MENGATASI MASALAH *MULTIKOLINEARITAS* DAN *OUTLIER* DENGAN PENDEKATAN ROBPCA

(STUDI KASUS ANALISIS REGRESI ANGKA KEMATIAN BAYI DI JAWA TIMUR)

Sony Sunaryo (sonny_s@statistika.its.ac.id)
Setiawan
Jurusan Statistika, Institut Teknologi Sepuluh Nopember

Tiodora Hadumaon Siagian (theo@bps.go.id)
Badan Pusat Statistik

ABSTRACT

Multicollinearity and outliers existence in data can be detected by various techniques. Principal Component Analysis (PCA) is one of the statistical techniques that can be used to handle data reduction and multicollinearity problem. However, PCA is very sensitive to outliers as it based on the mean and the covariance matrix. Hubert et al. (2005) developed ROBPCA, a robust PCA to the outliers existence. The ROBPCA combines PP technique and Minimum Covariant Determinant (MCD) method for solving outliers problem. In the present study, ROBPCA is applied to the study case of the regression analysis of infant mortality rate in East Java Province in 2009. The result shows that ROBPCA is more robust compare to PCA when data contains outlier. ROBPCA can explain 85.6 percent of variation by 2 principal components, whereas, PCA needs 3 principal components to explain 86.6 percent of variation. Moreover, ROBPCA produces higher coefficient determination which means the regression model using ROBPCA is better in explaining response variable. The study findings also revealed that the average of duration of exclusive breastfeeding has the largest contribution in lowering infant mortality rate followed by percentage of delivery assisted by medical provider and percentage of households that have access to safe drinking water.

Key words: infant mortality rate, multicollinearity, outlier, regression analysis, ROBPCA

Pada proses kalibrasi multivariat ditemui berbagai permasalahan statistik, antara lain masalah efek kesalahan acak, kolinearitas, penentuan statistik distribusi, pencocokan model pada data, pencilan (*outlier*), kekuatan (*robustness*) dan sebagainya. Dua permasalahan statistik yang didiskusikan dalam penelitian ini adalah masalah multikolinearitas dan *outlier*.

Multikolinearitas didefinisikan sebagai suatu kondisi dimana dua atau lebih variabel prediktor pada *Multiple Linier Regression* (MLR) berkorelasi tinggi. Masalah multikolinearitas seringkali membuat hasil analisis regresi menjadi tidak sesuai atau bertentangan dengan teori. Dalam MLR, multikolinearitas pada matriks $\mathbf{X}$ menghasilkan ($\mathbf{X}^T\mathbf{X}$) yang tidak berpangkat penuh (*singular*) sehingga dengan metode *Ordinary Least Square* (OLS) tidak dapat diperoleh koefisien regresi yang unik. Dampak lain dari adanya multikolinearitas pada analisis regresi antara lain; 1) satu atau lebih variabel prediktor adalah *redundant* artinya satu variabel prediktor menjelaskan tentang variabel respon persis sama dengan yang dijelaskan oleh variabel prediktor lainnya, 2) mempengaruhi kemampuan model untuk mengestimasi koefisien regresi, dan 3) varian dari estimasi parameter dengan OLS menjadi tinggi ketika satu atau lebih nilai *eigen* matriks **X** mendekati 0, yang berarti probabilitas rendah pada nilai vektor koefisien regresi β.

Outlier didefinisikan sebagai sebagian dari data pengamatan yang memiliki pola yang berbeda dari sebagian besar data pengamatan (Hadi, Imon, & Werner, 2009). *Outlie*r pada data dapat menyebabkan ketidakhomogenan matriks varian kovarian. Selain itu Hadi et al (2009) menyebutkan *outlier* memberi efek *masking* (mengaburkan data) dan *swamping* (kesalahan mengidentifikasi data non *outliers* sebagai *outliers*). Sebaiknya jumlah *outlier* dalam data tidak melebihi dari 50 persen. Pada analisis regresi, adanya *outlier* dapat menyebabkan berbagai penyimpangan antara lain; 1) residual yang besar dari model yang terbentuk, 2) varian data menjadi lebih besar, dan 3) rentang yang lebar pada *confidence region.*

