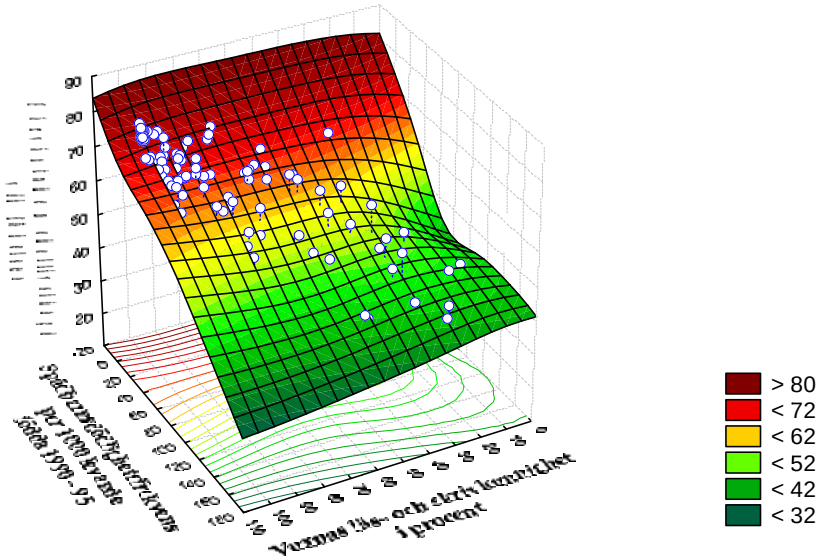
EKONOMIHÖGSKOLAN
Lunds universitet
Statistiska Institutionen

Principalkomponentsregression



2010-05-10

Magisterexamen 15 hp

91-120 nivå

Farhad H. Alinaghizadeh

Handledare: Mats Hagnell

Innehåll

Inledning

”Principal Components Regression” (PCREG) som en metod från ” biased regression metoder” kan används för att modellera data med högt korrelerade förklaringsvariabler. Metoderna redan diskuteras av ett antal författare *(Marquardt 1970, Hoerl and Kennard 1970, Lott 1973, Hawkins 1973, Webster, Graces and Mason 1974, Hoerl, Kennard and Baldwin 1975, Marquardt and Snee 1975, Hocking, Speed and Lynn 1976, Smith and Campbell 1980, Park 1981).*

Under de senaste årtionden har statistiker och forskare använt sig av multipla regressionsmodeller som är relaterade till ”Partial Least Squares regression” (PLSREG) och SIMPLS- algoritm för att undvika multikollinjäritet i data. Dessa analysmetoder används för att skatta värde av en responsvariabel från ett stort antal högt korrelerade förklaringsvariabler. PCREG använder de skattade komponenterna på samma sätt som (PLSREG) vid skattning av regressionskoefficienter, dvs som en linjär kombination av de ursprungliga variablerna. Dessa metoder diskuteras i detalj av (*Wold 1966, Horel and Kennard 1970, Dijkstra 1983, 1985, de Jong 1993 SIMPLS algoritm, Rännar et al. 1994, van der Voet 1994*) och har under de senaste år utvecklats i ett antal mjukvaror (t.ex. SAS, MiniTab, Statistica och SPSS där SPSS kräver ett särskilt tillägg) och är fortfarande under utveckling. PLSREG- metoderna använder principalkomponenterna i ordning från den 1-a med största variationsförklaring och 2-a med näst största osv. Syftet med denna uppsats är att visa hur man kan välja dessa komponenter beroende på deras vikt och betydelse och inte automatiskt i ordning efter egenvärde.

En allmän teori av PCREG i jämförd med PLSREG, SIMPLS-algortim och ”ridge” regression (RREG) diskuteras i denna uppsats. Metoderna illustreras med två olika dataset i syftet att bygga en regressionsmodell (PCREG) med ett antal utvalda skattade komponenter som ger mer relevant tolkning av data och dessutom låga medelkvadratsfel. Därefter skattar vi regressionskoefficienterna, standardfelen av koefficienterna och t-värdena i PCREG med hjälp av PROC IML i SAS 9.2.

Tolkning av regressionskoefficienter i PCREG sker exakt på samma sätt som i sedvanliga regressionsanalys. Vår numeriska resultat tyder på att teorin även kan användas för att förbättra PLSREG skattningar som bygger på roterade principalkomponenter. Även om detta arbete har motiverats av problemet med analys av kvantitativa variabler inom socioekonomisk och demografisk statistik, är metoden också tillämpbar på problem vid skattning av regressionskoefficienter baserade på högt korrelerade kvantitativa förklaringsvariabler. Metoden är relevant för ett brett spektrum av problem inom de fysiska, kemiska, medicinska, ekonomiska och tekniska vetenskaperna, vilket också kommer att diskuteras i denna uppsats.

Nyckelord: Multiple Linear Regression (MLREG), Principalkomponentsregression (PCREG), Partial Least Squares Regression (PLSREG), SIMPLS- algorithms, Reduced Rank Regression (RRR), Ridge regression (RREG), Ridge coefficient (δ), "Predicttive Error Sum of Squares" (PRESS), Variation Inflation Factor (VIF).

