PENAKSIRAN PARAMETER PLS DENGAN METODE ALGORITMA NIPALS MENGGUNAKAN BAHASA PEMROGRAMAN STATISTICA

ESTIMATING THE PARAMETERS OF PLS WITH NIPALS ALGORITHM METHOD USING STATISTICA PROGRAMMING LANGUAGES

Hedi
UP MKU Politeknik Negeri Bandung
hedi@polban.ac.id

ABSTRAK

Prediksi nilai variabel respon didasarkan pada beberapa variabel prediktor dinamakan analisis regresi. Dalam praktiknya, permasalahan yang sering muncul adalah variabel-variabel prediktor yang saling berkorelasi (multikolinieritas). Untuk mengatasi masalah tersebut diterapkan metode *Partial Least Squares Regression* (PLS). Penaksiran parameter pada PLS diperoleh dengan menggunakan algoritma NIPALS (*Non Linier Iterative Least Square*) yang langkah-lanngkah perhitungannya ditentukan dengan bahasa pemrograman STATISTICA.

Kata Kunci: Regresi, PLS, NIPALS, STATISTICA

ABSTRACT

Regression analysis is a technique of statistical analysis that describes the relationship between the response variable and predictor variables. In practice, problems frequently arise are predictor variables were correlated (multicollinearity). To overcome these problems applied method of Partial Least Squares Regression (PLS). Parameter estimation on the PLS obtained by using algorithms NIPALS (Non Linear Iterative Least Square) is determined by through STATISTICA programming language.

***Keywords**: Regression, PLS, NIPALS, STATISTICA*

PENDAHULUAN

Analisis regresi adalah teknik analisis secara statistik yang menjelaskan hubungan di antara variabel respon dan variabel-variabel prediktor. Dengan menerapkan data sampel yang merupakan pasangan prediktor dan respon, taksiran parameter model ditentukan dengan metode *ordinary least squares* (OLS**)** (Ronald E. 2012). Permasalahan yang sering ditemukan dalam analisis regresi adalah multikolenieritas yaitu, adanya korelasi antara variabel prediktor yang berimplikasi pada solusi *least square* tidak unik sehingga persamaan regresi tidak dapat menjelaskan dengan baik hubungan antara variabel respon dan variabel-variabel prediktor (Gujarati, D. N, 2009)**.**

Metode untuk mengatasi masalah multikolenieritas, salah satunya adalah *Principal Coponent Regression (*PCR). Satu metode baru diperkanalkan sebagai perbaikan dari metode PCR yaitu metode *Partial least square Regression* (PLS). Metode ini dalam proses reduksi variabel independen telah mengakomodasi korelasi antara variabel respon dengan variabel bebasnya (Nyoman, 2009).

PLS adalah mendekomposisikan matriks variabel prediktor menjadi matriks komponen dan matriks loading komponen-komponen tersebut akan berperan sama seperti variabel prediktor. Selanjutnya, metode pembentukan dan seleksi komponen menggunakan algoritma NIPALS (Wold 1966)

Penaksiran parameter PLS dengan algoritma NIPALS dapat ditentukan dengan menggunakan beberapa software komputer seperti STATISTICA, EVIEW, dan SAS. Akan tetapi, *output* yang dijalankan oleh *software* tersebut tikak selalu memenuhi berbagai bentuk teori statistika. Oleh karena itu, perlu bahasa pemrograman yang dapat menyelesaikan algoritma NIPALS.

Prediksi nilai variabel respon didasarkan pada beberapa variabel prediktor dinamakan analisis regresi. Model regresi dengan k variabel prediktor $X_1, X_2, X_3 ... X_k$ dan variabel respon Y

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + .. \beta_k X_k + \varepsilon$$

(1) dengan ε : Eror

Untuk menentukan taksiran parameter model regresi berdasarkan data sampel prediktor $X_1, X_2, ..., X_K$ dan respon Y yang disajikan dalam tabel 1.