Metode *Principal Component Analysis* (PCA) adalah salah satu teknik statistik yang dikenal bertujuan mereduksi dimensi dan sekaligus mengatasi masalah multikolinearitas pada data. Pada dasarnya PCA mentransformasi secara linier variabel asal yang umumnya saling berkorelasi menjadi sejumlah variabel yang lebih sedikit yang tidak berkorelasi dan disebut komponen utama (*Principal Component*). Metode PCA yang berdasarkan rata-rata dan matriks varian kovarian sangat sensitif terhadap *outlier* data pengamatan. Sehingga muncul kebutuhan PCA yang *robust* terhadap *outlier*. *Robust Principal Component Analysis* (ROBPCA) adalah suatu metode yang kuat (*robust*) untuk PCA terhadap keberadaan *outlier* pada data (Hubert, Rousseeuw, & Branden, 2005). Metode ROBPCA menggabungkan konsep *Projection Pursuit* (PP) dengan penduga *robust Minimum Covariance Determinant* (MCD). Identifikasi *outlier*-nya menggunakan kombinasi *orthogonal distance* dan *score distance* yang juga menggunakan penduga MCD.

Penelitian ini bertujuan mengatasi masalah multikolinearitas dan *outlier* dengan pendekatan ROBPCA pada studi kasus analisis regresi Angka Kematian Bayi (AKB) di Provinsi Jawa Timur tahun 2009. Diharapkan hasil penelitian ini dapat memberikan suatu gambaran bagaimana mengatasi masalah multikolinearitas dan *outlier* melalui pendekatan ROBPCA. Program ROBPCA dapat diunduh pada toolbox Matlab di *Robust Calibration* pada website http://www.wis.kuleuven.ac.be/stat/robust.html.

Model Regresi

Pada model regresi, variabel respon (y) dan p variabel prediktor $\left(x_1, x_2, \ldots, x_p\right)$ dinyatakan dalam persamaan matematis:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i$$

di mana:

$i = 1, 2, \ldots, n$

$\beta_0, \beta_1, \ldots, \beta_k$ adalah parameter model.

$\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ adalah kesalahan yang diasumsikan identik, independen dan berdistribusi normal dengan rata-rata nol dan varian konstan.

Penaksir parameter model dengan metode OLS dalam bentuk vektor dapat ditulis sebagai berikut:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

di mana:

$\hat{\beta} =$ Vektor parameter yang ditaksir yang berukuran $n \times (p+1)$

$X =$ Matriks data berukuran $n \times (p+1)$, variabel prediktor yang elemen pada kolom pertama bernilai 1

$Y =$ Vektor variabel respon yang berukuran $n \times 1$

$k =$ Banyaknya variabel prediktor ($k = 1, 2, \ldots, p$)

Deteksi Multikolinearitas dan *Outlier*

Ada beberapa cara mendeteksi multikolinearitas pada data pengamatan antara lain (dalam Naes, Isakson, Fearn, & Davies, 2002):

1. Dengan menghitung koefisien korelasi antara sesama variabel prediktor. Jika nilai koefisien korelasi melebihi 0,8 maka ini mengindikasikan adanya masalah kolinearitas di dalam regresi.
2. Dengan menghitung *Variance Inflation Factor* (VIF). Jika nilai VIF melebihi 10, maka hal ini menunjukkan adanya masalah multikolinearitas antar variabel prediktor. Semakin tinggi nilai VIF-nya maka semakin serius permasalahan multikolinearitasnya.