Abstract

"Principal Components Regression" (PCREG), as one of methods in "biased regression methods" can be used to model data with highly correlated explanatory variables. This well-known method has been discussed by many authors (*Marquardt 1970, Hoerl and Kennard 1970, Lott 1973, Hawkins 1973, Webster, Grace and Mason 1974, Hoerl, Kennard and Baldwin 1975, Marquardt and Snee 1975, Hocking, Speed and Lynn 1976 Smith and Campbell 1980, Park 1981*).

In recent decades, when modeling data with highly correlated explanatory variables, statisticians and researchers have frequently used multiple regression models that are related to the "Partial Least Squares regression (PLSREG) and SIMPLE algorithm to avoid multicollinearity in the data. PCREG uses the estimated components in the same way as PLSREG to estimate the regression coefficients. Both of these biased regression methods are discussed in detail by *Wold 1966, Horel and Kennard 1970, Dijkstra 1983, 1985, de Jong 1993 SIMPLE algorithm, Rännar et al. 1994, van der Voet 1994.* Most available software; however, only gives users the opportunity to run PLSREG analyses. Examples include SAS, Minitab, Statistica and SPSS (although SPSS require that a special component be downloaded and installed to run PLSREG analyses). All these programs are under development and each provides slightly different output.

All programs that run PLSREG analyses automatically rank principal components in descending order of eigenvalues. One purpose of this essay is to clarify why this descending order may not always be the most useful order in all kinds of research. A second purpose is to show that it is possible to use a newly developed program that runs PCREG in SAS to choose components on the basis of their weight and importance in relation to the explanatory variables. Additionally, the general theory of PCREG, PLREG and “ridge” regression (RREG) will be discussed.

Two data sets (one large and one small) will be used to compare the results of analyses using PCREG, PLSREG, and RREG. Two data sets were necessary because the more variables in a dataset, the more difficult it is to choose how many estimated components must be used in the analysis and the more potentially misleading the automatic eigenvalues choice of the existing software programs can be. The PCREG analyses of the two data sets (run using PROC IML in SAS 9.2) will include an estimation of the regression coefficients, standard errors of the coefficients, and t-test and (p) values.

Interpretation of regression coefficients in PCREG is done in exactly the same way as in regression analysis. The numerical results of the PCREG analyses suggest that the PCREG algorithm used to write the new program can also be used to improve PLSREG estimates based on rotated principal components.

Although this work has been motivated by the problem of analyzing quantitative variables in socio-economic and demographic data sets, the method is also applicable to problems that involve estimating regression coefficients that are based on highly correlated quantitative explanatory variables. Thus, the method is probably relevant to a wide range of problems in the physical, chemical, medical, financial, and technical sciences.

Keywords: Multiple Linear Regression (MLREG), Principal components regression (PCREG), Partial Least Squares Regression (PLSREG), SIMPLS-algorithms, the Reduced Rank Regression (RRR), Ridge regression (RREG), Ridge coefficient (δ), "Predicttive Error Sum of Squares" (PRESS),, Variation Inflation Factor (VIF).

Bakgrund

Att använda skattade principalkomponenter (PC) som förklaringsvariabler i en multipel linjär regression (MLREG) har både fördelar och nackdelar. I de flesta fall, när man tittar på förklaringen till en responsvariabel är det acceptabelt att använda viktigaste komponenten i en regressionsmodell som kännetecknar ett antal variabler, men i viss mån, är önskan eller målet med en regressionsmodell att skatta parametrar som är direkt kopplade till de ursprungliga förklaringsvariablerna. Problemet med vanlig MLREG som använder de beräknade komponenterna som förklaringsvariabler i en regressionsmodell är tolkningen då man drar slutsatser från en regressionsparameter, där flera förklaringsvariabler ingår samtidigt. Vi kan underlätta tolkningen av regressionsparametrar med hjälp av PCREG. Genom att standardisera data och använda dem för att skatta regressionsparametrar som i sin tur påverkar kovariansmatrisen vilket i sin tur ger mindre skattade medelfel. Det andra skälet till att använda (PCREG) är att standardisering av de ursprungliga variablerna undviker multikollinjäritetsproblemet i data.

Kravet på PCREG i denna uppsats är att de slutliga modellerna ska vara jämförbara med den sedvanlig MLREG och koefficienterna ska vara tolkbara på samma sätt. Denna uppsats försöker samtidigt påvisa hur man med hjälp av metoder som PLSREG, PCREG eller RREG kan undvika problemet med redundant information eller multikollinjäritet vid MLREG metoder som under de senaste år har utvecklats med mer fokus på PLSREG. Men när antalet förklaringsvariabler och observationer är stort och korrelationsmatrisen speglas av höga egenvärden (positiv och negativ) är det svårt att välja antal roterade komponenter eller ridge- koefficientvärde. Dessa kritiska moment diskuteras i denna uppsats och några tillgängliga metoder diskuteras. Samtidigt vet vi att en multipel regressions modell (OLS) med många variabler skattar alla koefficienterna ändå och rapporterar en hyfsad determinationskoefficient (R2). Men huvudfrågan är kan man på ett enkelt sätt tolka dessa skattade parametrar? Och den andra frågan är hur vid PLSREG eller PCREG ska välja antal komponenter: beroende på deras variationsförklaring eller egenvärde? Vid PLSREG har vi inte möjlighet att själv välja de vettiga komponenterna, detta görs automatisk, och huvudmålet med denna uppsats är att påvisa hur kan vi själv styra analysen enligt vissa kriterier och andra praktiska skäl.

