Indeks Sosio-ekonomi Menggunakan Principal Component Analysis

Iwan Ariawan*

Abstract

In household survey, we could measure socio-economic status through income, expenditure and ownership of valuable goods. Measuring income and expenditure in developing countries has many weaknesses, therefore many researchers prefer to use the ownership of valuable goods as proxy of socio-economic status. Using ownership of valuable goods as proxy indicator creates another problem of having many variables for the socio-economic proxy. To show how to simplify many variables of ownership of valuable goods into 1 socio-economic index. Using prinicpal component analysis with Stata. Using Indonesia Demographic & Health Survey 2002-2003 data, 7 binomial variables of ownership of valuable goods and 3 ordinal variables of housing condition to construct socio-economic indices using principal component analysis (PCA), tetrachoric and polychoric correlation.We used Stata to construct the socio-economic index. Correlation matrices were derived using tetrachoric command for tetrachoric correlation and polychoric command for polychoric correlation. Two socio-economic indices were constructed, 1 index was based only on 7 binomial variables of ownership of valuable goods and 1 index was based on 7 binomial variables of ownership of valuable goods and 3 ordinal variables of housing conditions. PCA was used to construct those 2 indices. In 7 variables model, the socio-economic index could explain 57% variance and in 10 variables model, the socio-economic index could explain 54% variance. We also showed the use of xtile command to regroup the subjects based on quintile of socio-economic indices. PCA, tetrachoric and polychoric correlation could be used to construct socio-economic indices based on information of ownership of valueable goods and housing conditions.

Key words: Socio-economic indices, principal component analysis, tetrachoric correlation, polychoric correlation.

Abstrak

Pada penelitian survei, kita dapat mengukur tingkat status sosio-ekonomi rumah tangga melalui pemasukan, pengeluaran dan kepemilikan barang-barang berharga. Penggunaan variabel pemasukan dan pengeluaran di negara berkembang memiliki banyak kelemahan, sehingga banyak peneliti lebih suka menggunakan variabel kepemilikan barang berharga untuk mengukur status sosio-ekpnomi. Namun, penggunaan variabel kepemilikan barang berharga menimbulkan masalah lain, yaitu banyaknya variabel untuk mengukur status sosio-ekonomi. Tujuan penulisan ini adalah menyederhanakan banyak variabel kepemilikan barang berharga menjadi 1 indeks sosio-ekonomi. Data yang digunakan adalah data Survei Demografi Kesehatan Indonesia 2002-2003 yang memiliki 7 variabel binomial tentang kepemilikan barang berharga dan 3 variabel ordinal tentang keadaan rumah untuk membuat indeks sosio-ekonomi. Indeks dibentuk dengan menggunakan *principal component analysis (PCA)*, korelasi tetrakorik dan polikorik. Kami memperlihatkan bagaimana membuat indeks sosio-ekonomi dengan bantuan perangkat lunak Stata. Matriks korelasi tetrakorik dibentuk dengan perintah tetrachoric dan matriks korelasi polikorik dibentuk dengan perintah polychoric. Dua indeks sosio-ekonomi dibentuk, 1 indeks berdasarkan 7 variabel binomial kepemilikan barang berharga dan 1 indeks lagi berdasarkan ke 7 variabel binomial tersebut ditambah 3 variabel ordinal kondisi rumah. Kedua indeks dibentuk dengan prosedur PCA. Pada model 7 variabel binomial, indeks yang terbentuk dapat menjelaskan 57% varians kepemilikan barang berharga dan pada model 7 variabel binomial ditambah 3 variabel ordinal, indeks dapat menjelaskan 54% varians kepemilikan barang berharga dan kondisi rumah. Kami juga memperlihatkan penggunaan perintah xtile untuk membagi subyek penelitian menurut kuintil indeks sosio-ekonomi. PCA, korrelasi tetrakorik dan polikorik dapat digunakan untuk membentuk indeks sosio-ekonomi berdasarakan informasi tentang kepemilikan barang berharga dan kondisi rumah.

Kata kunci: indeks sosio-ekonomi, *principal component analysis*, korelasi tetrakorik, korelasi polikorik.