   $VIF_k = \dfrac{1}{1 - R_k^2}$ dimana R_k^2 adalah kuadrat dari koefisien korelasi.
3. Dengan menghitung nilai TOL yaitu suatu ukuran *tolerance* untuk deteksi multikolinearitas, di mana:

   $$TOL_k = \frac{1}{VIF_k} = 1 - R_k^2 = \begin{cases} 1 & \text{Akan bernilai 1 jika } X_k \text{ tidak berkorelasi dengan variabel prediktor lainnya.} \\ 0 & \text{Akan bernilai 0 jika } X_k \text{ berkorelasi dengan variabel prediktor lainnya.} \end{cases}$$
4. Dengan menghitung *Condition number* (CN).

   $CN = \mathrm{K} = (\hat{\lambda}_1 / \hat{\lambda}_{\mathrm{K}})^{1/2}$

   dimana $\hat{\lambda}_1$ adalah nilai *eigen* terbesar dan $\hat{\lambda}_{\mathrm{K}}$ adalah nilai *egien* terkecil dari matriks kovarian. Jika $K \geq 30$, ini mengindikasikan terjadi masalah multikolinearitas.

Di dalam menghadapi masalah identifikasi *outlie*r terdapat dua pendekatan (Hadi et al, 2009) yaitu: estimasi yang *robust* dan metode yang khusus mengidentifikasi *outlier*. Metode yang *robust* dirancang secara khusus untuk mengatasi *outlier,* di mana kemudian saat penghitungan, estimasi parameter yang *robust* dapat digunakan untuk mengidentifikasi *outliers*. Di sisi lain, prosedur identifikasi *outliers* dapat pula digunakan untuk mendapatkan *estimator* yang *robust*. Berbagai cara mendeteksi keberadaan *outlier* pada data pengamatan antara lain:

1. Membandingkan nilai F_i dengan $F_{tabel.}$

   Suatu data pengamatan dikatakan *outlier* jika nilai $F_i > F_{\alpha;p,n-p-1}$

   atau $\dfrac{(n-p-1)nd_i^2}{p(n-1)^2 - npd_i^2} > F_{\alpha;p,n-p-1}$
2. Menggunakan Leverage yang berkaitan dengan jarak Mahalanobis. Nilai Leverage untuk sampel ke-i didefinisikan sebagai:

$$h_i = \frac{1}{n} + \mathbf{X}_i^T (\mathbf{X}^T \mathbf{X})^{-1} x_i .$$

Di mana X adalah matriks data. Jika nilai Leverage > $(2p-1)/n$ maka dianggap sebagai *outlier*.

Metode ROBPCA

Metode ROBPCA adalah suatu metode PCA yang *robust* terhadap keberadaan *outlier* pada data. Komponen *Loading* dihitung dengan menggunakan teknik *Projection-Pursuit* (PP) dan metode *Minimum Covariance Determinant* (MCD). Teknik PP digunakan untuk reduksi dimensi awal kemudian estimator MCD diaplikasikan menghasilkan estimasi yang lebih akurat. Deskripsi komplit dari metode ROBPCA dapat dilihat pada lampiran tulisan Hubert et al (2005) namun secara sederhana metode ROBPCA dapat dideskripsikan sebagai berikut; jika diasumsikan data asal berupa suatu matriks $\mathbf{X}$ berukuran $n \times p$ dimana n jumlah pengamatan dan p jumlah variabel asal maka metode ROBPCA dilakukan dalam 3 langkah utama. Pertama, dilakukan pre-proses data sehingga transformasi data berada pada sebuah subruang dengan dimensi paling tinggi $n-1$. Kemudian dibentuk matriks kovarian S_0 yang digunakan untuk memilih jumlah komponen k menghasilkan subruang berdimensi k yang cocok dengan data. Titik-titik data kemudian diproyeksikan ke subruang ini kemudian lokasi dan matriks sebarannya diestimasi secara *robust* dan dihitung nilai *eigen* $l_1, l_2, \ldots, l_k$. Maka didapat vektor e*igen* yang bersesuaian adalah sejumlah k komponen utama yang *robust.*