De ovan nämnda metoderna jämförs med PCREG där förklaringsvariabler blir tolkbara enligt Rawlings 2002. Flera biased regressionsmetoder har föreslagits som lösningar på kollinjäritets problem. Dessa metoder diskuteras av (*Stein, 1960, d.s.k Stein krympning*), ”ridge” regression (*Hoerl och Kennard, 1970*) och PCREG och andra varianter av modell skattningar av (*Lott, 1973, Hawkins, 1973; Hocking, Speed och Lynn, 1976; Marquardt, 1970; Webster, Gunst och Mason, 1974*). ”Ridge” regression har fått störst acceptans jämfört med andra metoder, men alla dessa metoder används med uppenbar framgång för olika problem. *”Biased” regressionsmetoder är inte allmänt accepterade och bör användas med försiktighet.* Det är medelkvadratfelet som är grundmotiveringen till att använda ”biased” regressionsmetoder. Sådana metoder kan ge en bättre parameterskattning i den mening som avses för skattning av medelkvadratfelet. Detta innebär inte nödvändigtvis att en ”biased” regressionsmodell är godtagbar eller rent av ”bättre” än en minsta-kvadratlösning (OLS) för andra ändamål än beräkning av modellparametrar (Rawlings 2002).

Även om kollinjäritet inte påverkar precisionen i den skattade responsen (och prognoser) i observerade punkter i X-rymden vid multipel regression, kan den orsaka variansinflation [VIF] av den skattade responsen på olika nivåer av förklaringsvariabler. Park (1981) visar att beräkning av medelkvadratfel (SEM) för skattning av en respons över vissa förklaringsvariabler i X-utrymmet och restriktioner för parameterskattningar i PCREG är optimal. Detta tyder på att ”biased” regressionsmetoder också kan vara till nytta i vissa fall för skattning av respons. Men försiktighet måste iakttas vid användning av kollinjära data för parameter skattningar av respons på andra ställen än i de observerade punkterna.

Data

I detta dokument använder vi två olika dataset för att illustrera dessa metoder. Det första datasetet omfattar både större antal observationer och antal förklaringsvariabler än det andra datasetet. När antalet observationer och antalet variabler är få ger alla metoder (PCREG, PLREG, SIMPLS) nästan samma svar, men när stora dataset används blir det lite svårare att välja den slutliga modellen.

Det första datasetet som används är från ”Philips Geographical Digest” och ”World Resources” årsböcker 96-97. Av 61 möjliga variabler har 16 valts att ingå som förklaringsvariabler i en och samma modell (Alinaghizadeh 2009). Förklaringsvariablerna är högt korrelerade med varandra, utan bortfall. Dessutom är de högt korrelerade med responsvariabeln. Syftet med detta dataset är en uppföljning av en studie där vi försökte anpassa en regressionsmodell för förväntad medellivslängd för världens länder (μ=64,4 , σ=10,4 , N=144), utan att ta bort någon förklaringsvariabel och samtidigt använde ett antal valda komponenter som regressionsparametrar.

Det andra datasetet demonstrerar ett "biased" linjär regressionsproblem med ekonomiska data (Longley, 1967) med 6 förklaringsvariabler utan bortfall och är högt korrelerad. Med variabeln Employment som responsvariabel (μ=65317, σ =3511,97, N=16) i en studie som visar hur ett regressionsprogram fungerar med uppgifter som är kända att vara svår hanterbara och beräkningsintensiva. Analysresultat och beskrivande analyser från detta dataset redovisas i appendix.

Metoder

Vi börjar med en beskrivning av en ”biased” regression metod med (PCREG) och en sammanfattning av de övriga metoderna (PLS, NLPLS, SIMPLE). Att använda en ”biased” regressionsmetod är redan välkänt i litteraturen, men denna uppsats är fokuserad på PCREG. För mer information om de övriga metoderna hänvisas till angivna referenser.

Sedan principalkomponentsanalys (PCA) blivit känd har PCREG diskuterats av många och så småningom dyker metoden upp i litteraturen (*Kendall 1957, Hotelling 1957, Jeffers 1967, Mosteller and Tukey 1977, Mardia et al. 1979, Gunst and Mason 1980*). Misstolkningar av principalkomponenter i samband med PCREG att ”*utelämnade principalkomponenter minskar variationsförklaringen*” diskuteras av Jolliffe (1982) som argumenterar att alla skattade komponenter har sin egen betydelse och detta beror på den bakomliggande informationen som ligger i data. I sedvanliga metoder vid PCREG väljer man automatiskt ett antal komponenter i ordning från den första med högsta variationsförklaring och nedåt. I denna studie demonstreras hur man kan välja vilka komponenter som ska ingå i PCREG- modellen (Alinaghizadeh 2008) med tonvikt på betydelsen av skattade komponenter.