*Staf Pengajar Departemen Biostatistik Fakultas Kesehatan Masyarakat Universitas Indonesia

Salah satu masalah dalam penelitian yang melibatkan status sosio-ekonomi pada tingkat keluarga atau individu adalah bagaimana cara mengukur status sosio-ekonomi tersebut. Variabel status sosio-ekonomi umumnya berperan sebagai variabel bebas untuk analisis lanjut tentang pengaruh kemiskinan atau inekuitas dalam pelayanan kesehatan, kesakitan maupun kematian. Selain itu, variabel sosio-ekonomi juga dapat berperan sebagai *confounder* yang penting pada analisis faktor risiko tertentu terhadap pelayanan kesehatan, kesakitan ataupun kematian.

Ada 3 cara untuk mengukur status sosio-ekonomi dalam penelitian, yaitu dengan menanyakan penghasilan, pengeluaran dan kepemilikan barang-barang berharga. Pengukuran status sosio-ekonomi berdasarkan penghasilan memiliki banyak kelemahan pada penelitian di negara berkembang, karena banyak orang yang bekerja di sektor informal dan penghasilannya tidak tetap. Pengukuran melalui pengeluaran hanya akurat jika ditanyakan secara rinci pengeluaran untuk berbagai keperluan spesifik pada jangka waktu tertentu (mingguan, bulanan atau tahunan). Pengukuran melalui kepemilikan barang berharga banyak dilakukan pada penelitian di negara berkembang karena pengukuran ini relatif lebih mudah dilakukan dibandingkan dengan 2 pengukuran di atas dengan akurasi yang baik pula.[1]

Masalah yang kemudian timbul dari penggunaan kepemilikan barang berharga untuk mengukur status sosio-ekonomi adalah bagaimana menyatukan berbagai informasi tentang kepemilikan barang ini menjadi 1 indeks sosio-ekonomi. Banyak peneliti membuat indeks ini dengan cara menjumlahkan semua barang berharga yang dimiliki oleh keluarga tanpa membedakan nilai barang yang dimiliki. Pada cara ini nilai mobil sama dengan televisi, sehingga cara ini banyak dikecam oleh para ahli ekonomi. Cara lain yang lebih baik adalah dengan memberikan bobot sesuai dengan nilai barang yang dimiliki, misalkan mobil memiliki bobot 5 dan televisi memiliki bobot 1. Masalah yang timbul dari pembobotan ini adalah bagaimana menentukan besarnya bobot untuk masing-masing barang. Filmer dan Pritchett[2] mengusulkan penggunaan *principal componen analysis (PCA)* untuk membuat indeks sosio-ekonomi berdasarkan data kepemilikan barang berharga, sehingga subyektifitas peneliti dalam menentukan bobot dapat dihindari. Makalah ini membahas tentang cara penggunaan PCA untuk membuat indeks sosio-ekonomi berdasarkan kepemilikan barang menurut Filmer dan Pritchett[2] yang disempurnakan oleh Kolenikov dan Angeles.[3]

Analisis Faktor dan Principal Component Analysis (PCA)

Analisis faktor merupakan teknik statistik yang digunakan untuk menyederhanakan matriks korelasi antar variabel. Royce[4] membuat definisi faktor sebagai suatu dimensi atau konstruk yang merupakan bentuk ringkas dari korelasi antar variabel. PCA merupakan salah satu teknik analisis faktor untuk menyederhanakan matriks korelasi antar variabel. Perbedaan utama PCA dengan teknik analisis faktor lainnya adalah PCA akan berusaha untuk menjelaskan variasi korelasi antar variabel secara maksimal pada komponennya yang pertama.[5] Data dasar untuk PCA adalah matriks korelasi antar variabel, dan pada awalnya PCA digunakan untuk menyederhanakan korelasi antar variabel kontinyu.