Metode *Projection Pursuit*

Biasanya transformasi reduksi dimensi dilakukan dengan proyeksi linier atau kombinasi linier dari variabel-variabel asal. Namun reduksi dimensi berdasar metode *Projection Pursuit* (PP) dilakukan dengan cara memaksimumkan fungsi obyektif yang dikenal dengan *projection index*. Secara singkat metode PP didefinisikan oleh Jimenez dan Landgrebe (1995). Jika diasumsikan $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ adalah matriks data berukuran $d \times \mathrm{N}$ dan $\mathbf{Y}$ hasil proyeksi data dengan dimensi yang sudah tereduksi berukuran $m \times \mathrm{N}$ dan $\mathbf{A}$ adalah matriks orthonormal parametrik berukuran $d \times m$ dengan $\mathbf{Y} = \mathbf{A}^T \mathbf{X}$. Metode PP adalah metode yang menghitung $\mathbf{A}$ yang mengoptimalkan *projection index* $I(\mathbf{A}^T \mathbf{X})$.

Penduga MCD

Penduga MCD adalah penduga yang sangat *robust* untuk lokasi dan sebaran multivariat dan *resistant* terhadap keberadaan *outlier*. Meski sebenarnya sudah diperkenalkan sejak tahun 1984 namun penggunaannya baru dimulai sejak ada algoritma FAST-MCD yang dibangun oleh Rousseeuw dan Van Driessen (1999). Penduga MCD adalah berdasarkan determinan matriks varian kovarian minimum. Penduga ini didapat dengan cara mencari h pengamatan (*halfset*) yang memberikan nilai minimum dari matriks varian kovarian (Hubert, Rousseeuw, & Van Aelts, 2008). Jika diasumsikan $X = \{x_1, \ldots, x_n\}$ adalah sampel dari sejumlah n pengamatan dalam suatu ruang berdimensi R^k maka metode MCD berupaya mendapatkan h pengamatan $(h \le n)$ yang memiliki

determinan matriks varian kovarian terkecil dengan $h_0 \le h \le n$ dimana h_0 adalah bilangan bulat terkecil dari $\left((n+p+1)/2\right)$

$$MCD \approx \min\left\{\det\left(C(X)\right)_j\right\},\ j=1,2,\ldots,\binom{n}{h}$$

dengan:

$\mathbf{T(X)} = \frac{1}{h}\sum_{i=1}^{h}\mathbf{x}_i$ adalah penduga parameter lokasi berdasarkan MCD yaitu rata-rata dari subsampel h. Sedangkan $\mathbf{C(X)} = \frac{1}{h-1}\sum_{i=1}^{h}\left(\mathbf{x}_i - \mathbf{T(X)}\right)\left(\mathbf{x}_i - \mathbf{T(X)}\right)^T$ adalah penduga sebaran atau matriks varian kovarian, matriks $p \times p$ simetris definit positif yang berasal dari subsampel h dan $\mathbf{x}_i$ adalah vektor pengamatan ke-i serta n adalah jumlah seluruh pengamatan.