Multikollinjäritet

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 påverks av detta. Redundant information betyder att en X variabel förklarar Y på exakt samma sätt som en annan X variabel förklarar. I detta fall kan två eller flera redundanta förklaringsvariabler bli icke-relevanta 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 högre multikollinjäritet, desto färre tolkningsbara parametrar. Kort sagt, multikollinjäritet påverkar alla skattningar vid regressionsanalys, minsta kvadratskattning av regressionskoefficienter, medelkvadratfel (MSE), t-test, kvadratsummor och till sist determinationskoefficienten R2.

Det finns många metoder för att hantera multikollinjäritet t.ex ridge regression, principalkomponentsanalys, PLREG (Geladi 1986) och continuum regression (Stone 1990, Belsley 1980). Som nämnts har koefficientskattningar i multipel linjär regression kravet 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 minsta-kvadratskattningen, blir ekvationen $\hat{\beta} = (X'X)^{-1}X'Y$ mycket känslig för slumpmässiga fel i den observerade responsvariabeln (Y), vilka ger en stor varians. Multikollinjäritet kan uppstå, till exempel när data samlas in utan en experimentell design.

Regression

Givet en uppsättning av förklaringsvariabler **X1, X2, ...,Xp** och en beroendevariabel eller responsvariabel **Y** finner multipel linjär regressionsmetoden den bästa minstakvadrat linjär prediktionen av responsvariabel baserad på alla förklaringsvariabler. Om **X** är en (**n x p**) matris av **n** observationer på förklaringsvariablerna och **Y** är en (**n x 1**) vektor av observationer på responsvariabeln blir multipel regressionsmodellen:

$$Y = XB \quad (1)$$

där B är en (p x 1) vektor av regressions koefficienter.

Minstakvadrat skattning för **B** blir:

$$B=(X'X)^{-1}X'Y \quad (2)$$

och "lack-of-fit" av denna skattning d.v.s. medelkvadratfel skattning blir:

$$\left(\frac{Y'Y-B'X'Y}{n-p-1}\right)^{1/2} \quad (3)$$

Bland många problem relaterade till multipel regression är multikollinjäritet svåraste situation där förklaringsvariabler är högt korrelerade med varandra. Detta gör att X'X inte har någon invers även om dessa variabler inte är kollinjära men är de fortfarande korrelerade vilka i sin tur skapar dessa problem:

1- Att skatta en stabil invers för X'X blir svårt.
2- Ju mer korrelerade förklaringsvariabler är desto större blir medelkvadratfelet vid skattning av regressionskoefficienterna blir vilket gör att regressionskoefficienterna blir mer korrelerade. Detta gör det svårt att tolka dessa koefficienter. Det "sanna" sambandet mellan förklaringsvariabler och responsvariabel blir mer beroende av relationerna mellan förklaringsvariablerna sinsemellan.

Många metoder har förslagits som lösning, till exempel, stegvis regression eller ett subset av variabler, men dessa metoder tas inte upp i denna uppsats. Vi fokuserar på en multipel regression med alla förklaringsvariabler samtidigt i modellen.

Principalkomponentsregression

”Biased” skattning av en regressionsekvation har under de senaste åren diskuterats inom olika vetenskapliga områden. En sådan metod är regression med principalkomponenter. PCREG-analys är särskilt användbar för att förstå sambanden mellan X-variablerna och för att tolka regressionsparametrarna. Kunskap om variabler kan hjälpa till att avgöra vilka som ska tas med och hur viktiga de är (Rawlings 2002).

För att tillämpa PCREG, börjar vi med att beräkna singulära värden för en matris med centrerad och skalade oberoende variabler (standardiserade variabler), dvs. egenvektorer av Z.

$$Z=UL^{1/2}V' \quad (4)$$

$U_{(n\,x\,p)}$ och $V_{(p\,x\,p)}$

$L^{1/2}$ är diagonalmatris med singulära värden eller roten ur egenvärdena.

Principalkomponenten ges av

$$W=ZV \text{ eller } W=UL^{1/2} \quad (5)$$

Varje kolumn i W ger värdena för *n* observationer i en av principalkomponenterna. Kvadratsumman och multiplikations matris of principalkomponenterna W är en diagonalmatris av egenvärden

$$W'W=(UL^{1/2})'(UL^{1/2})=L^{1/2}U'UL^{1/2}=L \quad (6)$$

där $L = Diag(\lambda_1^2, \lambda_2^2, \cdots, \lambda_p^2)$ samt principalkomponenterna är ortogonala mot varandra, eftersom produkten av summor är lika med o och kvadratsummor av varje principalkomponent är lika med motsvarande egenvärde λ_j^2.

Principalkomponenten som motsvarar det minsta egenvärdet är dimensionen av Z som har minsta spridning. Dessa dimensioner av Z med liten spridning är orsaken till kollinjäritet. Den linjära modellen ges av

$$Y = 1\beta_0 + Z\beta + \varepsilon \quad (7)$$

vilka kan skrivas i form av principalkomponent W som

$$Y = 1\beta_0 + W\gamma + \varepsilon \quad (8)$$

Dessa använder $VV' = I$ för att transformera $Z\beta$ till $W\gamma$.