Kepemilikan barang-barang berharga pada umumnya diukur dalam bentuk binomial (seperti memiliki televisi ya/tidak) atau dalam bentuk ordinal (seperti dinding utama rumah : 0=bambu, 1=kayu, 3=bata/batako) yang kurang sesuai untuk PCA. Filmer dan Pretcher[2] menggunakan metode PCA untuk variabel kontinyu dan tetap menggunakan matriks korelasi Pearson antar variabel kepemilikan yang bersifat binomial. Sedangkan untuk variabel yang bersifat ordinal, Filmer dan Pretcher membuat variabel *dummy*. Metode yang disarankan oleh Filmer dan Pretcher ini telah banyak digunakan peneliti lainnya untuk menilai hubungan antara status sosio-ekonomi dengan berbagai macam keluaran kesehatan, sosial dan pendidikan.[6-9]

Kolenikov dan Angeles[3] menyatakan penggunaan matriks korelasi Pearson tidak sesuai dengan sifat data kepemilikan yang binomial atau ordinal. Penggunaan variabel *dummy* pada matriks korelasi juga menimbulkan masalah adanya korelasi palsu (*spurious correlation*) antara variabel *dummy*.

Kolenikov dan Angeles[3] mengusulkan pemakai matriks korelasi tetrakorik untuk PCA dengan variabel kepemilikan binomial saja, atau korelasi polikorik untuk campuran variabel kepemilikan binomial dan ordinal. Uebersax[10] juga mempertegas pentingnya penggunaan korelasi tetrakorik dan polikorik untuk mengukur korelasi antar variabel kategori binomial dan ordinal.

Contoh Aplikasi PCA pada Data Survei Demografi Kesehatan Indonesia 2002-2003

Penggunaan PCA untuk membuat indeks sosio-ekonomi berdasarakan variabel kepemilikan yang bersifat binomial dapat dilakukan dengan prosedur korelasi tetrakorik pada perangkat lunak Stata.[11] Pembuatan indeks sosio-ekonomi dengan korelasi tetrakorik dapat dilakukan dengan perintah:

```
. tetrachoric var
```

Stata akan mengeluarkan matriks korelasi tetrakorik antar variabel binomial. Langkah selanjutnya adalah

menyimpan matriks korelasi ini ke dalam satu *array* 2 dimensi dalam memori komputer, dengan cara:

```
. matrix C = r(Rho)
```

Hasil dari perintah ini adalah *array* C yang berisi matriks korelasi tetrakorik. Langkah selanjutnya adalah perhitungan indeks sosio-ekonomi berdasarakan PCA:

```
.pcamat C, n(jumlah sampel) comp(1)
```

Perintah di atas meminta Stata untuk melakukan PCA dari *array* C yang berisi matriks korelasi tetrakorik. Hasil dari perintah ini adalah komponen/faktor mendasar (*underlying component/factor*) dari kepemilikan barang, yang kita anggap sebagai indeks sosio-ekonomi. Langkah selanjutnya adalah perhitungan indeks sosio-ekonomi untuk tiap subyek penelitian berdasarkan hasil PCA:

```
. predict  var
```

Hasil perintah di atas adalah variabel baru *var* yang berisi indeks sosio-ekonomi dan bersifat kontinyu. Jika kita menginginkan pengelompokan subyek menurut kuintil sosio-ekonomi, maka kita dapat melanjutkan dengan perintah:

```
. xtile var2 = var1, nq(5)
```

Hasil perintah di atas adalah variabel baru *var2* yang berisi kuintil indeks sosio-ekonomi menurut variabel *var1*.

Sebagai contoh, indeks sosio-ekonomi dibentuk berdasarkan 7 variabel kepemilikan, yaitu: listrik (HV206), radio(HV207), televisi(HV208), lemari es(HV209), sepeda/ sampan(HV210), sepeda motor/perahu motor(HV211), mobil/truk(HV212). Semua variabel diberi kode 0=tidak memiliki dan 1=memiliki. Prosedur pembentukan indeks sosio-ekonomi dengan PCA dan korelasi tetrakorik:

```
. tetrachoric  hv206 hv207 hv208 hv209 hv210
hv211 hv212
```