Tabel 1. Pasangan Data Pengamatan

Y	X_1	X_2	…	X_K
y_1	x_{11}	x_{12}	…	x_{1k}
y_2	x_{21}	x_{22}	…	x_{2k}
	.	.	.	.
	.	.	.	.
	.	.	.	.
y_n	x_{n1}	x_{n2}	…	x_{nk}

selanjutnya bangun matriks

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \\ . \\ . \\ . \\ \mathbf{y_n} \end{pmatrix}, \mathbf{X}$$

$$= \begin{pmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2k} \\ . & . & . & & & & . \\ . & . & . & & & & . \\ . & . & . & & & & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{nk} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{pmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_k \end{pmatrix}$$

sehingga model persamaan 1 dalam bentuk matriks adalah

$$Y = X\beta + \varepsilon \qquad (2)$$

Jika diasumsikan X^tX adalah invertible artinya mempunyai rank penuh (kondisi ini dipenuhi apabila *k* cukup besar), dengan menerapkan *ordinary least square* (OLS), estimasi B (Ronald E. 2012) adalah

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad (3)$$

Apabila pada model persamaan 3, X^tX singular yang disebabkan terjadinya korelasi antara variabel prediktor sehingga estimasi parameter regresi tidak ada atau dengan standar error sangat besar. Selanjutnya diterapkan model PLS, untuk itu terlebih dahulu model persamaan 1 diubah kedalam bentuk model mean centered sebagai berikut :

$$Y - \overline{Y} = \beta_1(X_1 - \overline{X_1}) + \beta_2(X_2 - \overline{X_2}) +$$
$$\beta_3(X_3 - \overline{X_3}) + .. \beta_1(X_k - \overline{X_k}) + \varepsilon$$

dengan $\overline{X_i}$: rata-rata data prediktor ke i = 1, 2, 3, ... , k, dan $\overline{Y}$: rata-rata Respon

$$\tilde{Y} = \tilde{X}\beta + \varepsilon \qquad (4)$$

PLS adalah mendekomposisikan matriks variabel prediktor menjadi matriks komponen orthogonal T dan matriks *loading* P. Komponen- komponen tersebut akan berperan sama seperti variabel prediktor. Selanjutnya, metode pembentukan dan seleksi komponen menggunakan algoritma NIPALS.

Langkah pertama adalah memberi bobot variabel prediktor pada pembentukan komponen PLS . Pilihlah bobot yang menjelaskan variasi prediktor sekaligus variasi respon melalui fungsi kovariansi antara setiap variabel prediktor dengan variabel respon sehingga komponen yang terbentuk tidak hanya baik dalam menerangkan variasi prediktor juga relevan dan baik dalam memprediksi respon. Algoritma NIPALS bersifat iteratif, jadi vektor komponen pertama dapat menunjukan seberapa baik menerangkan variasi dari prediktor dan memprediksi respon. Kemudian algoritma ini akan mengiterasi pembentukan vektor komponen berikutnya.

Langkah-langkah algoritma NIPALS (korkmazoglu Ozlem berak, 2012) adalah

1. Inisialisasi j = 1 (dengan j = 1, 2, 3 ... m)
2. Tentukan Bobot w_j , $w_j = \dfrac{B_j}{\|B_j\|}$ dengan $B_j = X_j^t Y$
3. Tentukan *X-scores* t_j, $t_j = X_j w_j$, dan *X-loadings* , $p_j = \dfrac{X_j^t t_j}{t_j^t t_j}$
4. Tentukan *Y-loading* , $\hat{c}_j = \dfrac{t_j^t Y_j}{t_j^t t_j}$
5. Tentukan residual, $X_{j+1} = X_j - t_j p_j^t$ dan $Y_{j+1} = Y_j - t_j \hat{c}_j$
6. Ulangi langkah 1 sampai dengan 5 untuk j = 1, 2, 4, ... m
7. Bentuk Matriks $W = [w_1\ w_2\ w_3\ ...\ w_m]$, $T = [t_1\ t_2\ t_3\ ...\ t_m]$, $P = [p_1\ p_2\ p_3\ ...\ p_m]$, dan $C = [c_1\ c_2\ c_3\ ...\ c_m]$

Taksiran parameter PLS adalah

$$\hat{\beta}_{pls} = W(P^t W)^{-1}\,\hat{c} \qquad (5)$$

METODE

Metodologi penelitian ini dilakukan empat tahap yaitu Pengumpulan Data, Pengujian data, Nipals, dan Penaksiran. Perangkat komputer yang digunakan dalam pengolahan data adalah STATISTICA versi 8.