$$Z\beta = ZVV'\beta = W\gamma \quad (9)$$

Lägg märke till att $\gamma = V'\beta$ är vektor av regressionskoefficienter för principalkomponenter och β är en vektor av regressionskoefficienten för alla Z. Transformation av γ tillbaka till β fås genom:

$$\beta = V\gamma \quad (10)$$

Den sedvanliga minstakvadrat skattningen fås genom att använda principalkomponenter som oberoende variabler:

$$\hat{\gamma} = (W'W)^{-1}W'Y = L^{-1}W'Y \quad (11)$$

$$= \begin{bmatrix} \left(\sum_i W_{i1}Y_i\right)/\lambda_1^2 \\ \left(\sum_i W_{i2}Y_i\right)/\lambda_2^2 \\ \vdots \\ \left(\sum_i W_{ip}Y_i\right)/\lambda_p^2 \end{bmatrix} = \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \vdots \\ \hat{\gamma}_p \end{pmatrix}$$

Regressionskoefficienter för principalkomponenter kan beräknas individuellt för att principalkomponenterna är ortogonala och att W´W är en diagonal matris.

Liksom varians-kovarians matris av $\hat{\gamma}$ är en diagonal matris

$$Var(\hat{\gamma}) = L^{-1}\sigma^2 \quad (12)$$

där variansen av $\hat{\gamma}_j$ är $\sigma^2(\hat{\gamma}_j) = \sigma^2/\lambda_j^2$ och alla kovarianser är noll.

Eftersom principalkomponenter är ortogonala, så blir kvadratsummorna för varje principalkomponent lika och varje regressions kvadratsumma kan beräknas individuellt som

$$SS(\gamma_j) = \hat{\gamma}_j^2\lambda_j^2 \quad (13)$$

Om alla principalkomponenter används i samma modell, blir resultatet lika men den sedvanliga regression modellen (OLS regression). Skattning av β ges från $\hat{\gamma}$ som

$$\hat{\beta} = V\hat{\gamma} \quad (14)$$

och regressionsekvationen kan skrivs som

$$\hat{Y} = 1\overline{Y} + W\hat{\gamma} \quad (15)$$

eller

$$\hat{Y} = 1\overline{Y} + Z\hat{\beta} \quad (16)$$

Idén bakom PCREG är att eliminera de dimensioner (*s*) som orsakar kollinjäritet, dvs. dimensioner som har minsta λ_j. I denna uppsats kommer vi själv att välja dessa dimensioner (g = p - s) oavsett värdet för λ_j. Därefter skattar vi regressionsparametrar med (g*)* antal principalkomponent γ', variansen $s^2(\gamma')$ och kvadratsumman för regressionen $ss(\gamma'_j)$ i en regressionsmodell av Y på W. Medelkvadratresidualen från en full modell används för att skatta σ^2 som är $s^2(\gamma')$.

Vi testar hypotesen $\gamma'_j = 0$ för varje *j*. Student t-test eller F-test används för att fastställa detta. Enligt teorin reducerar vi principalkomponenter från modellen som skapar kollinjäritets problem (ett förslag till detta kriterium kan vara konditions index > 10).

Signifikanta slutsatser från modellen eller modellparametrarna görs genom

γ' (g) är vektor av skattade regression koefficienter som skattas med

$$SS(reg) = \text{summa av } SS(\gamma'_j) \quad (17)$$

där summeringen är över g komponenter som är skattade från SS(reg) med g frihetsgrader.

I nästa steg transformerar vi regressions koefficienterna från principalkomponentsregression till regressions koefficienten till de ursprungliga oberoende variabler (som är centrerade och skalade) genom:

$$\underset{(p\times 1)}{\beta^{+}_{(g)}} = \underset{(p\times g)}{V_{(g)}} \cdot \underset{(g\times 1)}{\hat{\gamma}_{(g)}} \quad (18)$$

Vi använder beteckningen β^+ i stället för $\hat{\beta}$ för att urskilja principalkomponent skattning av β från minsta-kvadratskattningar. Lägg märke till att det finns (*p)* element i $\beta^{+}_{(g)}$ fast det finns bara (g) element i V och $\hat{\gamma}$.

vilka har skattade variansen

$$Var\left[\beta^{+}_{(g)}\right] = V_{(g)}L^{-1}_{(g)}V'_{(g)}\sigma^2 \quad (19)$$

Dessa varianser består av bara stora egenvärdena. De minsta som orsakar varians inflation i OLS är eliminerade. Kvadratsummorna för regression erhölls från (g) principalkomponenter som tillskott med (g) frihetsgrader:

$$SS(\text{Regression}) = \sum_{j\epsilon(g)} SS(\gamma_j) \quad (20)$$

där summering är över en delmängd av (g) principalkomponenter som erhölls i modellen.

Regressions ekvationen är antingen

$$\hat{Y}_{(g)} = 1\overline{Y} + Z\beta^{+}_{(g)} \quad (21)$$

eller

$$\hat{Y}_{(g)} = 1\overline{Y} + W_{(g)}\hat{\gamma}_{(g)} \quad (22)$$

där $W_{(g)}$ är matris of erhållna principalkomponenter; $\hat{\beta}_0 = \overline{Y}$ och är ortogonal till varje $\beta^{+}_{(g)}$.