Stata memperlihatkan matriks korelasi tetrakorik antara 7 variabel kepemilikan. Interpretasi korelasi tetrakorik sama dengan interpretasi korelasi Pearson. Pembentukan indeks sosio-ekonomi dilakukan dengan PCA dengan memanfaatkan matriks korelasi tetrakorik yang sudah dihitung:

```
. matrix C=r(Rho)
. pcamat C, n(33088) comp(1)

Principal components/correlation                 Number of obs    =     33088
                                                 Number of comp.  =         1
                                                 Trace            =         7
    Rotation: (unrotated = principal)            Rho              =    0.5698

    --------------------------------------------------------------------------
       Component |   Eigenvalue   Difference         Proportion   Cumulative
    -------------+------------------------------------------------------------
           Comp1 |      3.98886      2.99862             0.5698       0.5698
           Comp2 |      .990238      .321926             0.1415       0.7113
           Comp3 |      .668311         .105             0.0955       0.8068
           Comp4 |      .563311     .0834187             0.0805       0.8872
           Comp5 |      .479892      .316611             0.0686       0.9558
           Comp6 |      .163281     .0171684             0.0233       0.9791
           Comp7 |      .146112            .             0.0209       1.0000
    --------------------------------------------------------------------------

          Principal components (eigenvectors)

                  -----------------------------------------------------
                           Variable |        Comp1 | Unexplained
                  ------------------+------------- -+------------------
                              hv206 |       0.4089 |       .3332
                              hv207 |       0.3045 |       .6302
                              hv208 |       0.4629 |       .1453
                              hv209 |       0.4557 |       .1717
                              hv210 |       0.1721 |       .8819
                              hv211 |       0.3608 |       .4807
                              hv212 |       0.3980 |       .3681
                  -----------------------------------------------------
```

Perintah `matrix C=r(Rho)` meminta Stata untuk menyimpan matriks korelasi ke *array* C. Hasil analisis menunjukkan indeks sosio-ekonomi yang terbentuk dapat menjelaskan 56,98% variasi kepemilikan 7 jenis barang dengan persamaan indeks sosio-ekonomi:

0,41*listrik+0,30*radio+0,46*tv+0,46*kulkas+0,17*sepeda+0,36*motor+0,40*mobil

Perhitungan indeks dapat dilakukan secara otomatis dengan prosedur:

```
. predict ses
```

Maka akan terbentuk variabel ses yang berisi indeks sosio-ekonomi berdasarkan kepemilikan dari 7 jenis barang dan memiliki skala pengukuran kontinyu. Jika indeks ini ingin dijadikan kuintil, dapat digunakan prosedur:

```
. xtile sespct = ses, nq(5)

. tab sespct

5 quantiles |
      of ses|      Freq.     Percent        Cum.
------------+-----------------------------------
           1 |      7,633       23.07       23.07
           2 |      6,272       18.96       42.02
           3 |      6,125       18.51       60.54
           4 |      7,689       23.24       83.77
           5 |      5,369       16.23      100.00
------------+-----------------------------------
       Total |     33,088      100.00
```

Maka pada variabel sespct akan berisi kode kuintil untuk sosio-ekonomi, mulai dengan kode 1 untuk kuintil terendah, atau kelompok paling miskin dan kode 5 untuk kuintil tertinggi, atau kelompok paling kaya.

Korelasi tetrakorik hanya dapat digunakan untuk variabel yang bersifat binomial, sedangkan jika variabel kepemilikan yang akan digunakan untuk membentuk indeks sosio-ekonomi bersifat ordinal, korelasi yang harus digunakan adalah korelasi polikorik. Prosedur `polychoric` tersedia gratis di internet yang dapat diunduh dengan perintah `findit polychoric`. Komputer Anda harus terkoneksi dengan internet pada saat perintah `findit` diberikan ke Stata.

Pembuatan indeks sosio-ekonomi dengan korelasi polikorik dapat dilakukan dengan perintah:

```
.polychoricpca var_kepemilikan,
score(var_skor) ns(1)
```

var_kepemilikan merupakan variabel-variabel tentang kepemilikan barang berharga dan *var_skor* merupakan variabel baru yang digunakan untuk menyimpan indeks sosio-ekonomi yang terbentuk. Pembagian merurut kuintil dapat dilakukan dengan perintah `xtile`, sama seperti pada korelasi tetrakorik.