Pengumpulan Data

Penelitian ini menggunakan data sekunder, dengan variabel respon data diasumsikan bergantung pada 13 variabel *predictor*. Sebagai *respon* adalah kandungan lemak tubuh dan sebagai *predictor* adalah umur dalam tahun, berat badan dalam pon, tinggi badan dalam inci, kemudian dalam satuan cm, lingkar leher, lingkar dada, lingkar perut, lingkar pinggul, lingkar paha , lingkar lutut, lingkar pergelangan kaki , lingkar bisep, lingkar lengan, dan lingkar pergelangan tangan.

Pengujian Data

Berdasarkan data, dilihat pola hubungan variabel prediktor dengan menggunakan software STATISTICA, Langkah ini, menguji keberadaan multikolinieritas antara variabel predictor.

Nipals

Selanjutnya untuk menentukan berapa komponen yang masuk kedalam model diterapkan algoritma NIPALS menggunakan sofware STATISTICA.

Penaksiran

Berdasarkan banyaknya komponen yang masuk kedalam model, akan dilakukan penaksiran parameter regresi, dengan software STATISTICA
dan bahasa pemrograman STATISTICA.

HASIL DAN PEMBAHASAN

Penerapan *Partial least square Regression* (PLS), menggunakan data dari StatLib-Datasets Archive, *Lib.stat.cmu.edu/datasets/Under filename bodyfat*. Sebagai respon Y adalah kandungan lemak tubuh yang bergantung pada 13 variabel predictor yaitu X_1: umur dalam tahun, X_2 : berat badan dalam pon, X_3 : tinggi badan dalam inci, kemudian dalam satuan cm, X_4 : lingkar leher, X_5 : lingkar dada, X_6 : lingkar perut, X_7 : lingkar pinggul, X_8 : lingkar paha , X_9 : lingkar lutut, X_{10} : lingkar pergelangan kaki , X_{11} : lingkar bisep, X_{12} : lingkar lengan, dan X_{13} : lingkar pergelangan tangan. Langkah pertama analisis data dilakukan dengan model regresi *linier multiple*, estimasi parameter menggunakan metode Ordinary Least Square. Hasilnya adalah sebagai berikut :

Tabel 2. Estimasi Parameter Regresi

Regression Summary for Dependent Variable: Y (LemakTubuh) R= .86547673 R²= .74904997 Adjusted R²= .73534261 F(13,238)=54.646 p

	Beta	Std.Err. - of Beta	B	Std.Err. - of B	t(238)	p-level
Intercept			-18.1885	17.34857	-1.04841	0.295511
X1	0.093481	0.048712	0.0621	0.03235	1.91904	0.056176
X2	-0.310598	0.187970	-0.0884	0.05353	-1.65238	0.099775
X3	-0.030459	0.042020	-0.0696	0.09601	-0.72485	0.469254
X4	-0.136698	0.067526	-0.4706	0.23247	-2.02437	0.044049
X5	-0.024040	0.099878	-0.0239	0.09915	-0.24069	0.809999
X6	1.230220	0.111388	0.9548	0.08645	11.04444	0.000000
X7	-0.177666	0.124906	-0.2075	0.14591	-1.42239	0.156223
X8	0.148112	0.090560	0.2361	0.14436	1.63552	0.103262
X9	0.004404	0.069736	0.0153	0.24198	0.06315	0.949699
X10	0.035239	0.044853	0.1740	0.22147	0.78565	0.432853
X11	0.065562	0.061779	0.1816	0.17113	1.06123	0.289663
X12	0.109145	0.048081	0.4520	0.19913	2.27001	0.024102
X13	-0.180792	0.059677	-1.6206	0.53495	-3.02954	0.002720