Diagnostic Plot

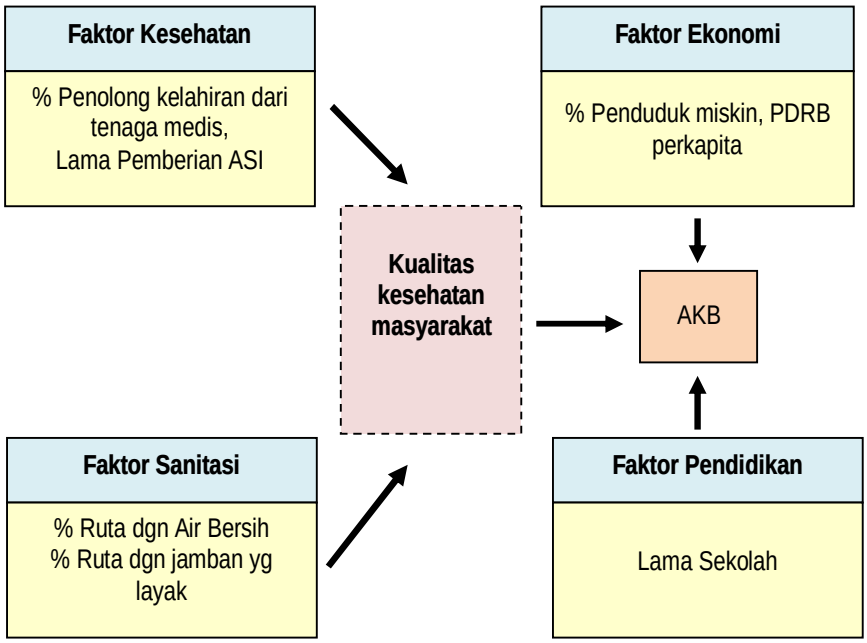
Metode ROBPCA bertujuan: 1) mencari kombinasi linier dari variabel asal yang mengandung paling banyak informasi walaupun data mengandung *outlier*, dan 2) mendeteksi dan mengklasifikasi *outlier* (Hubert et al, 2005). Berdasar tujuan ini maka dibentuk *Diagnostic Plot* atau peta *outlier* yang berguna untuk membedakan data pengamatan. Di dalam *Diagnostic Plot*, data pengamatan dibedakan menjadi 4 tipe yaitu:

1. Pengamatan regular (*regular observations*) yaitu kelompok data yang homogen yang dekat dengan subruang PCA.
2. Titik Leverage baik (*good leverage points*) yaitu data yang dekat dengan ruang PCA namun jauh dari *regular observations.*
3. *Orthogonal outliers* yaitu data yang memiliki *orthogonal distance* yang besar terhadap ruang PCA.
4. Titik Leverage buruk (*bad leverage points*) yaitu data yang memiliki *orthogonal distance* yang besar dan proyeksinya pada subruang PCA terpencil dari *typical projections*.

Studi Kasus: Analisis Regresi Angka Kematian Bayi (AKB) di Jawa Timur Tahun 2009

Menurunkan AKB merupakan salah satu target yang ingin dicapai dalam *Millenium Development Goals* (MDGs), suatu komitmen bersama masyarakat internasional untuk mempercepat pembangunan manusia dan mengentaskan kemiskinan. Estimasi AKB menjadi penting mengingat AKB merupakan salah satu indikator pembangunan bidang kesehatan di suatu wilayah. Mengingat pentingnya mengestimasi AKB guna menunjang pembangunan, maka studi kasus yang dipilih dalam penelitian ini adalah analisis regresi AKB per kabupaten/kota di Provinsi Jawa Timur pada tahun 2009.

Berdasar kerangka teori (Gambar 1) dan berbagai literatur tentang AKB, maka diambil 7 variabel prediktor yang diperkirakan mempengaruhi AKB. Data bersumber dari Badan Pusat Statistik (BPS) Provinsi Jawa Timur dengan unit analisis penelitian kabupaten/kota. Sehingga variabel respon dan variabel prediktor dalam penelitian ini tampak pada Tabel 1 berikut.



Sumber: Winarno (2009)

Gambar 1. Kerangka Teori Penelitian

Tabel 1. 7 variabel prediktor yang diperkirakan mempengaruhi AKB

Variabel	Deskripsi Variabel
Y	Angka Kematian Bayi (AKB) yaitu jumlah bayi meninggal per 1000 kelahiran hidup
X1	Persentase jumlah penolong kelahiran dari tenaga medis (%)
X2	Rata-rata lama pemberian Air Susu Ibu (ASI) eksklusif (bulan)
X3	Rata-rata lama sekolah (tahun)
X4	Persentase rumah tangga yang memiliki sumber air minum bersih (%)
X5	Persentase rumah tangga yang memiliki fasilitas jamban yang layak (%)
X6	Persentase penduduk miskin (%)
X7	Produk Domestik Regional Bruto (PDRB) per kapita (Rp)