Sebagai contoh, indeks sosio-ekonomi dibentuk berdasarkan 7 variabel kepemilikan binomial, yaitu: listrik (HV206), radio(HV207), televisi(HV208), lemari es(HV209), sepeda/sampan(HV210), sepeda motor / perahu motor(HV211), mobil/truk(HV212), sama seperti contoh untuk korelasi tetrakorik dan 3 variabel ordinal. Variabel-variabel ordinal yang diikutsertakan untuk pembentukan indeks sosio-ekonomi adalah: jenis jamban (HV205), bahan utama lantai rumah (HV213) dan bahan utama dinding rumah (HV214). Pilihan jawaban untuk jenis jamban: 0=halaman/semak-semak/hutan, 1=jamban cemplung, 2=sungai/danau, 3=jamban umum, 4=jamban pribadi tanpa tangki septik dan 5=jamban pribadi dengan tangki septik. Pilihan jawaban untuk bahan utama lantai rumah: 0=tanah, 1=bambu, 2=kayu, 3=batu bata, 4=ubin, 5=keramik/marmer/granit. Pilihan jawaban untuk bahan utama dinding rumah: 0=bambu, 1=kayu, 3=batu bata/batako. Kode jawaban untuk variabel ordinal sebaiknya dimulai dengan kode terendah untuk kelompok yang terendah dan harus sejalan dengan kode variabel binomial.

Prosedur pembentukan indeks sosio-ekonomi dengan PCA dan korelasi polikorik:

```
. polychoricpca hv205 hv206 hv207 hv208
hv209 hv210 hv211 hv212 hv213 hv214,
score(ses) nscore(1)
```

Stata akan memperlihatkan matriks korelasi polikorik 10 variabel kepemilikan dan PCA. Interpretasi korelasi polikorik sama dengan interpretasi korelasi Pearson. Pada tabel PCA, terlihat indeks sosio-ekonomi yang terbentuk dapat menjelaskan 54% variasi 10 variabel kepemilikan. Persamaan indeks sosio-ekonomi ditampilkan pada tabel *scoring coefficient.*

Persamaan indeks sosio-ekonomi dapat dituliskan sebagai:

$$SES = \gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 + \gamma_5 + \gamma_6 + \gamma_7 + \gamma_8 + \gamma_9 + \gamma_{10}$$

γ_1 merupakan koefisien untuk HV205, dengan nilai koefisien -0,534963 untuk HV205=0 (jenis jamban semak-semak/hutan), -0,306630 untuk HV205=1 (jenis jamban cemplung) dan seterusnya. Sedangkan γ_2 merupakan koefisien untuk HV206, γ_3 untuk HV207, dan seterusnya dengan nilai koefiesien sesuai tabel seperti untuk γ_1.

Stata secara otomatis menghitung indeks sosio-ekonomi ini dan disimpan dalam variabel ses1 dan-memiliki skala pengukuran kontinyu. Jika indeks ini ingin dijadikan kuintil, dapat digunakan perintah `xtile` sama seperti untuk korelasi tetrakorik.

```
Scoring coefficients

    Variable     |  Coeff. 1  |  Coeff. 2  |  Coeff. 3
-------------------------------------------------------
 hv205
              0  | -0.534963  |   0.277964 |   0.211176
              1  | -0.306630  |   0.159323 |   0.121042
              2  | -0.167047  |   0.086797 |   0.065942
              3  | -0.064265  |   0.033392 |   0.025369
              4  |  0.010589  |  -0.005502 |  -0.004180
              5  |  0.294198  |  -0.152864 |  -0.116135
 hv206
              0  | -0.531435  |   0.133106 |   0.274846
              1  |  0.101462  |  -0.025413 |  -0.052474
 hv207
              0  | -0.207256  |  -0.332861 |   0.196247
              1  |  0.175759  |   0.282275 |  -0.166423
 hv208
              0  | -0.366418  |  -0.089148 |   0.172410
              1  |  0.249750  |   0.060763 |  -0.117514
 hv209
              0  | -0.141822  |   0.021766 |   0.095287
              1  |  0.529370  |  -0.081246 |  -0.355672
 hv210
              0  | -0.086592  |  -0.539229 |  -0.297985
              1  |  0.114522  |   0.713151 |   0.394097
 hv211
              0  | -0.156427  |  -0.125512 |   0.050477
              1  |  0.328136  |   0.263287 |  -0.105887
 hv212
              0  | -0.055134  |   0.012915 |   0.016693
              1  |  0.643169  |  -0.150668 |  -0.194740
 hv213
              0  | -0.567728  |   0.483158 |  -0.826605
              1  | -0.400749  |   0.341053 |  -0.583485
              2  | -0.251819  |   0.214308 |  -0.366645
              3  |  0.023940  |  -0.020374 |   0.034856
              4  |  0.258983  |  -0.220405 |   0.377076
              5  |  0.501285  |  -0.426613 |   0.729865
 hv214
              0  | -0.506599  |   0.387525 |  -1.012133
              1  | -0.178806  |   0.136779 |  -0.357237
              2  |  0.220735  |  -0.168852 |   0.441006


. xtile sespct = ses1, nq(5)

. tab sespct

5 quantiles  |
    of ses1  |      Freq.     Percent        Cum.
-------------+-----------------------------------
           1 |      6,515       20.00       20.00
           2 |      6,578       20.20       40.20
           3 |      6,460       19.84       60.04
           4 |      6,530       20.05       80.09
           5 |      6,485       19.91      100.00
-------------+-----------------------------------
       Total |     32,568      100.00
```