Tabel 3. Matriks Korelasi Antara Prediktor

Correlations (LemakTubuh) Marked correlations are significant at p < .05000 N=252 (Casewise deletion of missing data)

	Means	Std.Dev.	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13
X1	44.9	12.6	1.00	-0.01	-0.17	0.11	0.18	0.23	-0.05	-0.20	0.02	-0.11	-0.04	-0.09	0.21
X2	178.9	29.4	-0.01	1.00	0.31	0.83	0.89	0.89	0.94	0.87	0.85	0.61	0.80	0.63	0.73
X3	70.1	3.7	-0.17	0.31	1.00	0.25	0.13	0.09	0.17	0.15	0.29	0.26	0.21	0.23	0.32
X4	38.0	2.4	0.11	0.83	0.25	1.00	0.78	0.75	0.73	0.70	0.67	0.48	0.73	0.62	0.74
X5	100.8	8.4	0.18	0.89	0.13	0.78	1.00	0.92	0.83	0.73	0.72	0.48	0.73	0.58	0.66
X6	92.6	10.8	0.23	0.89	0.09	0.75	0.92	1.00	0.87	0.77	0.74	0.45	0.68	0.50	0.62

X7	99.9	7.2	-0.05	0.94	0.17	0.73	0.83	0.87	1.00	0.90	0.82	0.56	0.74	0.55	0.63
X8	59.4	5.2	-0.20	0.87	0.15	0.70	0.73	0.77	0.90	1.00	0.80	0.54	0.76	0.57	0.56
X9	38.6	2.4	0.02	0.85	0.29	0.67	0.72	0.74	0.82	0.80	1.00	0.61	0.68	0.56	0.66
X10	23.1	1.7	-0.11	0.61	0.26	0.48	0.48	0.45	0.56	0.54	0.61	1.00	0.48	0.42	0.57
X11	32.3	3.0	-0.04	0.80	0.21	0.73	0.73	0.68	0.74	0.76	0.68	0.48	1.00	0.68	0.63
X12	28.7	2.0	-0.09	0.63	0.23	0.62	0.58	0.50	0.55	0.57	0.56	0.42	0.68	1.00	0.59
X13	18.2	0.9	0.21	0.73	0.32	0.74	0.66	0.62	0.63	0.56	0.66	0.57	0.63	0.59	1.00

Berdasarkan hasil pada tabel 2, respon Y hanya dipengaruhi X_4 , X_6 , X_{12} , dan X_{13} dengan p-level dibawah 5%, ini berarti model tidak cocok. Selajutnya korelasi antar predictor diperlihatkan pada tabel berikut:

Berdasarkan matriks korelasi pada tabel 3, nilai korelasi antara predictor hampir setengahnya diatas 50% , ini berarti mengindikasikan keberadaan multikolinieritas. Berikutnya diterapkan Metode PLS (*Partial least square Regression).* Langkah pertama adalah menentukan banyaknya komponen PLS dengan cara menghitung koefisien determinasi. Berdasarkan hasil perhitungan dengan software STATISTICA diperoleh hasil sebagai berikut.

Tabel 4. Koefisien Determinasi

Partial Least Squares Analysis Summary (LemakTubuh) Number of components is 3 69.2858% of sum of squares of the dependent variables has been explained by all the extracted components.							
	R²X	**R²X(Cumul.)**	**R²Y**	**R²Y(Cumul.)**	**Q²**	**Q²(Cumul.)**	**Significance**
1	0.61	0.61	0.46	0.46	0.37	0.37	S
2	0.10	0.72	0.18	0.64	0.25	0.53	S
3	0.05	0.76	0.05	0.69	-0.16	0.45	NS

Hasil perhitungan R^2 pada tabel 4, kolom 2 menyatakan variansi Prediktor X yang dijelaskan oleh komponen ke j, dan kolom 3 adalah bentuk kumulatif, banyaknya komponen utama yang dipilih adalah 3 buah.