Mengingat bahwa variabel-variabel sosial seringkali saling berkorelasi maka peneliti melakukan pra-pemrosesan data dengan metode PCA untuk mengatasi masalah multikolinearitas. Namun metode PCA memiliki keterbatasan pada data yang mengandung *outlier*. Untuk itu digunakan pendekatan ROBPCA pada analisis regresi AKB di Jawa Timur. Secara umum langkah-langkah analisis data dalam penelitian adalah sebagai berikut:

1. Standarisasi variabel prediktor.
2. Uji multikolinearitas data pengamatan.
3. Uji keberadaan *outlier* pada data pengamatan.
4. Uji kenormalan data pengamatan.
5. Cari komponen utama dengan metode ROBPCA.
6. Lakukan regresi dengan komponen utama terpilih.

HASIL DAN PEMBAHASAN

Terlebih dahulu variabel prediktor distandarisasi karena ada perbedaan satuan pengukuran, kemudian dilakukan uji multikolinearitas dengan melihat nilai *tolerance* dan VIF nya dengan bantuan software SPSS 16. Hasil analisis ditunjukkan pada Tabel 2. Dari tabel ini diketahui bahwa ada multikolinear pada data pengamatan karena terdapat nilai VIF yang lebih dari 10 dan *tolerance* yang kurang dari 0,1.

Tabel 2. Nilai *Tolerance* dan VIF

Model	Collinearity Statistics	
	Tolerance	VIF
X1	0,189	5,305
X2	0,366	2,736
X3	0,084	11,847
X4	0,498	2,007
X5	0,111	9,010
X6	0,315	3,171
X7	0,735	1,361

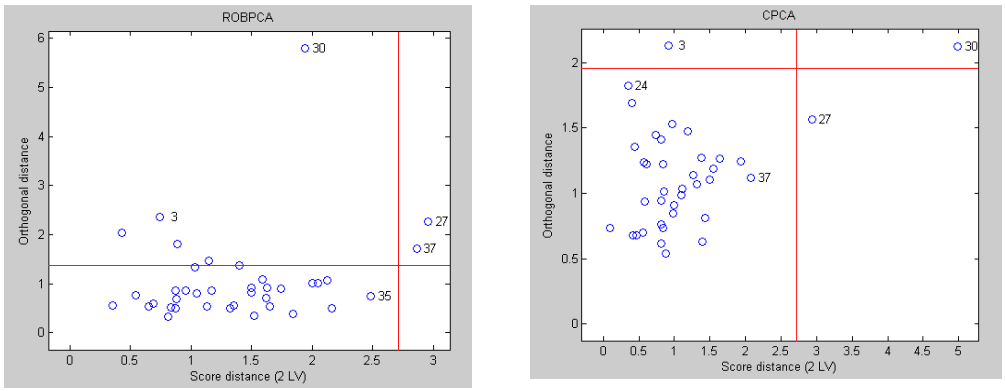
Penghitungan koefisien korelasi antar variabel prediktor juga menemukan adanya korelasi yang tinggi antara variabel X_1-X_3, X_1-X_5, X_3-X_5, dan X_3-X_6 dimana koefisien korelasinya bernilai lebih dari 0,8 dan *p-value* kurang dari α=5%. Sehingga semakin kuat dugaan adanya multikolinearitas pada data. Sementara itu, hasil pengujian keberadaan *outlier* dengan membandingkan antara nilai F_i dengan F_{tabel} pada data pengamatan menunjukkan adanya *outlier* di dua wilayah yaitu Kabupaten Sampang dan Kabupaten Kediri. Setelah mengetahui ada multikolinearitas dan *outlier* pada data pengamatan, aplikasi metode ROBPCA dianggap tepat dilakukan pada analisis regresi.