Variansen av $\hat{Y}_{(g)}$ kan skrivas på flera sätt. Den enklaste är:

$$\mathrm{Var}\left[\hat{Y}_{(g)}\right] = \left[\frac{J}{n} + W_{(g)}L_{(g)}^{-1}W_{(g)}^{'}\right] \quad (23)$$

Koefficienterna från principalkomponentsregressionen kan skrivas som en linjär funktion av en MK- skattning

$$\beta_{(g)}^{+} = V_{(g)} \cdot V_{(g)}^{'}\hat{\beta} = \left[I - V_{(s)} \cdot V_{(s)}^{'}\right]\hat{\beta} \quad (24)$$

där $V_{(s)}$ är matris av (s) egenvektor som elimineras från analysen.

Eftersom $\hat{\beta}$ är ”unbiased”, förväntade och ”biasen” of principalkomponentsregressions koefficienterna följer ekvationen enligt ovan

$$\epsilon\left[\beta_{(g)}^{+}\right] = \beta - V_{(s)} \cdot V_{(s)}^{'}\beta \quad (25)$$

eller biasen är

$$\mathrm{Bias} = \epsilon\left[\beta_{(g)}^{+}\right] - \beta = -V_{(s)} \cdot V_{(s)}^{'}\beta \quad (26)$$

Det är faktum att $\beta_{(g)}^{+}$ har (g) element, en regressionskoefficienten för varje oberoende variabel, även om endast (g) regressionskoefficienter $\hat{\gamma}_{(g)}$ uppskattades innebär att det finns linjära restriktioner för $\beta_{(g)}^{+}$. Det finns en linjär begränsning för varje eliminerad principalkomponent. Den linjära restriktion för $\beta_{(g)}^{+}$ definieras av $V_{(s)}$ som

$$V_{(s)}^{'}\beta = 0 \quad (27)$$

Till slut kan koefficienterna tolkas som de sedvanliga regressionskoefficienterna och jämförs med koefficienter från OLS. Minimering av medelkvadratsumman gör att konfidensintervallet blir mindre vilka i sin tur gör att modellen blir mer stabil. Skillnaden mellan R2 från OLS och PCREG borde inte vara så stor.

Partial Least Squares Regression (PLSREG)

Regression med partiell minstakvadratmetoden ”Partial least squares regression” utvecklades ursprungligen av Wold (1966) som en ekonometriteknik för att modellera ”paths” av kausala relationer mellan ett antal ”block” av variabler. PLSREG har använts i olika sammanhang inom kemi, ekonomi, medicin, psykologi och farmaciutbildningen där linjära modeller, speciellt när ett stort antal förklaringsvariabler är nödvändig. Särskilt i kemometri, har PLSREG blivit ett vanligt verktyg för modellering av linjära förhållandet mellan flera mätningar (de Jong, 1993).

Kort sagt, PLSREG är förmodligen den minst restriktiva av olika multivariata utvidgningar av MLREG. Denna flexibilitet gör att man kan använda den i situationer där traditionella multivariata metoder är starkt begränsad, till exempel när det finns färre observationer än förklaringsvariabler. Dessutom kan PLSREG användas som en inledande analys för att välja lämpliga förklaringsvariabler och identifiera extrema observationer inför en klassisk linjär regressionsanalys.

Observera att PLSREG gäller även för mer allmänna statistiska metoder som :

- Enkel och multipel regression
- Factorial regression
- Polynomial regression
- ANOVA, ANOCOVA, Factorial ANOVA, MANOVA och MANCOVA
- Hierarkiska metoder
- Överlevnadsanalys (PLS-Cox)

Dessa metoder tas ej upp i denna uppsats.

SAS erbjuder en ytterligare PLSREG metod som heter ”**R**educed **R**ank **R**egression” (RRR) som inte heller tas upp i denna uppsats. Men resultatet av en RRR analys presenteras i Tabell-2a och Tabell-2b för den nyfikne läsaren.

Nonlinear iterative partial least squares (NIPALS) Algoritm

Standardalgoritm för beräkning av komponenter för NIPALS är ett icke-linjärt iterationsförfarande (Nonlinear Iterative Partial Least Squares (NIPALS)). Det finns många varianter av NIPALS-algoritmen som normaliserar eller inte normaliserar vissa vektorer.

SIMPLS Algoritm

Ett alternativ skattningsmetoden för komponenter i partiell minsta kvadratmetod regression är SIMPLS-algoritm (de Jong, 1993) som kräver mindre beräkningstid.

Den viktigaste delen i PCREG eller PLSREG metoden är valet av antal komponenter som inkluderas i en PCREG eller PLSREG modell. Detta kan utföras genom valideringskriterier se nästa sektion. Optimalt antal komponenter bestäms genom en empirisk metod som heter ”korsvalidering” av en PLSREG modell genom ökning av antalet komponenter. Modell med minsta PRESS värde antas vara den ”bästa” modellen enligt Allen (1971).

PRESS-kriterier och R2(pred)

PRESS står för ”**Pr**edictive **E**rror **S**um of **S**quares”, eller ”**PRE**diction **S**um of **S**quares”. Den beräknas genom att summera alla prediktorers residualer under korsvalidering. Ett lågt PRESS-värde anger en bra predicerad modell (0 är optimal).

PRESS kan användas för att hitta det optimala antalet komponenter med ett stegvist urvalsförfarande. Den ”bästa modellen” består av så få prediktorer som möjligt med det lägsta (eller nästan lägst) PRESS - värdet. I Figur-5a och Figur-5b ser man ett exempel på en hypotetisk variabel urvalsförfarandet, vilket resulterade i en ”bästa” modell med 3 prediktorer i båda fallen, i vårt fall tre komponenter i PLSREG eller PCREG.

