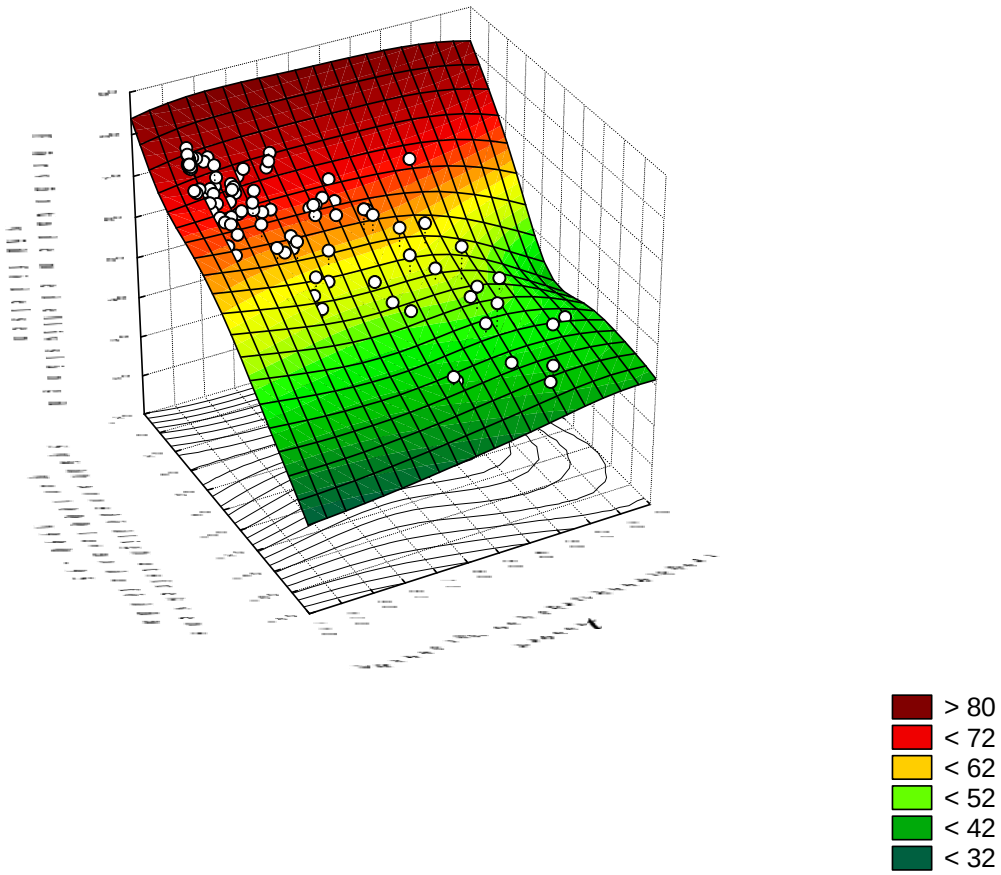
EKONOMIHÖGSKOLAN
Lunds universitet
Statistiska Institutionen

En analys av förväntad medellivslängd i världens länder 1996–97



Examensuppsats i statistik 61–90 hp
2009-05-18
Farhad H. Alinaghizadeh
Handledare: Mats Hagnell

Inledning

Många studier visar att medellivslängden hos befolkningar i världen varierar mellan länderna och orsakerna till det kan vara både socioekonomiska och demografiska. I denna uppsats beskrivs en analysteknik för att visa förväntad medellivslängd för människor över hela världen. Jag har använt en multivariat modell med multipelvariabler. Detta kan dock medföra *multikollinjäritetsproblem*; varför jag har valt *principalkomponentsanalys* för att undvika reducering av data. Resultatet visar att en enkel regressionsmodell med bara en principalkomponent som förklaringsvariabel i modellen går bra.

I uppsatsen presenterar jag en generell selektion av alla utvalda variabler (16 variabler) i en och samma modell, och enstaka beskrivande statistik av förklaringsvariabler presenteras med medelvärde, standardavvikelse och multipel korrelationsmatris.

Analyserna utförs först med analysmetoden multipel regression, därefter följer en analys av korrelationsstruktur (Gabriel's biplot) för att påvisa variabler som troligen har liten inverkan i detta sammanhang. För att undvika multikollinjäritetsproblem används ”ridge regression” som ett förslag, men detta löser inte problemet eftersom variablerna är högt korrelerade. Resultatet från analyserna jämförs med principalkomponentsanalys.

Slutligen jämförs resultaten från de olika metoderna med determinationskoefficienter R2 och R2(pred). PROC REG, ODS och PROC PRINCOMP i SAS version 9.2 används för att analysera data.

Nyckelord: Multipel regression (REG), principalkomponentsanalys (PCA), Multikollinjäritet, Kondition index, Ridge regression (RREG), Variance Inflation Factor (VIF), Predicted residual (PRESS), determinationskoefficienter (R2), Gabriel's biplot

Abstract

Many studies show that socioeconomic and demographic variables have a strong link with human health which in turn affects life expectancy and age. This paper presents the results of a principal component regression analysis, in order to describe a good regression model for mean age of the population of all the countries in the world. The data are from “Philips Geographical Digest” and “World Resources” Annuals 96–97).

This paper does not search for individual explanatory variables, but a multivariate model in which the entire variables (16 variables) are included in the same model. Use of all variables at the same time create a multicollinearity problem, therefore, principal component analysis were selected as the main analytical techniques to avoid reduction of data. The results of analysis show that we can use a simple regression model with only one principal component as the explanatory variable.

This is a general assessment of all the selected variables in one simple model from an statistical point of view, and occasional descriptive statistics are presented with mean, SE and intercorrelation matrix to improve multicollinearity.

The analysis starts with the classical method of multiple regressions, followed by ridge regression, principal component analysis and an analysis of the correlation’s structure by “Gabriel’s biplot” to detect the variables that have less impact in this context. To avoid problems with multicollinearity, “ridge regression” was used as the first alternative, but ridge regression did not solve the problem with multicollinearity because all the explanatory variables were strongly correlated. The results of ridge regression and multiple regression models were compared with results of principal component regression (pc1, pc1-pc4). Results from the different methods were compared by the determination coefficients R2 and R2(pred.). PROC REG, PROC PRINCOMP in SAS version 9.2 is used to analyze the data.

Bakgrund

Människors medellivslängd skiljer sig åt från land till land, och spannet ligger mellan 40 och 80 år (enligt ”Philips Geographical Digest” och ”World Resources” årsböcker 96–97). Det finns säkert många orsaker till skillnaderna, och i min uppsats försöker jag på ett enkelt sätt hitta den ”bästa modellen” som kan förklara detta, med data som kommer från dessa källor. Jag har valt ett praktiskt exempel av en statistisk metod (principalkomponentregression) från framför allt tillgängligheten av demografiska och ekonomiska förklaringsvariabler som har ett starkt samband med medellivslängd. Det finns mycket skrivet om medellivslängd, dels i olika samhällsvetenskapliga skrifter och dels i media t.ex. EU:s statistikbyrå Eurostat och världssamfunden m.m. och i samband med att man analyserar socioekonomiska faktorer.

Om vi bortser från medicinska och epidemiologiska faktorer och det faktum att ett land inte är homogent, finns det många faktorer som har ett starkt samband med landets genomsnittliga medellivslängd. Denna uppsats antar utmaningen och försöker på ett konkret sätt anpassa en modell med ett *fåtal förklaringsvariabler*, men som innehåller så *mycket information från data* som möjligt.

Data

Syftet med denna uppsats är att hitta en modell som kan beskriva den förväntade medellivslängden i världens länder under perioden 1996–1997 med hjälp av ett antal kontinuerliga förklaringsvariabler. Urvalet av förklaringsvariabler i denna studie är mer subjektivt, och när det gäller åldersgrupper, yrkesområden eller inkomstkällor i länderna är variablerna klassindelade och anger en andel av de observerade totala frekvenserna.

Källpresentation

I mitt arbete med att samla in data: har jag besökte flera bibliotek och ringt till föreningar/institutioner såsom Miljöinstitutet i Lund och Föreningen Miljöbiblioteket i Lund för att få ta på böcker med statistik som passade till studien. Sedan har jag registrerat data i datafiler för bearbetning.

De 16 variablerna som valts ut för analys, kommer från nedanstående källor:

1-WORLD RESOURCES (WR) the urban environment 1996–97

Denna bok publiceras en gång om året och enligt boken kommer de flesta inkluderade källorna från World Resources Institute, United Nations Environment Programme, United Nations Development Programme och World Bank.

2-Philips Geographical Digest (PG) 1996–97

Skillnaden mellan dessa två källor är att WR har reducerat observationerna till 152 länder och i PG finns över 174 länder, även utsprida öar som tillhör t.ex. USA, England och Frankrike.

Urval av data och antal förklaringsvariabler

Från ungefär 61 möjliga variabler har jag valt att använda 16 variabler (tabell 1) med X11 som responsvariabel (en gemensam variabel för båda könen). Första skälet är bortfall i de borttagna variablerna, vilka orsakar problem vid multivariata dataanaly-

ser och det andra är att korrelationen mellan dessa valda variabler är strak. Ett tredje skäl är variablernas skaltyp där de valda variablerna har ungefär samma skaltyp och mindre spridning (standardavvikelser). Responsvariabeln X11 är i hög grad korrelerad med alla valda förklaringsvariablerna. Korrelationen mellan X15 och responsvariabeln X11 ger det högsta korrelationskoefficienten (r=-0,96), vilket betyder att den första valda variabeln för modellanpassning i en regressionsanalys bör vara X15.

Tabell 1 Beskrivning av variabler och källor

X2: Befolkningsförändring Philips Geographical digest 1996 – 97 Heinemann
Befolkningsökning eller -minskning i % per år.
Anledning: Har storleken på befolkningsökningen eller -minskningen någon inverkan på landets medellivslängd?

X3: Stadsbefolkning i procent av befolkningen Philips Geographical Digest 1996 – 97 Heinemann
X4: Antal födda barn per kvinna Philips Geographical digest 1996 – 97 Heinemann
Antalet barn som en kvinna i genomsnitt föder under sin livstid.
Anledning: Har antalet barn som en kvinna föder under sin livstid någon betydelse för hur gammal hon blir?

X11: Förväntade medellivslängd vid födelsen 90 – 95 World Resources Tabell 8.2
Förväntad livslängd vid födelsen är medelvärdet av de antal år som ett nyfött barn förväntas leva om den åldersspecifika dödligheten är tillämplig under hela hans eller hennes livstid.
Anledning: Denna variabel är mer jämförbar med de andra som vi har i databasen eftersom den är från ungefär samma tid. Denna variabel är den mest väsentliga i hela uppsatsen. De flesta jämförelser skall göras med utgångspunkt från denna variabel.

X12: Andel av befolkningen i åldern 0 – 14 år Philips Geographical Digest 1996 – 97 Heinemann
X13: Andel av befolkningen i åldern 15 – 59 år Philips Geographical Digest 1996 – 97 Heinemann
X14: Andel av befolkningen i åldern 60+ år Philips Geographical Digest 1996 – 97 Heinemann
Befolkningen indelad i olika åldersgrupper.
Anledning: Har det någon betydelse för ett lands medellivslängd hur åldersfördelningen ser ut?