Pendekatan metode ROBPCA mensyaratkan data berdistribusi simetris. Dilakukan uji multinormal pada data pengamatan. Hasil *scatter plot* menunjukkan kecenderungan membentuk garis lurus dan sekitar 55,26 persen nilai $d_i^2 \le X^2_{7;0,50}$. Dengan demikian dapat disimpulkan data berdistribusi Multinormal dan metode ROBPCA dapat diaplikasikan pada data. Pendekatan dengan metode ROBPCA dilakukan dengan bantuan program ROBPCA yang tersedia di toolbox Matlab dengan mengetikkan perintah:

```
result=robpca(x,'k',7,'kmax',10,'alpha',0.75,'mcd',1,'plots',1,'labsd',3,'labod',3,'classic',0).
```

Gambar 2 menunjukkan perbandingan hasil *Diagnostic Plot* dengan 2 komponen utama dan Tabel 2 menyajikan perbandingan hasil reduksi dimensi antara metode ROBPCA dan PCA. Pada Gambar 2a dapat dibedakan kelompok data *orthogonal outliers* (berlabel 3 dan 30) dan satu kelompok *bad leverage points* (27 dan 37), namun tidak ada yang masuk dalam kelompok *good leverage points*. Hasil ini kemudian dibandingkan dengan analisis PCA yang klasik (CPCA) yang disajikan pada Gambar 2b. Meski *outlier* yang sama tetap terdeteksi namun tampak bahwa Gambar 2b sangat berbeda dibanding plot yang *robust* pada Gambar 2a. Perbedaannya terletak pada kasus *bad leverage points* dari ROBPCA malah dikonversi menjadi *good leverage points* oleh klasikal PCA.

(a) *Robust* PCA (b) Klasikal PCA



Gambar 2. Perbandingan *diagnostic plot* dari data pengamatan

Dari Tabel 3 dapat diketahui bahwa hanya dengan menggunakan 2 komponen utama, keragaman yang dapat dijelaskan dari metode ROBPCA sudah mencapai 85,6 persen dan bahkan mencapai 93,4 persen dengan 3 komponen utama. Sedangkan dengan 3 komponen utama, keragaman yang dijelaskan metode PCA baru mencapai 86,6 persen.

Tabel 3. Perbandingan Nilai *Eigen* dan Keragaman Kumulatif

Komponen Utama	**Nilai *Eigen***		**Keragaman Kumulatif**	
ke	**PCA**	**ROBPCA**	**PCA**	**ROBPCA**
1	4,5740	3.5923	0,653	0,737
2	0,9444	0.57813	0,788	0,856
3	0,5423	0.38489	0,866	0,934
4	0,4635	0.17752	0,932	0,971
5	0,2996	0.10273	0,975	0,992
6	0,1227	0.031709	0,992	0,998
7	0,0536	0.0075369	1,000	1,000

Ada berbagai cara untuk menentukan jumlah komponen utama yang akan dipakai. Misalnya dengan melihat keragaman kumulatifnya jika sudah lebih dari 80 persen (Johnson & Wichern, 2002), atau dengan melihat nilai *eigen*-nya yang lebih besar dari 1 atau bisa juga melalui pengamatan patahan siku pada *screeplot* dari nilai *eigen* dan jumlah komponennya. Berdasar nilai keragaman kumulatif, diputuskan menggunakan 3 komponen utama pada metode PCA dan 2 komponen utama pada metode ROBPCA.

Analisis regresi dengan variabel baru yang bebas dari multikolinearitas menunjukkan bahwa model regresi dengan 3 komponen utama berdasar metode PCA menghasilkan koefisien determinasi (R^2) sebesar 76,1 persen. Artinya variabel penelitian dalam model regresi yang dibentuk berdasar metode reduksi PCA mempengaruhi AKB sebesar 76,1 persen sedangkan sisanya (23,9 persen) dipengaruhi oleh variabel prediktor lainnya yang tidak diteliti. Jika dibandingkan model regresi dengan hanya 2 komponen utama berdasar metode ROBPCA menghasilkan R^2 sebesar 77.4 persen. Hal ini menunjukan variabel penelitian dalam model regresi yang dibentuk berdasar metode reduksi ROBPCA mempengaruhi AKB sebesar 77,4 persen sedangkan sisanya yaitu 22,6 persen AKB dipengaruhi oleh variabel prediktor lainnya yang tidak masuk di dalam model.