Maka pada variabel sespct akan berisi kode kuintil untuk sosio-ekonomi, mulai dengan kode 1 untuk kuintil terendah, atau kelompok paling miskin dan kode 5 untuk kuintil tertinggi, atau kelompok paling kaya.

Kesimpulan

Kami telah memperlihatkan bagaimana cara menggunakan *Principal Component Analysis* (PCA) untuk membentuk indeks sosio-ekonomi berdasarkan informasi kepemilikan barang, sehingga dihasilkan satu variabel indeks sosio-ekonomi yang bersifat kontinyu. Perhitungan PCA berdasarkan matriks korelasi antar variabel kepemilikan dan karena variabel-variabel ini bersifat binomial atau ordinal, maka disarankan untuk menggunakan matriks korelasi tetrakorik atau polikorik.

Daftar Pustaka

1. Shimeles A, Thoenen R. *Poverty Profile: A Methodological Note on Measuring Poverty*. Makalah dipresentasikan pada ODI/ESPD Conference on Addressing Inequalities: Policies for Inclusive Development, Addis Ababa, 11-12 Juli 2005.
2. Filmer D, Pritchett L (2001). Estimating Wealth Effect Without Expenditure Data or Tears: An Application to Eductional Enrollments in States of India. *Demograpy 38, 115-132*.
3. Kolenikov S, Angeles G (2004). *The Use of Discrete Data in PCA: Theory, Simulations and Applications to Socioeconomic Indices*. Working paper WP-04-85, MEASURE/Evaluation Project, Carolina Population Center, University of North Carolina, Chapel Hill.
4. Royce JR(1963). Factors as Theoretical Constructs. pada Jackson DN, Messick S (editor), *Problems in Human Assessment*. McGrawHill, New York.
5. Kline P (1994). *An Easy Guide to Factor Analysis*. Routledge, London.
6. Bollen KA, Glanville JL, Stecklov G (2002). Economic Status Proxies in Studies of Fertility in Developing Countries: Does The Measure Matter?. *Population Studies*, 56, 81-96.
7. EQUINET 2003. *Deprivation and Resource Allocation: Methods for Small Area Research*. EQUINET, Harare.
8. Garenne M, Garenne SH (2003). A Wealth Index to Screen High-risk Families: Application to Morocco. *J. Health Population and Nutrition* 21(3), 235-242.
9. Barros AJ, Victora CG. *A Nationwide Wealth Score Based on The 2000 Brazillian Demographic Census*. Makalah dipresentasikan pada The VI Brazilian Congress of Epidemiology, Recfine, June 2004.
10. Uebersax JS (2006). *The Tetrachoric and Polychoric Correlation Coefficient*. Diunduh dari http://ourworld.compuserve.com/homepages/juebersax/tetra.htm pada 25 Juli 2006.
11. Stata Corp (2005). *Stata Statistical Software: Release 9*. Stata Press, College Station.