X15: Spädbarnsdödlighetsfrekvens per 1000 levande födda 1990 – 95 World Resources Tabell. 8.3
Spädbarnsdödlighetsfrekvensen per 1000 levande födda är sannolikheten att dö vid exakt 1 års ålder multiplicerad med 1000.
Anledning: Hög spädbarnsdödlighet ger troligen lägre medellivslängd än vad låg spädbarnsdödlighet ger. Denna variabel är intimt förknippad med medellivslängdsvariablerna.

X16: Födelsefrekvens per 1000 människor Philips Geographical digest 1996 – 97 Heinemann
Antal födda per 1000 människor.
Anledning: Påverkas medellivslängden om ett land har hög nativitet?

X17:Dödlighetsfrekvens per 1000 människor Philips Geographical digest 1996 – 97 Heinemann
Antal döda per 1000 människor.
Anledning: Denna variabeln tillsammans med den föregående påverkar landets befolkningstillväxt och därmed också landets förmåga att försörja sina medborgare med mat.

X42:Inkomstkällor i procent Jordbruk Philips Geographical digest 1996 – 97 Heinemann
X43:Inkomstkällor i procent Industri Philips Geographical digest 1996 – 97 Heinemann
X44:Inkomstkällor i procent Tjänstesektorn Philips Geographical digest 1996 – 97 Heinemann
Olika inkomstkällor. Dessa variabler ger svar på inom vilken typ av sysselsättning som människorna får sin inkomst.
Anledning: Spelar fördelningen av dessa tre variabler någon roll för människors förmåga att försörja sig?

X45: Andelen människor sysselsatta med jordbruk i procent av folkmängden Philips Geographical digest 1996 – 97 Heinemann
Motsvarande X42-X44.
Andelen av landets invånare sysselsatt med jordbruk.
Anledning: Har fördelningen av vad människor arbetar med någon betydelse för landets medellivslängd? Går det att påvisa något samband mellan t.ex. hög andel sysselsatta inom jordbruk och hög medellivslängd eller tvärtom?

X46: Andelen sysselsatta inom industrin i procent Philips Geographical digest 1996 – 97 Heinemann
Motsvarande X42–X44.
X47: Andelen sysselsatta med service i procent Philips Geographical digest 1996 – 97 Heinemann
Motsvarande X45.
X50: Vuxnas läs- och skrivkunnighet i procent Philips Geographical digest 1996 – 97 Heinemann
Denna variabeln behandlar människor över 15 år som kan läsa och skriva. UNESCO definierar en person som icke läs- och skrivkunnig om personen inte kan läsa och skriva en kort berättelse om sitt dagliga liv.
Anledning: Påverkar läs- och skrivkunnighet människors hälsa och därmed även deras livslängd?

Socioekonomiska variabler

I ett rikt land lever invånarna längre än invånarna i ett fattigt land. De ekonomiska variablerna är därför med för att de kan tänkas förklara en stor del av skillnaderna i

hälsa och medellivslängden mellan länder. Ett typiskt exempel är Kuwait och Afghanistan som ligger i samma geografiska område men skiljer sig åt dramatisk.

Demografiska variabler

Det har visat sig att alla demografiska förklaringsvariabler har ett starkt samband med i första hand hälsa och i andra hand medellivslängd. Bland de valda demografiska variablerna finns det ett antal förklaringsvariabler som kan förklara medellivslängden hos de olika länderna. Sådana variabler som spädbarnsdödlighet (X15), antal födda barn per kvinna (X4), födelsetal per 1 000 människor (X16) och dödstal per 1 000 människor (X17) har troligen stor betydelse när man skall förklara medellivslängden och självklart är de starkt korrelerade med varandra och med responsvariabeln (X11). Vuxnas läs- och skrivkunnighet i procent (X50) är en typisk variabel som ofta används för att visa skillnader mellan världsdelar.

Jag har med avsikt inte tagit med variabler av typen rökning, alkoholkonsumtion etc. eftersom studien är långt från en studie i medicin eller epidemiologi.

Problem med bortfall

Ett känt problem inom multivariata dataanalyser och multipelregression analys är bortfallet. Det räcker att en observation saknar ett värde så utesluts observationen från analysen. Det finns metoder för att ersätta saknade värden eller hantera problemet på ett korrekt sätt, men den tekniken har jag inte använt i denna uppsats. Eftersom variablerna i vårt fall skiljer sig väldigt mycket åt från land till land så blir inte det imputerade värdet representativt för landet ifråga. Anledningen till detta är att vissa länder är väldigt stora och andra är väldigt små så ifråga om rent numeriska värden skiljer sig t.ex. Ryssland och Danmark för mycket åt för att det skall vara meningsfullt att använda Ryssland som grund för att ersätta Danmarks saknade värde. Därför valde jag 144 av 174 möjliga länder, nämligen de som hade fullständiga data i de 16 valda variablerna (se appendix-3).

Metod

Denna uppsats demonstrerar och utvärderar användningen av principalkomponenter vid regressionsanalys samt beskriver metoderna för att välja de variabler som kan förklara medellivslängden i världens länder med så få förklaringsvariabler som möjligt i en regressionsmodell där alla förklaringsvariabler är starkt korrelerade med varandra. De underliggande dimensionerna av data kan belysas enklare och bättre med principalkomponentanalys än med konventionella regressionsmetoder. Förlusten av informationen är minimal och de variabler som inte är adekvata i principalkomponentanalys kan förkastas.

Multivariata metoder

Vid användning av principalkomponentsanalysmetod givet multikollinjäritets förhållande i data (redundant information t.ex. X42-X44 och X45-X47), kan man ändå behålla alla variabler i analysen, genom att varje principalkomponentsekvation karaktäriseras av ett antal variabler (genom deras tecken och proportionsförklaring av varians i principalkomponentsladdningar). Vidare kan man använda dessa skattade principalkomponentsladdningar som en förklaringsvariabel i en regressionsmodell eller andra analysmetoder (Rawlings 1988).

På jakt efter den bästa modellen

Vid regressionsanalys används ofta analys av residualer och inflytelserika observationer för att avgöra vilka modell som är bäst anpassade till data. Metoderna härleds inte här, t.ex. COOK'S D, DFFITS, DFBETAS och COVRATIO. Dessa diagnostiska metoder används ofta vid regressionsanalys och är tillgängliga via all statistisk mjukvara med tillgång till olika grafiska presentationer av analysresultat (t.ex. SAS, SPSS, STATISTICA, STATA, MiniTab).

Multikollinjäritetsproblem

Vid analyser av stora databaser med multivariata data gör man traditionellt regressionsanalyser för att förklara multipel relationen (R2) eller förklarad varians mellan förklaringsvariablerna mot en responsvariabel. Normalt innebär det att man utför många kontroller av normalitetsantagande, *multikollinjäritet*, residualer samt test av modellanpassning vilket leder till att den slutliga regressionsmodellen kan innehålla färre antal variabler eller nyskapade variabler. Att använda sig av olika tillämpade regressionsmetoder med hänsyn till dagens möjligheter är inte så svårt, men i själva verket är det tidskrävande och komplext eftersom man måste granska resultaten noga (residualer, diagnostiska test etc.) och överväga förhållandena mellan förklaringsvariabler (*multikollinjäritet, kollinjäritetsdiagnostik, normalitet etc.*).

Konditionsindex:

Konditionsindex är ett mått på multikollinjäritet i data.

$$\delta_k = \left(\frac{\lambda_{\max}}{\lambda_{\min}} \right) \qquad (1)$$

Belsley, Kuh och Welsch (1980) föreslår beräkningsmetoden ”konditionsindex” för att avgöra multikollinjäritet vid regressionsanalys, och vidare förslår de att ett tal på 30–100 indikerar måttlig till stark multikollinjäritet.

Skattning av VIF (Variance Inflation Factor) och Toleransvärde

Det finns många metoder för att upptäcka multikollinjäritetsproblemet, och en av dem är VIF (Variance Inflation Factor) som skattar *hur mycket av variansen i skattningar av regressionskoefficienten $\hat{\beta}$ är påverkad jämfört med när en oberoende variabel inte är linjärt relaterad*. Om vi anta att $\hat{\rho}$ är korrelationsmatrisen för de oberoende variablerna, diagonalelement i $\hat{\rho}^{-1}$ kallas för VIF.

$$VIF_j = \frac{1}{1-R_j^2} \qquad (2)$$

R_j^2=förklaringsgrad (determinationskoefficient) av X_i på övriga oberoende variabler

$$1 - R_j^2 = \text{tolerans} \qquad (3)$$

Toleransvärde ($1 - R_j^2$) mindre än 0,01 anger dålig anpassning av variabler (Berk 1977).

Ridge regression

Ridge regression är utvecklad av Hoerl och Kennard (1970). Metoden är baserad på förändringen av minsta kvadratmetoden som tillåter ensidigt ”biased” skattningar av regressionslinjens koefficienter. Reducerad varians vid ridge regressionsskattningar ger ofta mindre medelkvadratfel (MSE) jämfört med den sedvanliga minstakvadratmetoden (OLS eller BLUP). Dessa skattningar är förslagsvis mer objektiva, eftersom de har en större sannolikhet att vara närmare de sanna parametervärdena (John 1984).

I regressionsanalyser är *multikollinjäritet* ett problem när två eller flera förklaringsvariabler är korrelerade med varandra eller är beroende av varandra. Med multikollinjäritet menar vi att två eller flera av variablerna är redundanta i modellen eller har samma information. Linjärt beroende i skattningar av regressionskoefficienter påverkas av detta. Redundant information betyder att vad en variabel förklarar om Y är exakt samma som en annan variabel förklarar. I detta fall kan två eller flera redundanta förklaringsvariabler bli fullständigt opålitliga eftersom b_i skulle skatta samma effekt för x_i som andra b. Vidare skulle $(X'X)^{-1}$ inte existera för att nämnaren 1-r2 är noll. Som en konsekvens av detta kan värdet för $\hat{\beta}_i$ inte skattas eftersom elementen för inversmatriserna och koefficienterna blir ganska stora (Younger 1979). En annan konsekvens av multikollinjäritet är att variansskattningarna vid minsta kvadratskattningarna blir för stora, vilket i sin tur ger bredare konfidensintervaller; så ju mer multikollinjäritet det finns, desto mindre tolkningsbara parametrar. Kort sagt, multikollinjäritet påverkar alla skattningar vid regressionsanalys, minsta kvadratskattning av regressionskoefficienter, medelkvadratfel (MSE), t-test, kvadratsummor och sist determinationskoefficienten R2.