Bahasa Pemorograman STATISTICA

Berdasarkan data ukuran 252, banyaknya prediktor 13, banyaknya komponen PLS 3 buah , dan Algoritma NIPALS. Estimasi parameter PLS dengan bahasa pemrograman STATISTICA adalah sebagai berikut.

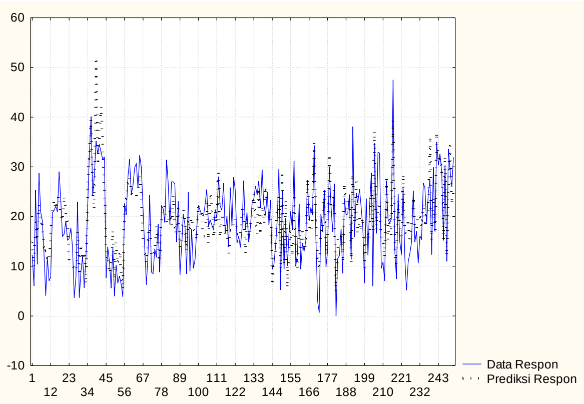
```
Option Base 1
Option Explicit
'#Uses "*STB.SVX"
'#Uses "*GRAPHICS.SVX"
Sub Main
        Dim X(252,13) As Double
        Dim Xt(13,252) As Double
        Dim Y(252,1) As Double
        Dim B(1,1) As Double
        Dim XtY(13,1) As Double
        Dim XtYt(1,13) As Double
        Dim W(13,1) As Double
        Dim T(252,1) As Double
        Dim spr As Spreadsheet
        Set spr = ActiveSpreadsheet

        Dim i As Integer
        Dim j As Integer
        Dim k As Integer
        Dim l As Integer
        Dim m As Integer
        Dim n As Integer
        Dim RT1 As Double
        Dim RT2 As Double
        Dim bbt As Double
        Dim C(1,1) As Double
        Dim D As Double
```

```
      Dim Tt(1,252) As Double
      Dim ST(1,1) As Double
      Dim Tk As Double
      Dim XtT(13,1) As Double
      Dim P(13,1) As Double
      Dim TtY(1,1) As Double
Dim PPt(13,13) As Double
Dim SS(13,13) As Double
Dim Stt As Double
Dim nn As Integer
Dim nnn As Integer
Dim bt(13,3) As Double
Dim Com(252,3) As Double
Dim Ldg(13,3) As Double
Dim CC(3,1) As Double
For n = 1 To 3
     MatrixTranspose(X,Xt)
     Call Bobot(Xt,Y,XtY,XtYt,C,D, W)
     Call Score(X,W,T)
     MatrixTranspose(T,Tt)
     Call SScore(Tt,T,ST)
     Call Loading(Xt,T,XtT,ST,Tk,P)
     MatrixMultiply(Tt,Y,TtY)
     TC(1,1) = TtY(1,1)/ST(1,1)
     MatrixTranspose(P,Pt)
     MatrixMultiply(T,Pt,TPt)
     MatrixSubtract(X,TPt,X)
     MatrixMultiply(T,TC,TTC)
     MatrixSubtract(Y,TTC, Y)

   For nn = 1 To 13
      bt(nn,n)=W(nn,1)
     Ldg(nn,n)=P(nn,1)
   Next nn

  For nnn = 1 To 252
     Com(nnn,n)=T(nnn,1)
  Next nnn
     Dim TC(1,1) As Double
     Dim TTC(252,1) As Double
     Dim Pt(1,13) As Double
     Dim TPt(252,13) As Double
  For i=1 To 13
     RT1 = 0
     For j=1 To 252
       RT1=RT1 + (spr.Cells (j,i+1))/252
     Next j
     For k =1 To 252
        X(k,i) = spr.Cells (k,i+1) - RT1
     Next k
 Next i
 RT2 =0
 For l =1 To 252
    RT2 = (spr.Cells (l,1))/252 + RT2
       Next l
  For m = 1 To 252
  Y(m,1)=spr.Cells (m,1) - RT2
  Next m
CC(n,1)=TC(1,1)
Next n
Dim btt(3,13) As Double
Dim bttXXt(3,252) As Double
Dim Comt(3,252) As Double
Dim Ldgt(3,13) As Double
Dim Lb(3,3) As Double
Dim Inv(3,3) As Double
Dim btInv(13,3) As Double
Dim Beta(13,1) As Double

MatrixTranspose(Ldg,Ldgt)
MatrixMultiply(Ldgt,bt,Lb)
MatrixInverse(Lb,Inv)
MatrixMultiply(bt,Inv,btInv)
MatrixMultiply(btInv,CC,Beta)
MatrixDisplay(Beta,"Parameter PLS")
End Sub
Sub Bobot(At() As Double, B() As Double, AtB() As Double,AtBt() As Double,C() As Double,D As Double, W() As Double )
      MatrixMultiply(At,B,AtB)
      MatrixTranspose(AtB,AtBt)
      MatrixMultiply(AtBt,AtB,C)
      D=1/Sqrt(C(1,1))
      MatrixElemMultiply(AtB,D,W)
End Sub

Sub Score(A()As Double,B() As Double,T() As Double)
      MatrixMultiply(A,B,T)
End Sub

Sub SScore(A()As Double,B() As Double,ST() As Double)
      MatrixMultiply(A,B,ST)
End Sub

Sub Loading(A()As Double,B()As Double,AB() As Double,ST() As Double,Tk As Double, P() As Double)
      MatrixMultiply(A,B,AB)
      Tk = 1/ST(1,1)
      MatrixElemMultiply(AB,Tk,P)
End Sub
```