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).

Variationsförklaring av X och Y vid PLSREG, PCREG och SIMPLS-algoritms modeller

PLSREG modellen beräknar dessa variationsförklaringar genom:

$$X = T P' + E \qquad (28)$$

och

$$Y = U Q' + F \qquad (29)$$

Där X och Y är matrisen av förklaringsvariabler och en responsvariabel. Matrisen på högre sidan av likhetstecknet definieras genom:

T = X- värden
P = X- laddningar
E = X- residualer

U = Y- värden
Q = Y- laddningar
F = Y- residualer

PLS algoritmer väljer ortogonala komponenter som maximerar kovariansen mellan varje X-värde och Y-värde. För en välanpassad modell visar de första skattade komponenterna en hög korrelation mellan X- och Y- värdena. Korrelationen brukar öka från en komponenten till den andra. I SAS kan man använda en särskild **`%macro`** för att rita detta i en graf (**`%plot_scr`**), men dessa beräkningar presenteras numeriskt i tabell.1a och tabell.1b.

Ridge regression

Ridge regression utvecklades 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).

Ridge regression löser problemet genom att skatta regressionslinjens koefficienter med hjälp av

$$\hat{\beta} = (X'X + \delta I)^{-1}X'y \qquad (30)$$

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-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

$$V'(X'X + \delta I)V = \begin{bmatrix} \lambda_1 + \delta & 0 & \cdots & 0 \\ 0 & \lambda_2 + \delta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k + \delta \end{bmatrix} \qquad (31)$$

Egenvärdet för den nya matrisen $(X'X + \delta I)$ är $\lambda_i + \delta$ för $i=1,2,\ldots, k$ 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

$$\sum_i \frac{\mathrm{Var}(b_{i,R})}{\sigma^2} = \sum_i \frac{\lambda_i}{(\lambda_i+\delta)^2} \qquad (32)$$

Därför kan δ i ridge regression dämpa de skadliga effekterna av små egenvärden som leder till kollinjäritet. Det finns flera metoder för att välja minskningsparametern δ. Ett alternativ vid val av δ värde kan vara den grafiska metoden som kallas för ”ridge trace” (se Figur2a - Figur3c).

Ridgespår (kriterier för att välja ”ridge” koefficient)

En metod för att bestämma värdet på $\boldsymbol{\delta}$ vid ”ridge” regression att erhålla ”ridge” regressionsresultat göra flera val av **δ** nära noll, t.ex inom intervallet 0,05- 0,20. Beräkningarna av de enskilda regressionslinjernas koefficienter plottas mot standardiserade **δ** eller **VIF,** givet en viss ”ridge” koefficient.

Figur av **R2**, **tr(Var[B(δ)])/q²** och **[B(δ)B(δ)]½**; mot **δ** kan också vara till hjälp (se Figur-2a - Figur-2c). Värdet för bästa **δ** är den minsta värde när stora förändringarna i **B (δ)** och **Var [B (δ)]** har realiserats och determinationskoefficienten **R2** inte har minskat för mycket. Detta synsätt är subjektivt och utmärkande för stabilitet hos modellen som till viss del beror på skalnivån av figuren.

Ett annat alternativ anges av Hoerl, Kennard och Baldwin (1975):

$$\boldsymbol{\delta} = ps^2/[\boldsymbol{B}(o)'\boldsymbol{B}(o)] \qquad (33)$$

där **p** är antalet parametrar utom $\mathbf{B_o}$ och $\mathbf{s^2}$ är kvadratmedelvärdet av residualerna från den sedvanliga regression (OLS) där **δ = o**. Nämnaren i ekvationen är kvadratsumman av regression koefficienterna **B(o)** från den sedvanliga minstakvadratmetoden, förutom intercept, som skattas med centrerade och standardiserade oberoende variabler **Z**.

Det finns en tendens att använda alltför stora värdet på **δ** när man väljer **δ** baserat på ”ridge trace” (Van Norstrand, 1980). Därför är det förmodligen bäst att lägga större vikt på värdet av **δ** som bestäms ur ekvationen enligt ovan. Regressionslinjens resultat med en utvalda värden bör dock underlätta valet av **δ**.

Korsvalideringsanalys

Ett alternativ metod för val av antal komponenter vid PCREG är korsvalidering för att förbättra modellskattningen (se Figur-5a och Figur-5b). Genom att välja ut endast en del av tillgängliga data (t.ex. demografiska variabler) och att mäta hur väl modeller med olika antal komponenter passar data jämfört med den andra delen av data (t.ex. socioekonomiska variabler) . Detta kallas validering test. Normalt är dock ovanligt att ha tillräckligt många förklaringsvariabler för att göra båda delarna för att vara stora nog som ett test för validering vilka i sin tur kan vara användbara och representativ för populationen. Ett annat alternativ till korsvalidering (Efron och Tibshirani 1993) är "Bootstrap" metoden vilken syftar till att minimera medelkvadratfel i skattningar.