Det finns många metoder för att hantera multikollinjäritet och några av dem är ridge regression, principalkomponentsanalys, partial least squares regression[1] och continuum regression[2] (Belsley, 1980). Som nämnts har koefficientskattningar i multipel linjär regression sitt krav på att modellparametrar ska vara oberoende (förklaringsvariablernas förhållande). När förklaringsvariablerna är korrelerade och kolumnerna i designmatris X har ett approximativ linjärt beroende, blir matrisen $(X'X)^{-1}$ nästan singulär. Som en följd av den minsta kvadratskattningen, blir ekvationen

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (4)$$

mycket känslig för slumpmässiga fel i den observerade responsvariabeln (y), vilka producerar en stor varians. Denna situation av multikollinjäritet kan uppstå, till exempel när data samlas in utan en experimentell design. Ridge regression löser problemet genom att skatta regressionslinjens koefficienter med hjälp av

$$\hat{\beta} = (X'X + \delta I)^{-1}X'y \qquad (5)$$

där δ>0 är ridge parameter (minskningsparameter) och I är identitetsmatris. Små positiva värden av δ förbättrar skattningarna och minskar variationen av skattningar, medan bias-skattning av den minskade kvadratsumman av ridge-skattningar ofta resulterar i ett mindre medelkvadratfel (SE) jämfört med de minsta kvadrat-

[1] Geladi and Kowalski (1986), “*Partial least-squares regression: A tutorial*,” Analytical Chimica Acta, 185, 1-17

[2] Stone, M. and Brooks, R. J (1990) “*Contium regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion)*”. J. R Statist. Soc. B, 52, 237-269; corrigendum, 54 (1992), 906-907.

skattningarna. Matris av $(X'X)$ ersätts med $(X'X + \delta I)$, och δ är en liten positiv kvantitet som i sin tur förändrar den V:s diagonalmatris där

$$V'(X'X + \delta I)V = \begin{bmatrix} \lambda_1 + \delta & 0 & \cdots & 0 \\ 0 & \lambda_2 + \delta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k + \delta \end{bmatrix} \quad (6)$$

Egenvärdet av den nya matrisen $(X'X + \delta I)$ är $\lambda_i + \delta$ för *i=1,2,…, k* där inkludering av δ till huvuddiagonalen ersätter λ_i med $\lambda_i + \delta$. En av egenskaperna hos δ i ridge-skattning är att dämpa variationen av skattningar. Effekterna av egenvärden på varianser av ridge-regressionskoefficienter kan illustreras som

$$\sum_i \frac{Var(b_{i,R})}{\sigma^2} = \sum_i \frac{\lambda_i}{(\lambda_i + \delta)^2} \quad (7)$$

Därför kan δ i ridge regression dämpa de skadliga effekterna av små egenvärden som leds till kollinjäritet. Det finns flera metoder för att välja minskningsparametern δ. Ett alternativ inför val av minskningsvärde för δ kan vara den grafiska metoden som kallas för ”ridge trace”.

Skattning av residual och standardiserade residual

Metoden används för att upptäcka inflytelserika observationer i ett regressionssammanhang. Antagandet för avstående av i;te data punkt från center av X-värdena förutsätts med v_{ii}, diagonalelement av p-matis (hat-matris)

$$e = y - \hat{y} = y - py = (1 - p)y \quad (8)$$

$$r_i = \frac{e_i}{s\sqrt{1 - v_{ii}}} \quad r_{ii} \approx N(0,1) \quad (9)$$

Principalkomponentanalys

Principalkomponentanalys presenterades för första gången av Karl Pearson (1901). Han trodde att denna metod var det korrekta sättet att hantera vissa uppgifter som innehåller för många variabler (X_i), vilka då var intressanta för biometriker (biostatistiker). Han tillämpade den aldrig med mer än två eller tre variabler. En praktisk förklaring av principalkomponentanalysmetoden kom först senare (Hotelling 1933). Tillämpning av de viktigaste beståndsdelarna diskuteras av (Rao 1964), (Cooley och Lohnes 1971), och (Gnanadesikan 1977). Ytterligare framställningar av komponentanalys finns i (Kshirsagar 1972), (Morrison 1976), och (Mardia 1979).

Principalkomponentanalys används också för att analysera polynomiala relationer och för att finna multivariata extremobservationer i data (Gnanadesikan 1977). Principalkomponentanalys är relaterad till faktoranalys, korrespondensanalys, och biased regressionsanalys (Mardia 1979).

Principalkomponentanalys är en linjär kombination av p observerade variabler, där den första komponenten antas innehålla den mesta variansen mellan dessa observerade variabler $X_1,\ldots,X_p$. Principalkomponentanalysens grundteori bygger på det som kallas för "ellipsfördelning" och standardavvikelsen. Som en analys av kovarians-

struktur, är principalkomponentanalys oberoende av lokaliseringsparametrar. Metoden antar, att alla variabler har medelvärdet lika med 0, annars måste variabel X_j transformeras enligt ($X'_j=X_j-\mu_j$). Kovariansmatrisen är inte ändrad i sina geometriska beräkningar, utan denna transformation motsvarar ett skifte där ursprungliga koordinater förskjuts till p-antal medelvärden (X_1, .., X_p). Den huvudsakliga tanken bakom detta är att hitta en linjär kombination som innehåller den största variansen, som kan betraktas som den linjära kombinationen med koefficienter b_{11}, ..., b_{1p} vilka ska definiera alla *p*-variabler.

$$Z_1 = b_{11}X_1 + b_{12}X_2 + .. + b_{1p}X_p \quad (10)$$

Koefficienterna (b_{1j}) bestäms så att Z_1 har den maximala variansen. I dessa sammanhang, är problemet ej väl definierade, varians av Z_1 kan bli godtyckligt stor, helt enkelt med det stigande värdet av (b_{1j})

$$b_{11}^2 + b_{12}^2 + \cdots + b_{1p}^2 = 1, \;\; el \;\; \sum_{j=1}^{p} b_{ij}^2 = 1 \quad (11)$$

Transformationen Z_1 kan beskrivas som en projektion av *p*-dimensionella punktkluster in i den räta linjen, vilka i resultatet har den största variabiliteten. Z_1 kallas första principalkomponent (eller principalkomponentspoäng) av variablerna X_1,....., X_p. Den andra principalkomponenten blir,

$$Z_2 = b_{21}X_1 + b_{22}X_2 + ... + b_{2p}X_p \quad (12)$$

dvs. den andra linjära kombinationen, vilken återigen har den största variansen och $Var(Z1)=\lambda_1$ är maximal inom alla linjära kombinationer av X

$$Var(Z_1) > Var(Z_2) >.....> Var(Z_p) \quad (13)$$

men denna gång under förutsättning att Z_2 är okorrelerade med Z_1, och Z_2 måste återigen blir normaliserad. Geometriskt betyder detta att den andra principalkomponenten måste vara lodrät mot den första. Proceduren är då analog. Då h;te steget (2<h<p) kan blir formulerat in på generellt sätt som

$$Z_i = b_{i1}X_1 + b_{i2}X_2 + ...+ b_{ip}X_p \quad (14)$$

och vi hoppas kunna fastställa principalkomponentskoefficienterna med linjärkombinationen eftersom Z_i har maximal varians mellan alla linjära kombinationer vilka är okorrelerade med Z_1, ... , Z_{i-1}. Denna beskrivning ville säga att h;te linjärkombinationer är lodräta mot alla de andra principalkomponenterna. Z_i kallas för i;te principalkomponent.

”Gabriel‘s Biplot”

Korrelationsstruktur mellan förklaringsvariabler presenteras med Gabriel’s Biplot (Gabriel 1971, 1972, 1978). Detta är en formativ presentation av data och förhållande mellan variabler och observationer som visar:

1) Relationerna mellan förklaringsvariabler (oberoende variabler)
2) Relativa likheter mellan förklaringsvariabler (oberoende variabler)
3) Relativt värde av observationer gentemot förklaringsvariabler

Namnet ”biplot” kommer från metoden där både rader (observationer) och kolumner (variabler) är uppsatta på samma graf. Man brukar rita en graf av de två första principalkomponenterna som tillsammans förklarar

$$\frac{\lambda_1+\lambda_2}{\sum \lambda_i} \qquad (15)$$

Biplot använder singularvärde av Z där

Z=UL$^{1/2}$V’
L$^{1/2}$ och V kan skattas från Z’Z
V är matris av egenvektor
L är diagonalmatrisen av kvadratrötterna av egenvärdena

”Scree plot”

Screeplot visualiserar egenvärde hos kovariansen eller korrelationsmatrisen. Cattell (1966) presenterade ”scree plot” som ett verktyg för att bestämma ett antal viktiga komponenter eller faktorer i multivariata analysmetoder. Ordet ”scree” används för att referera till högsta egenvärde som en högsta punkt i ett berg och minsta egenvärde som en nedförsbacke till lägsta nivån av bergen. Exempel på scree plot kan hittas i litteratur som handlar om principalkomponentsanalys. De flesta statistiska mjukvarorna t.ex SAS, SPSS, STATA eller MiniTab har en inbyggd funktion för att automatiskt efter en principalkomponentsanalys få grafen i olika varianter av scree plot som har utvecklats under de senaste åren. Jag använder metoden i syfte att visualisera styrkan hos de första principalkomponenterna.

Principalkomponentsregression

Principalkomponentsregression definierar alla skattade parametrar i form av centrerade och skalade prediktorer av den ursprungliga variabeln X. Till skillnad från andra metoder, som PLSR- och SIMPLS[3]-metoder, väljer principalkomponentsregression X-viktad/X-parametrar utan hänsyn till responsvariabel. X-parametrarna som är valda för att förklara så mycket variation i X som möjligt, X-viktad från principalkomponentsregressionsmetoden är egenvektorer av prediktorer i en kovariansmatris X’X. Återigen, X- och Y-laddningar definieras som i PLSR, men som i SIMPLS är det lätt att beräkna övergripande modellkoefficienter för den ursprungliga responsen Y (centrerad och skalad) i form av det ursprungliga prediktorerna X.

> ”Principalkomponentsregression” är särskilt användbart när användaren förstår sammanhangen tillräckligt bra mellan X-variabler för att kunna tolka regressionsparametrar. Det är viktigt med kunskap om variabler som kan hjälpa till att avgöra vilka av dem som påverkar den andra och vilka som anger samma information och varför.

Jag byggde min principalkomponentsregressionsmodell med skattade principalkomponentsfaktorer (Z_i) eftersom syftet var att ha en modell med minsta antalet förklaringsvariabler. Den slutliga modellen innehåller den första principalkomponenten, och parametern tolkas som en generell bedömning av några variabler samtidigt.

[3] SIMPLS algorithm of; de Jong, S. (1993), “SIMPLES: An alternative approach to partial least squares regression,” Chemometrics and Inteligent Laboratory System, 18, 252-263

PRESS-kriterier och R2(pred)

Som ett mått för att jämföra regressionsmodellerna är PRESS och R2(pred) ett bra alternativ. Om modellen är stabil och välanpassad borde inte R2 skilja sig mycket från R2(pred).

Skattning av kvadratsumman

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2 \qquad (16)$$

$\hat{y}_{(i)}$ är det förväntade värdet för den i:te observationer i modellen. PRESS används som ett mått på hur väl en regressionsmodell är anpassad. Det är en beräkning som liknar R2- statistik för prognostisering av modellen.

PRESS statistik beräkning är:

$$R^2_{prediction} = 1 - \frac{PRESS}{SS_T} \qquad (17)$$

Dataanalys och resultat

För att skatta principalkomponentskoefficienter eller latenta vektorn har jag två alternativ att utgå från, antingen från en *kovariansmatris* eller från en *korrelationsmatris*. Dessa båda matriser ger olika svar. Jag har valt korrelationsmatrisen eftersom en kontroll med kovariansmatris visade att skalskillnader kan skapa problem vid skattningar. Om de ursprungliga variablerna inte är korrelerade så gör principalkomponentanalysen ingen nytta (Bryan 1980). Vi vet redan att de 16 valda förklaringsvariablerna är starkt korrelerade med varandra och det finns heller inget bortfall.

