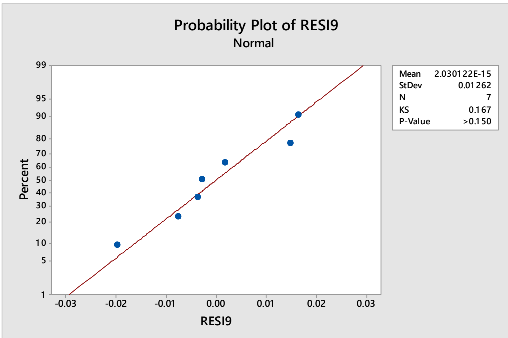
diterima, yang berarti data residual terdistribusi normal.

3.3.2. Heteroskedastisitas

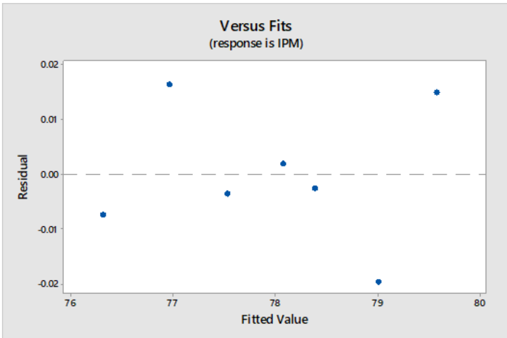
Plot antara residual dan nilai dugaan IPM tidak membentuk pola tertentu (Gambar 1.B). Hal ini mengindikasikan tidak terjadi heteroskedastisitas dalam model regresi. Untuk semakin menguatkan indikasi tersebut maka dilakukan perhitungan *Breusch-Pagan Test*. Hasil perhitungan didapatkan nilai sebesar 2.312 dengan nilai *p-value* sebesar 0.3147. Dengan nilai tersebut maka keputusannya adalah menerima H_0 sehingga dapat disimpulkan bahwa memang benar tidak terjadi heteroskedastisitas dalam model regresi.

3.3.3. Autokorelasi

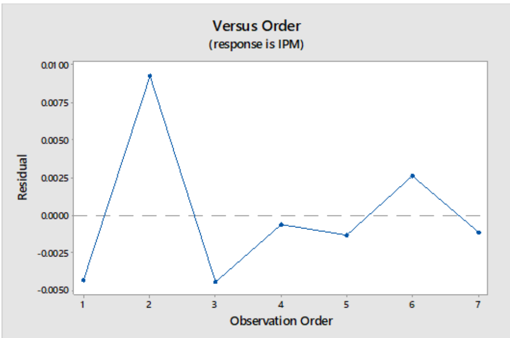
Plot antara residual dan waktu (*order*) tidak membentuk pola tertentu sehingga mengindikasikan tidak terjadi autokorelasi (Gambar 1.C). Untuk semakin menguatkan indikasi tersebut maka dilakukan perhitungan *Durbin Watson* dan diperoleh nilai 2.61742 dengan *p-value* sebesar 0.406. Dengan demikian maka keputusannya adalah H_0 diterima. Artinya, tidak terjadi autokorelasi.



A. Plot Quantile-Quantile



B. Plot residual dan nilai dugaan IPM



C. Plot antara residual dan *order*

Gambar 1 Plot asumsi *Gauss Markov*

3.3.4. Multikolinearitas

Tabel 8 mengindikasikan tidak adanya multikolinearitas antar peubah prediktor karena nilai korelasi yang sangat kecil (2.002 x 10^{-16}). Selain dengan matrik korelasi, multikolinearitas juga dapat dideteksi dengan perhitungan nilai VIF yang dapat dilihat pada Tabel 9. Nilai VIF untuk prediktor t_1 dan t_2 masing-masing adalah sebesar 1. Karena nilai VIF < 10 maka hasil dari penerapan PLS pada data IPM di Provinsi DKI Jakarta periode tahun 2010-2016 tidak lagi terjadi multikolinearitas.

Tabel 8. Korelasi antara t_1 dan t_2

	t_1	t_2
t_1	1.00	2.002 e-16
t_2	2.002 e-16	1.00

Sumber : Hasil pengolahan data

Tabel 9. Nilai *Variance Inflation Factor* (VIF) t_1 dan t_2

	t_1	t_2
VIF	1.00	1.00

Sumber : Hasil pengolahan data

Perhitungan nilai koefisien determinasi (R-square) pada model diperoleh nilai sebesar 99.99 persen. Artinya, sebesar 99.99 persen dari seluruh keragaman IPM pada data yang diteliti dapat dijelaskan oleh komponen-komponen pembentuknya.

4. Kesimpulan

Permasalahan multikolinearitas pada data Indeks Pembangunan Manusia (IPM) Provinsi DKI Jakarta periode tahun 2010–2016 dapat diatasi dengan penggunaan metode *Partial Least Square* (PLS). Selain itu, hasil akhir penerapan metode ini telah memenuhi semua uji asumsi *Gauss Markov* pada regresi linear berganda yaitu kenormalan, heteroskedastisitas, dan autokorelasi. Dengan demikian model yang dihasilkan oleh model PLS adalah model penaksir yang baik dengan nilai koefisien determinasi (R-square) yang tinggi.

Daftar Pustaka

Bastien P., Vinzi V.E., dan Tenenhaus, M. 2005. PLS generalised linear regression. *Computational Statistics and Data Analysis*, *48*, 17–46. https://doi.org/10.1016/j.csda.2004.02.005

BPS. 2016. *Berita Resmi Statistik Provinsi DKI Jakarta No. 19/04/31/Th. XVIX, 17 April 2017*.

Helland, I.S. 1990. Partial Least Squares Regression and Statistical Models. *Scandinavian Journal of Statistics*, *17*(2), 97–144.

Jarque, C.M., dan Bera, A.K. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, *6*(3), 255–259. https://doi.org/10.1016/0165-1765(80)90024-5

Kosasih N., dan Budiani, S. 2007. Pengaruh Knowledge Management Terhadap Kinerja Karyawan: Studi Kasus Departemen Front Office Surabaya Plaza Hotel. *Jurnal Manajemen Perhotelan*, *3*(2), 80–88. https://doi.org/10.9744/jmp.3.2.80-88

Noorbakhsh, F. 1998. The Human Development Index: Some Technical Issues and Alternative Indices. *Journal of International Development*, *10*, 589–605.

O'Brien, R.M. 2007. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality and Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Putra, D.M., dan Ratnasari, V. 2015. Pemodelan Indeks Pembangunan Manusia (IPM) Provinsi Jawa Timur Dengan Menggunakan Metode Regresi Logistik Ridge. *Jurnal Sains Dan Seni ITS*, *4*(2), 175–180.