Setelah dikembalikan ke variabel asal maka persamaan regresi yang didapat adalah sebagai berikut:

Dengan 3 komponen utama pada metode PCA:

$$\hat{Y} = 106 - 3,039X_1 - 4,443X_2 - 1,318X_3 - 2,338X_4 - 1,345X_5 + 1,849X_6 + 1,664X_7$$

Dengan 2 komponen utama pada metode ROBPCA:

$$\hat{Y} = 38,7 - 3,339X_1 - 5,247X_2 - 1,156X_3 - 2,585X_4 - 1,266X_5 + 0,168X_6 + 0,059X_7$$

Dari kedua persamaan regresi tersebut diketahui variabel yang paling mempengaruhi AKB adalah X_2 (rata-rata lama pemberian ASI eksklusif), X_1 (persentase jumlah penolong kelahiran dari tenaga medis) dan X_4 (persentase rumah tangga yang memiliki sumber air minum bersih).

KESIMPULAN

Kesimpulan yang dapat diambil dari hasil analisis adalah metode PCA dapat menjelaskan keragaman 86,6 persen dengan 3 komponen utama, sedangkan metode ROBPCA sudah dapat menjelaskan keragaman 85,6 persen dengan hanya 2 komponen utama saja dan bahkan keragaman mencapai 93,4 persen jika menggunakan 3 komponen utama. Dapat disimpulkan metode ROBPCA memang lebih *robust* untuk PCA terhadap keberadaan *outlier*. Meski perbedaannya tidak terlalu signifikan, namun metode ROBPCA menghasilkan model regresi dengan R^2 yang lebih tinggi dibanding dengan metode PCA, artinya model regresi yang dihasilkan lebih baik dalam menjelaskan variabel respon yaitu AKB. Tidak ditemukannya perbedaan yang cukup signifikan pada koefisien determinasi antara metode PCA dan metode ROBPCA mungkin disebabkan karena sedikitnya *outlier* pada data yang digunakan dalam contoh kasus. Untuk itu perlu dilakukan penelitian lebih lanjut mengenai aplikasi metode ROBPCA pada studi kasus dengan data yang mengandung lebih banyak *outlier* sehingga dapat diketahui seberapa jauh perbedaannya pada kebaikan model regresi. Selain itu pendekatan metode ROBPCA hanya mampu mengatasi masalah multikolinearitas dan *outlier* pada data prediktor, sehingga disarankan penelitian lebih lanjut dengan pendekatan yang berbeda yang sekaligus mampu mengatasi *outlier* pada variabel respon.

REFERENSI

Hadi, A.S. Imon, A.H.M.R, & Werner, M. (2009). Detection of Outliers. *WIREs Computational Statistics*, 1, 57-70.

Hubert, M. Rousseeuw, P.J. & Branden, K.V. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47, 64-79.

Hubert, M. Rousseeuw, P.J. & Van Aelts, S. (2008). High-Breakdown robust multivariate methods. *Statistical Science*, 23 (1), 92-119.

Jimenez, L.O. & Landgrebe, D, (1995), *High dimensional feature reduction via projection pursuit*. Purdue University.

Johnson, R.A & Wichern, D.W, (2002), *Applied multivariate statistical analysis* (5th ed). New Jersey: Prentice Hall.

Naes, T., Isaksson, T., Fearn, T., & Davies, T., (2002), *Multivariate calibration and classification*. West Sussex: NIR Publication.

Rousseeuw, P.J. & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212-223.

Winarno, D. (2009). *Analisis angka kematian bayi di Jawa Timur dengan pendekatan model regresi spasial*. Thesis master yang tidak dipublikasikan, Institut Teknologi Surabaya.