Deskriptiv analys av data

Medelvärde och spridning kring medelvärdet för utvalda variabler visas i tabell 2. Alla variabler är som redan nämnts signifikant korrelerade med varandra förutom de markerade variablerna i viss kombination i tabell 2. Detta fenomen är vanligt när man hämtar data från ett stängt system där alla förklaringsvariabler är korrelerade med varandra och kausalitet är ett faktum (t.ex. all möjlig information inom ett företag).

Eftersom alla variabler mäts på olika skalor, kommer jag använda korrelationsmatrisen i stället för kovariansmatrisen vid principalkomponentanalys. Dessa socioekonomiska och demografiska variabler har en stark interkorrelation och ett starkt samband med huvudvariabeln medellivslängd.

En annan anmärkning är att antalet länder i den ursprungliga databasen är reducerade från 174 till 144 länder på grund av bortfall och ej fullständiga data. För länder som är med analyserna se appendix-3.

Tabell 2 Deskriptiv analys- och korrelationsmatris

	Means	Std.Dev.	Y	X2	X3	X4	X12	X13	X14	X15	X16	X17	X42	X43	X44	X45	X46	X47	X50
X11	64,4	10,4	1,00	-0,39	0,73	-0,87	-0,76	0,60	0,63	-0,96	-0,88	-0,78	-0,72	0,44	0,46	-0,88	0,77	0,80	0,83
X2	2,0	1,8	-0,39	1,00	-0,30	0,65	0,61	-0,41	-0,60	0,43	0,60	0,10	0,21	-0,12	-0,14	0,40	-0,42	-0,32	-0,51
X3	50,2	23,2	0,73	-0,30	1,00	-0,66	-0,63	0,54	0,56	-0,71	-0,67	-0,48	-0,72	0,37	0,49	-0,82	0,66	0,78	0,64
X4	4,0	1,9	-0,87	0,65	-0,66	1,00	0,89	-0,71	-0,74	0,86	0,96	0,56	0,64	-0,41	-0,40	0,79	-0,77	-0,67	-0,83
X12	35,4	10,5	-0,76	0,61	-0,63	0,89	1,00	-0,74	-0,89	0,75	0,94	0,28	0,58	-0,37	-0,34	0,71	-0,76	-0,58	-0,72
X13	55,6	7,0	0,60	-0,41	0,54	-0,71	-0,74	1,00	0,48	-0,58	-0,73	-0,34	-0,46	0,37	0,21	-0,56	0,59	0,45	0,57
X14	8,8	5,5	0,63	-0,60	0,56	-0,74	-0,89	0,48	1,00	-0,64	-0,80	-0,09	-0,47	0,22	0,35	-0,61	0,63	0,51	0,62
X15	52,1	42,3	-0,96	0,43	-0,71	0,86	0,75	-0,58	-0,64	1,00	0,87	0,75	0,71	-0,46	-0,44	0,85	-0,78	-0,76	-0,85
X16	29,9	13,5	-0,88	0,60	-0,67	0,96	0,94	-0,73	-0,80	0,87	1,00	0,52	0,63	-0,41	-0,37	0,80	-0,81	-0,66	-0,82
X17	10,4	4,3	-0,78	0,10	-0,48	0,56	0,28	-0,34	-0,09	0,75	0,52	1,00	0,51	-0,40	-0,25	0,64	-0,50	-0,61	-0,65
X42	21,2	15,7	-0,72	0,21	-0,72	0,64	0,58	-0,46	-0,47	0,71	0,63	0,51	1,00	-0,54	-0,65	0,72	-0,63	-0,67	-0,59
X43	31,9	12,5	0,44	-0,12	0,37	-0,41	-0,37	0,37	0,22	-0,46	-0,41	-0,40	-0,54	1,00	-0,23	-0,51	0,59	0,38	0,39
X44	46,8	14,2	0,46	-0,14	0,49	-0,40	-0,34	0,21	0,35	-0,44	-0,37	-0,25	-0,65	-0,23	1,00	-0,39	0,21	0,45	0,35
X45	38,8	29,2	-0,88	0,40	-0,82	0,79	0,71	-0,56	-0,61	0,85	0,80	0,64	0,72	-0,51	-0,39	1,00	-0,83	-0,94	-0,78
X46	20,5	12,1	0,77	-0,42	0,66	-0,77	-0,76	0,59	0,63	-0,78	-0,81	-0,50	-0,63	0,59	0,21	-0,83	1,00	0,61	0,69
X47	40,5	20,3	0,80	-0,32	0,78	-0,67	-0,58	0,45	0,51	-0,76	-0,66	-0,61	-0,67	0,38	0,45	-0,94	0,61	1,00	0,70
X50	74,7	24,6	0,83	-0,51	0,64	-0,83	-0,72	0,57	0,62	-0,85	-0,82	-0,65	-0,59	0,39	0,35	-0,78	0,69	0,70	1,00

Markerade variabler är ej signifikanta vid korrelationstest. (X2 med X17, X43, X44 och X14 med X17). Alla variabler är starkt korrelerade med variabeln medellivslängd (X11).

Jag vill undersöka om kombinationen av två metoder, principalkomponentanalys och regressionsanalys kan vara användbara för att analysera multivariata data i en modell utan att förlora information givet multikollinjäritetsproblem i data. Principalkomponentanalysmetoden kan även användas i syfte att upptäcka extrema observationer i data.

Principalkomponentanalys

Principalkomponentsvektorer (b_{ij}) med hjälp av korrelationsmatrisen visas i tabell 4. Från tabell 3 kan man avläsa att den första principalkomponenten kan förklara 62 procent av den totala variansen mellan variablerna.

Egenvärdet (varians) för den första principalkomponenten är lika med 9,90 och eftersom vi utgår från korrelationsmatrisen utgör detta 62 procent av den totala variansen (sum Var(λ_i)=16,00). Det beräknade konditionsindexet enligt ekvation-1 för PC1-PC16 blir 48,53.

$$\delta_k = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) = \left(\frac{9{,}8953}{0{,}0042}\right) = \sqrt{2365{,}024} = 48{,}5389$$

Det beräknade konditionsindexet för PC1-PC4 blir 3,56 vilket indikerar icke kollinjära förhållanden mellan de första fyra komponenterna, därför kan en regressionsmodell med de första fyra komponenter accepteras. Dessa fyra komponenter tillsammans förklarar 86 procent av den totala variationen i data (se tabell 4).

Tabell 3 Egenvärde (λ_i)och proportionsförklaring av totalvariationen.

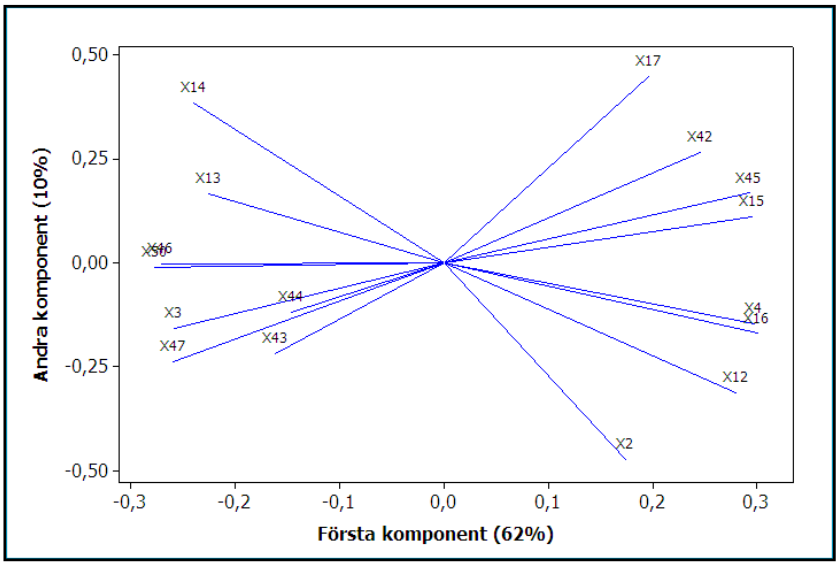
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
PC1	9,90	61,85	9,90	61,85
PC2	1,62	10,14	11,52	71,98
PC3	1,37	8,58	12,89	80,56
PC4	0,78	4,90	13,67	85,46
PC5	0,59	3,72	14,27	89,18
PC6	0,48	3,01	14,75	92,19
PC7	0,41	2,57	15,16	94,76
PC8	0,25	1,59	15,42	96,35
PC9	0,21	1,29	15,62	97,63
PC10	0,17	1,04	15,79	98,68
PC11	0,09	0,55	15,88	99,23
PC12	0,05	0,28	15,92	99,51
PC13	0,03	0,21	15,96	99,72
PC14	0,03	0,17	15,98	99,89
PC15	0,01	0,08	16,00	99,97
PC16	0,00	0,03	16,00	100,00

Gabriel‘s Biplot

Med hjälp av ”Gabriel‘s Biplot” (Gabriel 1971, 1972, 1978, 1981; Carroll, 1972) kan vi enkelt visualisera förhållandena mellan alla förklaringsvariabler för att se hur det euklidiska avståndet mellan variablerna ser ut. Figur 1 som förklarar 71,98 procent av variationen i data från första och andra principalkomponent och visar variabler som har negativ eller positiv laddning i den första principalkomponenten i förhållande till den andra principalkomponenten.

Här kan man tydligt avgöra vilka variabler som inte ger någon laddning vid de två första principalkomponenterna, dvs. variabler som ligger nära centrum i koordinatsystemet (t.ex. X44 som har den minsta laddningen vid de två första principalkomponenterna).

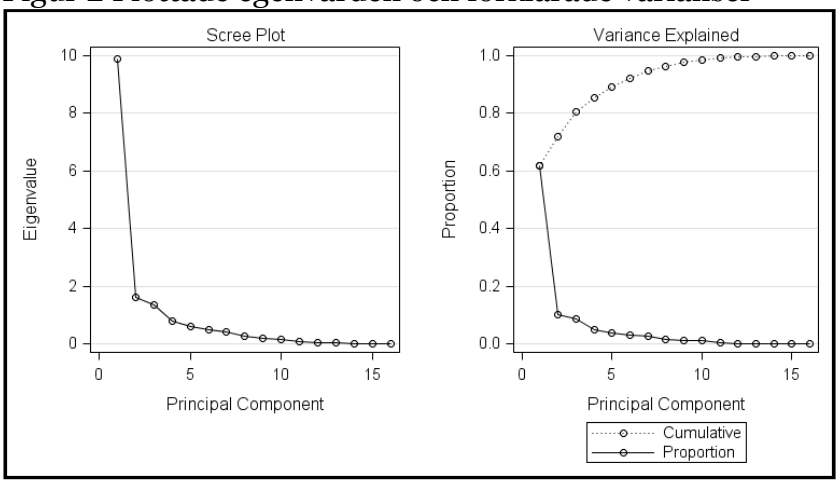
Figur 1 Gabriel‘s biplott av den första och andra principalkomponenten.



Grafen är gjord med MiniTab

Som redan nämnts kan variabler som är nära 0 vid den första och andra principalkomponenten betyda att dessa variabler vid detta sammanhang inte är så givna t.ex. X2, X14, X17, X44 och X43 som är signifikanta vid regressionsanalys (se tabell 6a). Man måste utgå från betydelsen för användning av dessa variabler i analysen för att bedöma om de ska vara med eller ej. De nämnda variablerna var även signifikanta vid korrelationsmatrisen också, detta betyder att dessa variabler ej kan inkluderas i samma modell som de övriga variablerna pga det gör modellen instabil.