Output program STATISTICA menghsilkan estimasi parameter PLS sebagai berikut

β_1 = -0.19; β_2 = -0.12; β_3 = -0.04; β_4 = -0.03; β_5 = 0.05; β_6 = -0.19; β_7 = 0.06; β_8 = 0.12; β_9 = 0.15; β_{10} = 0.16; β_{11} = 0.19; β_{12} = 0.27; β_{13} = 0.51

Dan perbandingan prediksi respon dengan data respon diperlihatkan pada gambar 1 secara multiple, dan terlihat jelas prediksi respon melingkupi nilai data respon.



Gambar 1. Prediksi Respon dan Data Respon

SIMPULAN

Permasalahan analisis regresi adalah adalah multikolenieritas yaitu, adanya korelasi antara variabel prediktor. Salah satu metode baru dapat diterapkan metode *Partial least square Regression* (PLS).

Dalam PLS matriks variabel prediktor didekomposisikan menjadi matriks komponen dan matriks loading. Komponen-komponen tersebut akan berperan sama seperti variabel prediktor. Selanjutnya metode pembentukan dan seleksi komponen menggunakan algoritma NIPALS.

Perlu ada penelitian lanjutan yang dapat memberikan hasil prediksi respon lebih baik dibandingkan dengan regresi PLS, baik dari nilai prediksi, standar error, koefisien determinasi sebagai ukuran kebaikan model dan juga dari bias prediksi parameter.

DAFTAR PUSTAKA

Abdi, Herve. 2010. *Partial Least Squares Regression and Projection on latent StructureRegression (PLS Regression)*. John Wiley & Sons,Inc

Gujarati, D. N., 2009, *Basic Econometrics Fourth Edition*, Mc Graw-Hill, New York.

Korkmazoglu Ozlem Berak Gulder, Kemalbay. 2012.*Econometrics application of partial least squares regression: anendogeneous growth model for Turkey*. Procedia - Social and Behavioral Sciences 62 (2012) 906 – 910

Nyoman, Mindra Jaya I Gede. 2009. Kajian Penanganan Multikolenieritas Dalam Analisis Regresi Menggunakan *Partial Least Square* Regression. *Prosiding Seminar Nasional Penelitian, Pendidikan dan Penerapan MIPA*. Yogyakarta Fakultas MIPA, Universitas Negeri Yogyakarta

StatLib-DatasetsArchive, Lib.stat.cmu.edu/datasets/Under filename bodyfat 2010

Ronald E, Walpole. 2012. *Probability & Statistics for enginer & Science*, Prentice Hall