Figur 2 Plottade egenvärden och förklarade varianser



Grafen är gjord med SAS.

Figur 2 visar hur egenvärdet minskar efter den andra och tredje principalkomponenten och hur förklarade variansen ökar efter första principalkomponenten. Här kan jag välja bland fyra principalkomponenter förutom den första. Dessa förklarar 86 procent av den totala variansen. Jämfört med utgångspunkt från kovariansmatrisen ger de fyra första principalkomponenterna 87 procents förklaringsgrad. Den första principalkomponentens varians är tillräckligt stor för att man ska acceptera analysresultatet, vilket tyder på att linjärkombinationer av data eller analys med 16 variabler kan vara acceptabelt.

Tabell 4 Principalkomponentsladdningar

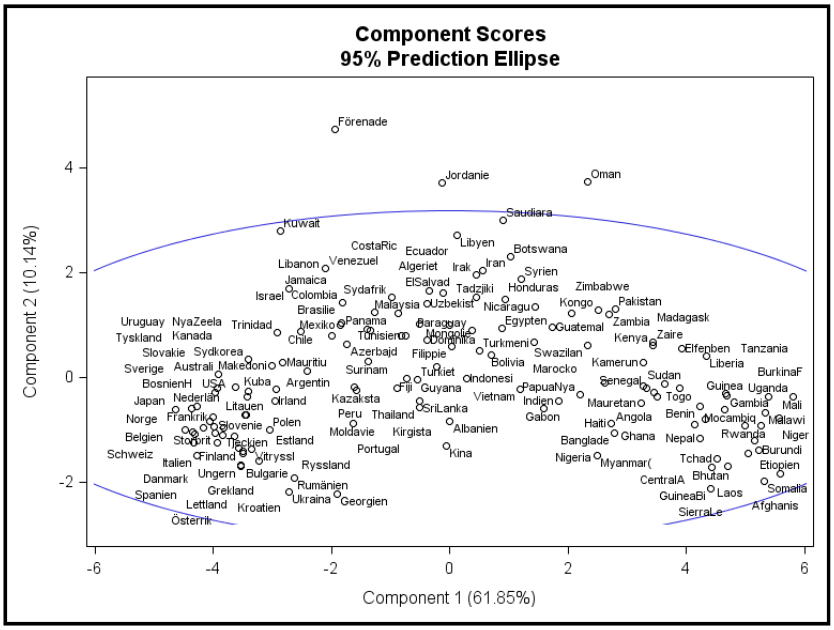
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16
X2	0,2	-0,5	0,0	-0,3	0,3	-0,2	0,7	0,2	0,0	0,0	-0,2	0,1	0,1	-0,1	0,0	0,0
X3	-0,3	-0,2	0,1	-0,2	-0,1	-0,4	-0,2	-0,1	-0,7	-0,3	-0,1	0,1	0,0	0,0	0,0	0,0
X4	0,3	-0,1	0,0	-0,2	-0,1	-0,1	0,0	0,0	-0,1	0,2	0,5	-0,3	-0,6	0,3	0,2	0,0
X12	0,3	-0,3	0,0	0,2	-0,1	0,0	-0,2	-0,1	-0,1	0,1	0,1	-0,1	0,5	-0,3	0,6	0,0
X13	-0,2	0,2	-0,1	-0,1	0,7	-0,4	-0,3	0,1	0,1	0,1	0,3	-0,2	0,1	-0,1	0,1	0,0
X14	-0,2	0,4	0,1	-0,2	-0,3	0,1	0,4	0,2	-0,1	-0,2	0,3	-0,5	0,3	0,0	0,1	0,0
X15	0,3	0,1	0,0	-0,2	0,0	-0,2	-0,2	0,0	0,0	0,2	-0,6	-0,6	0,1	0,2	0,0	0,0
X16	0,3	-0,2	0,0	-0,1	-0,1	0,0	-0,2	0,0	-0,1	0,2	0,3	-0,1	0,4	-0,1	-0,7	0,0
X17	0,2	0,4	0,1	-0,5	-0,2	-0,1	-0,1	0,1	0,1	0,3	0,0	0,5	0,0	-0,2	0,1	0,0
X42	0,2	0,3	-0,2	0,4	-0,1	-0,4	0,2	-0,1	0,0	-0,1	0,0	0,2	0,3	0,5	0,1	0,1
X43	-0,2	-0,2	-0,6	-0,4	-0,1	0,3	-0,2	0,2	0,0	0,0	0,0	0,1	0,2	0,4	0,1	0,0
X44	-0,1	-0,1	0,7	-0,1	0,2	0,2	-0,1	-0,1	0,1	0,1	0,1	0,1	0,2	0,5	0,1	0,0
X45	0,3	0,2	0,0	0,0	0,3	0,3	0,0	0,1	-0,3	-0,1	0,0	0,0	0,0	-0,1	0,0	0,8
X46	-0,3	0,0	-0,2	-0,1	-0,1	0,0	0,2	-0,8	0,0	0,3	0,0	-0,1	0,1	0,0	0,0	0,3
X47	-0,3	-0,2	0,1	0,0	-0,4	-0,4	-0,1	0,3	0,4	0,0	0,0	0,0	0,0	-0,1	0,0	0,5
X50	-0,3	0,0	0,0	0,3	0,0	0,1	0,1	0,4	-0,4	0,7	-0,1	0,0	0,0	0,0	0,0	0,0

PCi= Principalkomponent. Tabellen visar procent laddningar av varje variabel i principalkomponentskoefficienterna.

Från tabell 4 kan vi avläsa principalkomponentsladdningar eller det som kallas egenvektor av korrelationsmatrisen (Z' Z). Variabler som har mindre värden än 0,10 antas inte ha laddning i de första fyra principalkomponenterna, vilket betyder att dessa variabler ej heller ger rätt parameterskattningar i regressionsanalysen, men detta betyder inte att dessa variabler är meningslösa. Det framgår tydligt att alla variabler har ungefär samma laddning vid den första principalkomponenten.

Figur 3 visar spridning kring det anpassade koordinatsystemet för de två första principalkomponenterna. En annan tolkning av principalkomponentspoängen är att sambanden mellan principalkomponenter är icke-linjära, om man vill vara mycket noggrann.

Figur 3 Plottade principalkomponentspoänger (Z1 och Z2).



Båda graferna (Figur 2 och Figur 3) skapas med SAS 9.2 vid PROC PRINCOM.

Med hjälp av ellipslinjen visas det 95-procentiga konfidensintervallet kring alla observationer och i bästa fall ska alla observationer ligga inom detta ellipsområde. Observationer utanför ellipsområdet antas som extrema observationer. I detta fall ligger

Förenade arabemiraten, Oman och Jordanien utanför 95-procent ellipsfördelningen. En kombination av figur 1 och figur 3 kan hjälpa till att identifiera dessa extrema observationer mot variabler där de har extremt högt eller lågt värde. En subanalys har visat att dessa extrema observationer har det minsta värdet i X17, eftersom strecken X17 riktar sig mot detta håll. Jag har inte exkluderat dessa observationer från data, eftersom syftet med analysen var att behålla både variabler och observationer.

En undersökning visar att ju högre positivt värde för den första skattade principalkomponentsekvationen (Z_1) desto mindre värde för medellivslängd. Jag jämför detta påstående (tabell 5) med det ursprungliga värdet för medellivslängd (X_{11}) och den beräknade principalkomponentsekvationen (Z_i), eftersom denna variable så småningom ska användas som förklaringsvariabel.

Tabell 5 Beräknat värde för första principalkomponentsekvationen (Z_1)

Afghanistan	$X_{11} = 43,5$	$Z_1 = 5,0869553887$
Sverige	$X_{11} = 78,2$	$Z_1 = -4,318029803$

Ju högre värde för X3, X13, X14, X43, X44, X46, X47 och X50, desto högre negativt värde för Z_1.
Ju högre värde för X2, X4, X12, X15, X16, X42 och X45, desto högre positivt värde för Z_1.

Tabell 5 visar att Afghanistan som har det minsta värdet för X_{11} har det högsta positiva värdet för Z_1, och Sverige som har nästan högsta värdet för X_{11} har det nästan minsta värdet för Z_1.

Regression med principalkomponenter

Syftet med denna metod är att anpassa en så bra modell som möjligt där alla förklaringsvariablerna är med i den slutliga modellen.

Vi börjar med en multipel regression inkluderande alla variabler (tabell 6a) i syfte att visa svårigheten vid tolkning av resultatet, vidare gör vi en multipel regression med de första fyra principalkomponenter för att markera den skattade determinationskoefficienten (R2), och till sist en enkel regression med de skattade principalkomponenterna. Resultatet redovisas i slutet av detta avsnitt.

Multipel regressionsmodell med alla variabler

Vi kommer se tydligt att en regressionsanalys inte tar hänsyn till verkligheten, d.v.s. den kan skatta alla parametrar som man stoppar in i modellen och deras konfidensintervall kan dessutom vara utanför det verkliga intervallet. Om man inte är medveten om relationerna mellan förklaringsvariabler måste man vara mycket försiktig med sin tolkning av parametrar.

Vid icke-signifikanta förklaringsvariabler i modellen väljer man metoder för reduktion. Variationsinflationsfaktor (VIF) är det första förslaget för att välja en reduktion av modellen där anpassningen utgår från diagonalelementen från inversen av korrelationsmatrisen för de oberoende variablerna. När en förklaringsvariabel är nästan linjärt kombinerad med andra förklaringsvariablerna i modellen, blir skattningarna instabila med stora standardfel. Detta problem har vi kallat för kollinjäritet eller multikollinjäritet. Det är en bra idé att ta reda på vilka variabler som är kollinjära med varandra.

Tillvägagångssättet i PROC REG i SAS följer Belsley, Kuh och Welsch (1980). I PROC REG finns flera metoder för att upptäcka kollinjäritet med COLLIN, COLLINOINT, TOL, och VIF.

Tabell 6a Multipel regressionsmodell med alla 16 variabler i modellen (full modell)

	B	Std.Err.	t(127)	p-level	Partial Corr.	Semipart Corr.	Tolerance	R-square	VIF
Intercept	83,22	10,27	8,11	0,00					
X2	0,33	0,16	2,11	**0,04**	0,18	0,04	0,38	0,62	2,60
X3	0,00	0,01	-0,31	0,75	-0,03	-0,01	0,25	0,75	4,01
X4	-0,46	0,41	-1,12	0,26	-0,10	-0,02	0,05	0,95	20,26
X12	-0,26	0,10	-2,50	**0,01**	-0,22	-0,04	0,03	0,97	39,08
X13	-0,01	0,05	-0,15	0,88	-0,01	0,00	0,27	0,73	3,76
X14	0,24	0,11	2,26	**0,03**	0,20	0,04	0,09	0,91	11,33
X15	-0,07	0,02	-4,87	**0,00**	-0,40	-0,08	0,07	0,93	13,88
X16	0,03	0,09	0,38	0,70	0,03	0,01	0,02	0,98	45,83
X17	-1,00	0,13	-7,90	**0,00**	-0,57	-0,13	0,10	0,90	9,74
X42	-0,04	0,04	-0,85	0,40	-0,07	-0,01	0,06	0,94	16,26
X43	-0,05	0,04	-1,19	0,24	-0,10	-0,02	0,10	0,90	10,42
X44	0,00	0,04	-0,07	0,95	-0,01	0,00	0,08	0,92	12,37
X45	0,05	0,07	0,68	0,50	0,06	0,01	0,01	0,99	149,23
X46	0,07	0,07	0,89	0,37	0,08	0,01	0,04	0,96	26,38
X47	0,12	0,07	1,62	0,11	0,14	0,03	0,01	0,99	69,43
X50	-0,03	0,02	-1,77	0,08	-0,16	-0,03	0,22	0,78	4,65

R= ,98213970 R^2= ,96459840 Adjusted R^2= ,96013835

F(16,127)=216,28 p<0,0000 Std.Error of estimate: 2,0752

Det är svårt att bedöma regressionsresultatet i tabell 6a. Det räcker inte att nöja sig med R2 eftersom alla skattade regressionskoefficienterna är svåra att tolka (t.ex X44) i samband med responsvariabeln (X11). Koefficienternas positiva eller negativa påverkan på medellivslängden inte är pålitlig i jämförd med korrelations tabell. För en bättre jämförelse av resultat med principalkomponenter och korrelationskoefficient se appendix-2.

Tabell 6b Multikollinjäritet diagnostisk analys

Number	Eigenvalue	Condition Index
1	9,89529	1,00000
2	1,62161	2,47025
3	1,37232	2.68526
4	0,78413	3,55238
5	0,59482	4,07868
6	0,48153	4,53319
7	0,41155	4,90347
8	0,25466	6.23351
9	0,20562	6.93723
10	0,16679	7.70240
11	0,08860	10.56785
12	0,04511	14.81083
13	0,03351	17.18337
14	0,02682	19.20813
15	0,01348	27.09581
16	0,00415	48.80969

Resultat från SAS med ”COLLINOINT”

Den beräknade konditionsindexen i detta fall blir

$$\delta_k = \left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) = \left(\frac{9{,}89529}{0{,}00415}\right) = \sqrt{2384{,}407229} = 48{,}80969$$

Detta är ganska högt för att man ska kunna påstå att multikollinjäritet föreligger. Vilka innebär att normalt man börjar med en reduceringsprocess av förklaringsvariablerna. Diagnostiska analysmetoder och residual analyser uteblir eftersom modellen har redundant information.

”Ridge regression”

Innan vi gör principalkomponentsregressionen, kontrollerar vi data med ”ridge regression” (R. John 1984; Esteban 1988). En modell med ridge-koefficient δ=0,10 och de valda variablerna som enligt metoden är lämpliga markeras och presenteras i tabell 7. Många kontroller av antaganden och tester måste utföras innan man är klar med sin slutliga modell.

Resultatet från ridge regression med ridge-koefficient δ=0,10 ger inte något annorlunda svar än den vanliga multipelregressionen, utan proceduren markerar att variabel X47 också är en signifikant parameter. Den skattade R2 är 93 procent, lite mindre än tabell 6a. Fortfarande är modellen instabil och modellparametrarna är svåra att tolka.

Tabell 7 Ridge regressions resultat med δ=0,10

	B	SE(B)	t(127)	p-level	Toleran.	R-square	Partial Cor.	Semipart Cor.
Intercept	77,73	5,51	14,11	0,00				
X2	0,33	0,17	2,00	0,05	0,59	0,41	0,17	0,04
X3	*0,00*	0,02	0,19	0,85	0,39	0,61	0,02	0,00
X4	-0,50	0,28	-1,78	0,08	0,18	0,82	-0,16	-0,04
X12	-0,09	0,05	-1,69	0,09	0,16	0,84	-0,15	-0,04
X13	*0,01*	0,05	0,14	0,89	0,49	0,51	0,01	0,00
X14	**0,17**	**0,08**	**2,18**	**0,03**	**0,26**	**0,74**	**0,19**	**0,05**
X15	**-0,06**	**0,01**	**-4,96**	**0,00**	**0,20**	**0,80**	**-0,40**	**-0,11**
X16	-0,08	0,04	-1,95	0,05	0,15	0,85	-0,17	-0,04
X17	**-0,72**	**0,09**	**-7,58**	**0,00**	**0,32**	**0,68**	**-0,56**	**-0,17**
X42	-0,02	0,03	-0,76	0,45	0,25	0,75	-0,07	-0,02
X43	-0,02	0,03	-0,57	0,57	0,36	0,64	-0,05	-0,01
X44	0,03	0,03	0,97	0,33	0,32	0,68	0,09	0,02
X45	-0,02	0,02	-1,11	0,27	0,15	0,85	-0,10	-0,02
X46	0,02	0,03	0,58	0,56	0,31	0,69	0,05	0,01
X47	**0,05**	**0,02**	**2,13**	**0,03**	**0,24**	**0,76**	**0,19**	**0,05**
X50	*0,01*	0,02	0,35	0,73	0,33	0,67	0,03	0,01

R= 0,97 R^2= 0,94 Adjusted R^2= 0,93

F(16,127)=123,00 p<0,0000 Std.Error of estimate: 2,7156

Multipelregressionsmodell med principalkomponenter

För att kontrollera sambandet mellan de fyra första principalkomponenterna, görs en multipelregressionsmodell som visas i tabell 8. Man kan se att de två första och den fjärde principalkomponenten har högst signifikansvärde. Om man önskar att skapa en modell med bara dessa tre variabler finns en modell med bara tre variabler som i sin tur karaktäriseras av de nämnda dimensionerna från principalkomponenterna.

I detta fall ger regressionsmodellen med fyra principalkomponenter approximativt samma resultat som modellen med alla ursprungliga variabler (16 variabler) R2=93 procent. Skillnaden blir i antalet förklaringsvariabler. I detta fall när modellen är optimal och välanpassad undersöker vi de extrema observationerna vilka presenteras i nästa avsnitt. Som redan nämnts är tolkning av regressionskoefficienter mer generell

än enskilda tolkningar. Länder med låg socioekonomiska och demografisk status kommer att ha lägre värde för medellivslängd (intercept=64,41), eftersom deras principalkomponentvärde är lågt.

Tabell 8 Regressionsanalys med fyra principalkomponenter som utgår från korrelationsmatrisen och är valda enligt det kumulativa värdet

	B	Std.Err. - of B	t(139)	p-level
Intercept	64,40556	0,236743	272,0485	0,000000
PC1	-9,67993	0,237569	-40,7457	0,000000
PC2	-1,82979	0,237569	-7,7021	0,000000
PC3	0,29251	0,237569	1,2313	0,220303
PC4	1,74935	0,237569	7,3635	0,000000

R= ,96300969 R^2= ,92738767 Adjusted R^2= ,92529811
F(4,139)=443,82 p<0,0000 Std.Error of estimate: 2,8409

Extrema observationer i data

Vidare söker jag efter observationer som i detta sammanhang kan betraktas som några inflytelserika observationer. De flesta extremvärden antas ha negativa värden för residualerna vilka syns bättre i ett histogram. Resultatet av denna undersökning visas i tabell 9. Två av dessa länder rapporterades som extrema redan vid den första principalkomponentskattningen (se figur 3). INFLUENCE- alternativet i SAS begär att få statistik som Belsley, Kuh och Welsch (1980) föreslog för att mäta påverkan av varje observation på parameterskattningarna. Inflytelserika observationer är de som enligt olika kriterier, tycks ha ett stort inflytande på parameterskattningarna.

Tabell 9 Inflytelserika observationer enligt principalkomponentregression

obs	PC1	X11	Fit	SE Fit	Residual	St Resid
Före. arab	-1,94	73,8	72,13	1,16	1,68	0,65 X
Liberia	4,34	55,4	49,47	0,50	5,93	2,12 R
Oman	2,32	69,6	58,38	1,03	11,22	4,24 RX
Saudi. arab.	0,90	69,7	63,45	0,70	6,23	2,26 R
Soerra Leo	4,70	39,0	45,00	0,67	-6,00	-2,17 R
Sydafrika	-0,99	62,9	68,59	0,41	-5,69	-2,02 R
Zambia	2,68	48,9	57,94	0,41	-9,04	-3,21 R
Zimbawe	2,06	53,7	60,06	0,37	-6,36	-2,26 R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Eftersom dessa extrema observationer kommer från samma geografiska område och världsdel låter vi dem vara kvar i kommande analyser. Förslaget i ett sådant läge är att justera parameterskattningar för en slumpfaktor (variabel som grupperar data till de olika världsdelarna).

Enkel regressionsmodell med den första principalkomponenten

Resultat av enkel linjär regression med den första principalkomponenten ger 86 procents förklarad varians (R2=0,86). Den skattade modellen visas i figur 4. Vidare kan man se att ett optimalt läge för en linjär regression föreligger, där alla punkter koncentreras kring lutningslinjen.

Regressionsmodellen baserad på den första principalkomponenten blir:

Medellivslängd = 64,4056 – 3,04558(PC1)
R2=86 %

Vi kontrollerar detta för två utvalda länder, Afghanistan och Sverige:

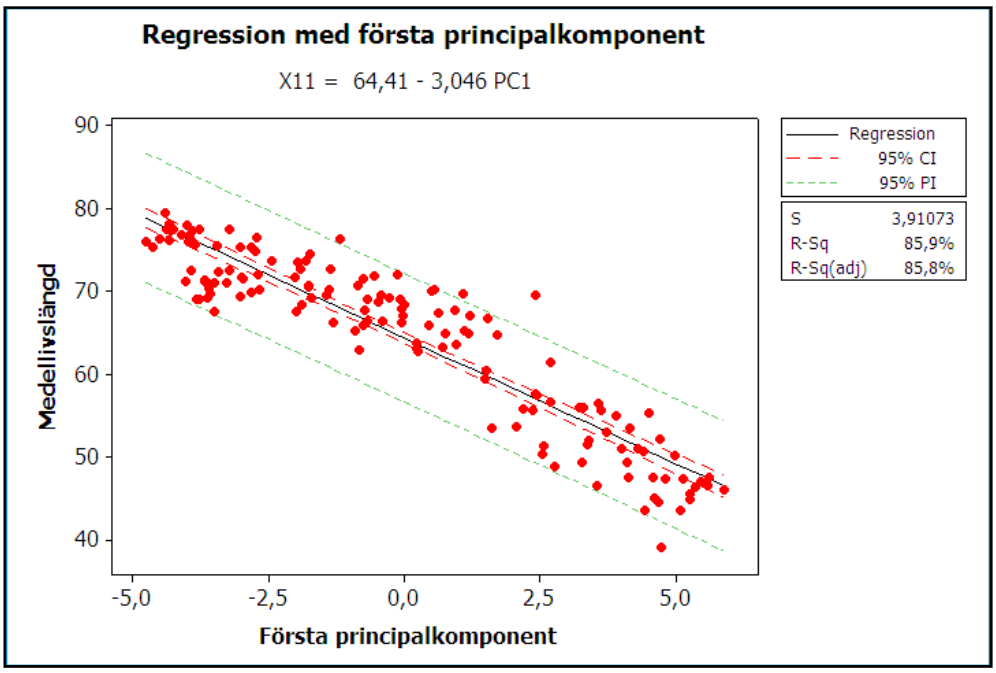
Observerade värden:
Afghanistan $X_{11} = 43,5$
Sverige $X_{11} = 78,2$

Skattade Z1:
Z1(Afghanistan)= 5,0869553887
Z1(Sverige) = -4,318029803

Skattade Y hat från regressionsmodell:
Medellivslängd (X11) = 64,4056 – 3,04558(5,0869553887) = 48,9129
Medellivslängd (X11) = 64,4056 – 3,04558(-4,318029803) = 77,56

Det är inte så illa, med bara en förklaringsfaktor och R^2=86 % jämfört med multipel-regression som gav R2=96 % kan vi nu skatta medellivslängd för vilket land som helst i hela världen. Det enda som behövs är värden på förklaringsvariablerna och en tabell med skattade proportionsegenvärden (tabell 4). Eller om man ska använda detta som ett verktyg kan man räkna om till Z_i-värden.

Figur 4 Enkel linjär regression med den första principalkomponenten



Grafen är skapad med MiniTab, 95 % CI = 95 % konfidensintervall, 95 % PI = 95 % predikted intervall (skattade intervall)
PC1= Första principalkomponent

Mitt förslag att använda principalkomponentsregression syftar inte bara till att lösa multikollinjäritet utan också att kringgå olika test och diagnostiska krav på kontroll av regressionsmodellen. Om syftet är att skatta regressionskoefficienter så är en regressionsmodell baserad på de ursprungliga variablerna en självklar lösning. I mitt fall är en enkel regression med den första principalkomponenten representativ för de variabler som har de högsta laddningarna i den första principalkomponenten. I de fall man undrar hur mycket medellivslängden förändras vid en enhets ökning i en förklaringsvariabel är det alltid själva regressionsekvationen som är baserad på de ursprungliga variablerna, som svarar på denna fråga.

Tabell 10 Sammanställning av determinationskoefficienter från olika regressionsmodeller

	FULL modell	PC1-PC4	PC1, PC2, PC4	PC1	PC1,PC2,PC4,PC7	Ridge modell
R	98,21	92,70	92,70	86,70	94,33	97,00
R2	96,46	92,50	92,50	86,60	94,17	94,00
R2(pred)	93,77	92,04	92,11	86,38	93,72	
PRESS	962,37	1230,02	1218,30	2104,30	970,30	

Modell med PC1, PC2, PC4 och PC7 är vald enligt en stegvis metod. Ridge regression med $\delta = 0{,}10$
FULL modell= Multipel regression

En sammanställning av determinationskoefficienter R2, R2 (pred) och PRESS vissas i tabell 10. Att jämföra determinationskoefficienterna från dessa regressionsmodeller ger inte så mycket information. Multipelregression med samtliga variabler ger 93,8 procents förklarad varians, medan en principalkomponentsregression med bara en förklaringsvariabel gav 86,4 procents förklarad varians. Med fyra principalkomponenter fick vi 93,7 procent och ridge regressionsmetod med δ = 0,10 ger 94 procents förklarad varians.

Om syftet var att öka den skattade andelen förklarad varians (R2), så används en principalkomponentsregression med 4 principalkomponenter som ger 94,2 procents förklarad varians. Med hänsyn till den höga andelen förklarad varians med bara den första principalkomponenten och icke-normalfördelad medellivslängd kan vi nog acceptera resultatet. En kontroll av regression med alla principalkomponenterna ger exakt samma determinationskoefficient, residual fördelning och DIFF-värde som multipelregression.

Anmärkning; R2 och R2(pred) från FULL-modell visar att modellen är ganska instabil (R2=96,5 % R2(pred)=93,7 %) jämfört med principalkomponentsregression. Skattade kvadratsumman minskar från enkel modell (PRESS=2104,3) med bara en komponent till en modell med fyra valda komponenter (PRESS=970,3). PC1, PC2, PC4 och PC7 är valda enligt en stegvis metod (alfa to enter=0,15 och alfa to remove=0,15).

Sammanfattning

Människor i de rika länderna kan förväntas uppnå en genomsnittlig ålder av 85 år, dvs. 30 år mer än för tre generationer sedan. Människor i de fattiga länderna lever i allmänhet till 63 års ålder, men livslängden ökar hela tiden, vilket bidrar till förväntade befolkningsökningar under de kommande 20–30 åren.

Att människor lever längre beror huvudsakligen på att man förebygger och hanterar smittsamma sjukdomar bättre i de fattiga länderna i synnerhet de som drabbar barn. Detta sker genom en kombination av bättre utbildning, särskilt för kvinnorna, tillgång till rent vatten, vaccinationer, bättre kost, bättre bostäder och enkel primärvård. I vissa länder det har förekommit påståenden om mycket hög uppnådd ålder t.ex. i Kina men det finns i alla länder personer som blir 100 år och det är vanligare bland kvinnor än män.

För att förklara skillnaden mellan människors medellivslängd i världens länder har vi gjort analyser i två steg, först med en gemensam faktormodell och därefter med en reducerad modell, där jag valde ut 16 av 61 möjliga variabler som hämtades från olika databaser. Vidare valde jag variabler genom en deskriptivanalys som hade höga korrelationskoefficienter och som hade mindre avstånd från sitt medelvärde, för att kraven för principalkomponent analys skulle vara uppfyllda.

Resultatet av denna analys visade huvudsakligen vilka variabler som kan vara de bästa förklaringsvariablerna om man vill undersöka människornas medellivslängd.
Ett fåtal demografiska variabler, t.ex. spädbarnsdödlighet, dödstal och variabler inom jordbrukssamhället kan förklara den underliggande dimensionen kring människornas medellivslängd i världen. Dessa blandningar av ekonomisk-, och demografiska variabler är möjliga att få för minst 144 länder

Vidare har icke normalfördelade variabler påverkat bedömningar och approximationer, dvs. både utfallsvariabeln och förklaringsvariablerna var ej normalfördelade vid den sista regressionsmodellen.

Jag har inte tolkat koefficienterna från vare sig regressionsanalysen eller principalkomponentsregressionen eftersom detta inte var syftet. En subanalys av data har visat att resultatet av principalkomponentanalysen med utgångspunkt från en korrelationsmatris ger högre informationsvärde till modellerna än med utgångspunkt från kovariansmatrisen.

Om förhållandena i de extremt fattiga länderna förändras till det bättre så kommer de närmare de länder som ligger omkring medelvärdet för medellivslängden. Däremot kan inte de extremt rika länderna minska sina värden eftersom de redan har noll som minsta värde (t.ex. X42 och X45). Detta innebär att man kanske i framtiden kan utgå från kovariansmatrisen och skatta nya modeller.

Resultatet visar att vi kan ha en enkel regressionsmodell (R2 = 86,6 %) med bara en principalkomponent (procent total varians = 62,0 %) som förklaringsvariabel i modellen. Detta är en generell bedömning av alla utvalda variabler (16 variabler) i en och samma modell. Slutsatsen blir att ju högre positivt värde det är för den första skattade principalkomponentsekvationen (Z1) desto mindre värde är det för medellivsläng-

den. Dessa positiva värden karaktäriseras av ”Befolkningsförändring i procent μ=2, σ=1,8”, ”Antal födda barn per kvinna μ=4, σ=1,9”, ”Andel av befolkningen i åldern 0-14 år μ=35,4, σ=10,5”, ”Spädbarnsdödlighetsfrekvens per 1 000 levande födda 1990–95 μ=52,1, σ=42,3”, ”Födelsefrekvens per 1 000 människor μ=29,9, σ=13,5”, ”Inkomstkällor i procent inom jordbruk μ=21,2, σ=15,7” och ”Sysselsatta inom jordbruk i procent av befolkningen μ=38,8, σ=29,2”.

Metoden är användbar inom medicinsk forskning, epidemiologi, sociometrik, psykometrik, ekonometri och industri statistik. Samt vid undersökningsmetodik t ex. vid enkätundersökningar där man helst vill välja de frågor (items) som tillsammans kan bygga en linjär kombination oavsett utfallsvariabler, och principalkomponentsanalys är mycket mer effektiv än Item respons teorier. Även i de fall där man försöker hitta underliggande dimensioner i data som innehåller många förklaringsvariabler kan metoden vara mycket användbar.

Metoden är ett bra verktyg att använda vid multivariata data, där man kan gruppera data eller variablerna till olika underliggande dimensioner. Tolkningen kan bli lite svårare, men att samlad information som grupperas till dimensioner som är karaktäristisk för bakgrunden till data är mer rimligt än att tolka masseffekterna av multipla förklaringsvariabler från regressionsmodeller. Till sist vill jag markera att dessa kombinationer av två olika analysmetoder kunde hjälpa till att på ett bra och korrekt sätt identifiera extrema observationer.

Att använda principalkomponentsanalyser i förväg kan vara bra och tidssparande för att kringgå multikollinjäritetsproblem och andra nämnda typer av problematik. Jag har demonstrerat att användning av principalkomponenter för att nå en bra skattning av en bra modell inte är så långt ifrån direkta skattningar med regressionsanalys.

Slutligen, att kombinera principalkomponent till regressionsanalys är effektivt och inte så komplicerat som tidigare. Med hjälp av dagens datorer och mjukvara kan man utföra analyser med stora dataset på ett korrekt sätt och på kortare tid.

Erkännande: En stor tack till professor Sven-Erik Johansson som granskade hela jobbet från ett statistiskt perspektiv, samtidigt som han uppmuntrade mig att bli klar med detta examensarbete. Tack även till min kollega May Blom, leg. Sjuksköterska, med.doktor, som hjälpte mig med språkgranskning.

Farhad H. Alinaghizadeh M., 2009, Stockholm
Farhad.Alinaghizadeh@ki.se

Referenser

1- Afifi, A. A., Virginia C. (1990), "*Computer-Aided Multivariate Analysis*", Second Edition, New York, Chapman & Hall
2- Akaike, H. (1974), "*A New Look at the Statistical Identification Model*", IEEE Transactions on Automatic Control, 19, 716 -723.
3- Akaike, H. (1987), "*Factor Analysis and AIC*", Psychometrika 52, 317 -332.
4- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), "*Regression Diagnostics*", New York: John Wiley & Sons, Inc.
5- Berk, K.N (1977), "*Tolerance and condition in regression computation*", Journal of the American statistical association 72, 863-866
6- Bryan, F. J. Manly (1994), "*Multivariate Statistical Methods*", Second Edition, New York, Chapman & Hall
7- Carroll, J. D. (1972), "*Individual Differences and Multidimensional Scaling*", in R. N. Shepard, A. K. Romney, and S. B. Nerlove, eds., "*Multidimensional Scaling: Theory and Applications in the Behavioral Sciences" (Volume 1)*, New York: Seminar Press.
8- Cattell, R. B. 1996, "*The scree test for the number of factors*". Multivariate Behavioral Research 1: 245-276
9- Cooley, W.W. and Lohnes, P.R. (1971), "*Multivariate Data Analysis*", New York, John Wiley & Sons, Inc.
10- de Jong, S. (1993), "*SIMPLES: An alternative approach to partial least squares regression,*" Chemometrics and Inteligent Laboratory System, 18, 252-263
11- Esteban, W, Jeffrey B. B (1988) "*Influence Measures in Ridge Regression*", JSTOR: Technometrics, Vol. 30, No. 2, pp. 221-227
12- Gabriel, K. R. (1981), "*Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis*," in V. Barnett, ed., "*Interpreting Multivariate Data*", London: John Wiley & Sons.
13- Gabriel, K. R. (1971), "*The biplot graphic display of matrics with application to principal component analysis*". Biometrika 58:453-467.
14- Geladi and Kowalski (1986), "*Partial least-squares regression: A tutorial,*" Analytical Chimica Acta, 185, 1-17
15- Gnanadesikan, R. (1977), "*Methods for Statistical Data Analysis of Multivariate Observations*", New York, John Wiley & Sons, Inc.
16- Hjorth, Urban (1989), "Statistisk slutledning", Studentlitteratur.
17- Hoerl A. E. and Kennard R. W., "*Ridge Regression: Biased Estimation to Nonorthogonal Problems*", Technometrics, 12 (1970), 56–67.
18- Hotelling, H. (1933), *"Analysis of a Complex of Statistical Variables into Principal Components"* Journal of Educational Psychology, 24, 417 -441, 498 -520.
19- Jackson, J.E. (1991), "*A user's guide to principal components*", New York, John Wiley & Sons, Inc.
20- John, R.C. St. (1984). "*Experiments With Mixtures in Conditioning and Ridge Regression*" Journal of Quality Technology 16, pp.81-96.
21- Kshirsagar, A.M. (1972), "*Multivariate Analysis*", New York, Marcel Dekker, Inc.
22- Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), "*Multivariate Analysis*", London: Academic Press.
23- Morrison, D.F. (1976), "*Multivariate Statistical Methods*", Second Edition, New York, McGraw-Hill Book Co.
24- Pearson, K. (1901), *"On Lines and Planes of Closest Fit to Systems of Points in Space"* , Philosophical Magazine, 6(2), 559 -572.
25- Rao, C.R. (1964), *"The Use and Interpretation of Principal Component Analysis in Applied Research"*, Sankhya A, 26, 329 -358.
26- Rawlings J. O. (1988), "*Applied Regression Analysis: A research tool*", Wadsworth & Brooks/Cole
27- Stephens, M. A. (1974). EDF "*Statistics for Goodness of Fit and Some Comparisons*", Journal of the American Statistical Association, Vol. 69, pp. 730-737.
28- Stone, M. and Brooks, R. J (1990) "*Contium regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion)*". J. R Statist. Soc. B, 52, 237-269; corrigendum, 54 (1992), 906-907.
29- Younger M. S., "*A Handbook for Linear Regression*", USA: DUXBURY Press, 1979.

Appendix-1 Sammanställning av regression, korrelation, tolerans, VIF och R2 från enkla regressionsmodeller

	B	Std.Err. B	t(127)	p-level	Partial Corr.	Semipart Corr.	Tolerance %	R-square	VIF	PC1	PC2	PC3	PC4
Intercept	83,22	10,27	8,11	0,00	Cor.	Cor.							
X2	0,33	0,16	2,11	**0,04**	0,18	0,04	0,38	0,62	2,60	0,17	-0,47	0,01	-0,30
X3	0,00	0,01	-0,31	0,75	-0,03	-0,01	0,25	0,75	4,01	-0,26	-0,16	0,15	-0,24
X4	-0,46	0,41	-1,12	0,26	-0,10	-0,02	0,05	0,95	20,26	0,30	-0,15	0,02	-0,15
X12	-0,26	0,10	-2,50	**0,01**	-0,22	-0,04	0,03	0,97	39,08	0,28	-0,31	0,01	0,15
X13	-0,01	0,05	-0,15	0,88	-0,01	0,00	0,27	0,73	3,76	-0,23	0,16	-0,13	-0,07
X14	0,24	0,11	2,26	**0,03**	0,20	0,04	0,09	0,91	11,33	-0,24	0,38	0,11	-0,21
X15	-0,07	0,02	-4,87	**0,00**	-0,40	-0,08	0,07	0,93	13,88	0,30	0,11	0,00	-0,20
X16	0,03	0,09	0,38	0,70	0,03	0,01	0,02	0,98	45,83	0,30	-0,17	0,03	-0,07
X17	-1,00	0,13	-7,90	**0,00**	-0,57	-0,13	0,10	0,90	9,74	0,20	0,45	0,10	-0,53
X42	-0,04	0,04	-0,85	0,40	-0,07	-0,01	0,06	0,94	16,26	0,25	0,27	-0,16	0,40
X43	-0,05	0,04	-1,19	0,24	-0,10	-0,02	0,10	0,90	10,42	-0,16	-0,22	-0,59	-0,37
X44	0,00	0,04	-0,07	0,95	-0,01	0,00	0,08	0,92	12,37	-0,15	-0,12	0,71	-0,09
X45	0,05	0,07	0,68	0,50	0,06	0,01	0,01	0,99	149,23	0,29	0,17	0,01	0,01
X46	0,07	0,07	0,89	0,37	0,08	0,01	0,04	0,96	26,38	-0,27	0,00	-0,21	-0,15
X47	0,12	0,07	1,62	0,11	0,14	0,03	0,01	0,99	69,43	-0,26	-0,24	0,12	0,05
X50	-0,03	0,02	-1,77	0,08	-0,16	-0,03	0,22	0,78	4,65	-0,28	-0,01	-0,03	0,32

Kommentar: Här kan man se hur ostabil multipelregression är när det gäller skattade parametrar. Där för har jag valt att lägga dem bredvid de första fyra komponenterna för att jämföra deras tecken(±). Om man vill göra en generell bedömning av förklaringsvariabler i en sådan situation där alla förklaringsvariabler är korrelerade med varandra, då är principalkomponenter bäst för en sådan bedömning. De gulmarkerade komponenterna antas ha en negativ påverkan och grönmarkerade komponenter har en positiv påverkan. Vi har redan räknat ut det att ju högre positivt värde desto mindre medellivslängd om vi ställer upp första komponenten mot responsvariabeln X11 (se figur 4.1).

Appendix-2 Sammanställning av multipelregression, ridge regression, första 4 principalkomponenterna

	Multipelregression			Ridge regression δ=0,10			Principalkomponent 1-4				Deskriptiv analys		
	BETA	Std.Err. BETA	p-level	BETA	Std.Err. BETA	p-level	PC1	PC2	PC3	PC4	Means	Std.Dev.	X11 Corr.
Intercept	83,22	10,27	0,00	77,73	5,51	0,00							
X2	**0,33**	**0,16**	**0,04**	**0,33**	**0,17**	**0,05**	0,17	-0,47	0,01	-0,30	2,04	1,77	-0,39
X3	0,00	0,01	0,75	0,00	0,02	0,85	-0,26	-0,16	0,15	-0,24	50,19	23,22	0,73
X4	-0,46	0,41	0,26	-0,50	0,28	0,08	0,30	-0,15	0,02	-0,15	3,98	1,91	-0,87
X12	-0,26	0,10	**0,01**	-0,09	0,05	0,09	0,28	-0,31	0,01	0,15	35,40	10,50	-0,76
X13	-0,01	0,05	0,88	0,01	0,05	0,89	-0,23	0,16	-0,13	-0,07	55,59	6,97	0,60
X14	**0,24**	**0,11**	**0,03**	**0,17**	**0,08**	**0,03**	-0,24	0,38	0,11	-0,21	8,80	5,52	0,63
X15	**-0,07**	**0,02**	**0,00**	**-0,06**	**0,01**	**0,00**	0,30	0,11	0,00	-0,20	52,06	42,33	-0,96
X16	0,03	0,09	0,70	-0,08	0,04	0,05	0,30	-0,17	0,03	-0,07	29,90	13,48	-0,88
X17	**-1,00**	**0,13**	**0,00**	**-0,72**	**0,09**	**0,00**	0,20	0,45	0,10	-0,53	10,37	4,26	-0,78
X42	-0,04	0,04	0,40	-0,02	0,03	0,45	0,25	0,27	-0,16	0,40	21,16	15,70	-0,72
X43	-0,05	0,04	0,24	-0,02	0,03	0,57	-0,16	-0,22	-0,59	-0,37	31,85	12,48	0,44
X44	0,00	0,04	0,95	0,03	0,03	0,33	-0,15	-0,12	0,71	-0,09	46,79	14,21	0,46
X45	0,05	0,07	0,50	-0,02	0,02	0,27	0,29	0,17	0,01	0,01	38,75	29,20	-0,88
X46	0,07	0,07	0,37	0,02	0,03	0,56	-0,27	0,00	-0,21	-0,15	20,55	12,12	0,77
X47	**0,12**	**0,07**	**0,11**	**0,05**	**0,02**	**0,03**	-0,26	-0,24	0,12	0,05	40,53	20,32	0,80
X50	-0,03	0,02	0,08	0,01	0,02	0,73	-0,28	-0,01	-0,03	0,32	74,67	24,60	0,83

Kommentar: Sammanställning av resultat från de tre metoderna. Värdena för PC1 är grupperad enligt deras positiva och negativa tecken. Vi ser här att ridge regression har lyckats att minimera SE av regressionskoefficienterna men resultatet är fortfarande ostabilt och tolkningen av koefficienterna är svår.

Anmärkning: Lägg märke till att regressionskoefficienterna från både multipelregression och ridge regression som är lika med 0.

Appendix-3 Alla länder som är med i databasen (144)

Afghanistan	Gambia	Madagaskar	Slovenien
Albanien	Georgien	Makedonien	Somalia
Algeriet	Ghana	Malawi	Spanien
Angola	Grekland	Malaysia	Sri Lanka
Argentina	Guatemala	Mali	Storbritannien
Australien	Guinea	Marocko	Sudan
Azerbajdzjan	Guinea-Bissau	Mauretanien	Surinam
Bangladesh	Guyana	Mauritius	Swaziland
Belgien	Haiti	Mexiko	Sverige
Benin	Honduras	Mocambique	Sydafrika
Bhutan	Indien	Moldavien	Sydkorea
Bolivia	Indonesien	Mongoliet	Syrien
Bosnien & Hercego	Irak	Myanmar (Burma)	Tadzjikistan
Botswana	Iran	Nederländerna	Tanzania
Brasilien	Irland	Nepal	Tchad
Bulgarien	Israel	Nicaragua	Thailand
Burkina Faso	Italien	Niger	Tjeckien
Burundi	Jamaica	Nigeria	Togo
Centralafr. rep.	Japan	Norge	Trinidad & Tobago
Chile	Jordanien	Nya Zeeland	Tunisien
Colombia	Kamerun	Oman	Turkiet
Costa Rica	Kanada	Pakistan	Turkmenistan
Danmark	Kazakstan	Panama	Tyskland
Dominikanska rep	Kenya	Papua Nya Guinea	Uganda
Ecuador	Kina	Paraguay	Ukraina
Egypten	Kirgistan	Peru	Ungern
El Salvador	Kongo	Polen	Uruguay
Elfenbenskusten	Kroatien	Portugal	USA
Estland	Kuba	Rumänien	Uzbekistan
Etiopien	Kuwait	Rwanda	Venezuela
Fiji	Laos	Ryssland	Vietnam
Filippinerna	Lettland (Latvia)	Saudiarabien	Vitryssland
Finland	Libanon	Schweiz	Zaire
Frankrike	Liberia	Senegal	Zambia
Förenade arabemir.	Libyen	Sierra Leone	Zimbabwe
Gabon	Litauen	Slovakien	Österrike