Alternativet är att man göra flera olika grupperingar av observerade data för att fastställa och testa antagandena. Detta kallas för *korsvalidering* och det finns flera olika typer. En korsvalideringsmetod, ”utelämna-en-i-taget” eller ”Leave-one-out cross-validation”, en observation i taget utlämnas. En annan metod är att utelämna ett block av observationer, till exempel observationer 1 till 7, sedan observation 8 till 14 och så vidare, vilket är känt som block-validering. En liknande metod är ”spilt-sample korsvalidering”, där successiva grupper enligt vissa kriterier bildar en testgrupp.. Till exempel observationer (1, 11, 21,. . .), sedan observationer (2, 12, 22,. . .) och så vidare. Slutligen kan ett slumpmässigt test uppsättningar väljas från de observerade data, detta kallas ”Random sample cross validation” (SAS User Guide 9.2, 2009).

Alla dessa metoder är implementerade i PROC PLS SAS 9.2. För mer detaljer om algoritmerna hänvisas till själva referenserna i SAS dokumentation. Grafen från denna procedur presenteras i Figur-5a och Figur-5b.

Vilken validerings metod man väljer för att minimera ”predicted residual sum of squares” (PRESS) beror helt enkelt på data, och en alternativ testmetod föreslås av van der Voet (1994) inför val av antalet komponenter. Metoden är lämplig där man har tillräklig många observationer i data för att skapa olika grupperingar, samt tillräkligt stor för att vara representativt för populationen.

Testmetoden ”van der Voet”

Anta $R_{i,jk}$ är j predicerade residualer för respons k vilka har i skattade komponenter, då PRESS ges av $\sum_{jk} R_{i,jk}^2$. Anta i_{min} är antal komponenter för vilka PRESS är minimerad. Kritiskt värde för *van der Voet* test är baserad på skillnader mellan skattade residualer i kvadrat

$$D_{i,jk} = R_{i,jk}^2 - R_{i_{min},jk}^2 \qquad (34)$$

Ett alternativ för kritisk värde är $C_i = \sum_{jk} D_{i,jk}$, vilket är skillnaden mellan PRESS för i och i_{min} komponenter. Ett alternativ som van der Voet föreslår är *Hoteling* T^2

$$C_i = d_{i,}' S_i^{-1} d_i \qquad (35)$$

där d_i är summan av vektor $d_{ij} = \{D_{i,j1}, \dots, D_{i,jNy}\}'$ och S_i är kvadratsumman och korsproduktsmatris

$$S_i = \sum_j d_{i,j} d_{i,j}' \qquad (36)$$

Betydelsen av *van der Voets test* erhålls genom att jämföra C_i med fördelning av värdena som uppstår på grund av slumpmässiga utbyten av $R^2_{i,jk}$ och $R^2_{i\ min,jk}$. I praktiken är *Monte Carlo* urval av sådana värden som simuleras och värde som signifikansnivån är approximerad som proportion av simulerade kritiska värden som är större än C_i.

”Gabriel‘s Biplot”

Korrelationsstrukturen mellan förklaringsvariabler presenteras med Gabriel’s Biplot (Gabriel 1971, 1981). Detta är en formativ presentation av data och förhållanden 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

$$\frac{\lambda_1+\lambda_2}{\sum \lambda_i} \tag{37}$$

Biplot använder singularvärden av $\mathbf{Z}=UL^{1/2}V'$ (se ekvation 4)

Resultat

De vanliga minsta-kvadratskattningarna (OLS) av regressionskoefficienterna från båda dataseten stämmer inte jämfört med den information som man får från korrelationsmatrisen vilka tyder på multikolinjäritet. Däremot är skattningar av regressionskoefficienter från PCREG mer trovärdiga. Om vi börjar med att jämföra tecknen på korrelationerna med tecknen för regressionskoefficienterna ser vi att överstämmelsen är bättre med PCREG. Ett annat argument är det skattade standardfelet, som är betydligt mindre vid PCREG. I tabell-1 redovisas ett exempel på resultatet från PCREG och PLS jämfört med MLREG för en typiskt observation (Afghanistan) från data mot den observerade värde (X)

Tabell-1 Beräknad medellivslängd för Afghanistan (X11=43,5) med tre olika regression metoder.

	REG (BETA)	REG B(X)	PLS (BETA)	PLS B(X)	PCREG (BETA)	PCREG B(X)	Observerade (X) Afghanistan
Intercept	**83,22**	83,22	**60,75**	**60,75**	**59,93**	**59,93**	
X2: Befolkningsförändring	**0,33**	1,089	**-0,24**	**-0,792**	**-0,3**	**-0,99**	**3,3**
X3:Stadsbefolkning i procent av befolkningen	**0**	0	**0,03**	**0,6**	**0,03**	**0,6**	**20**
X4: Antal födda barn per kvinna	**-0,46**	-3,22	**-0,48**	**-3,36**	**-0,48**	**-3,36**	**7**
X12:Andel av befolkningen i åldern 0 – 14	**-0,26**	-10,92	**-0,08**	**-3,36**	**-0,08**	**-3,36**	**42**
X13: Andel av befolkningen i åldern 15 – 59	**-0,01**	-0,53	**0,09**	**4,77**	**0,1**	**5,3**	**53**
X14:Andel av befolkningen i åldern 60+	**0,24**	1,2	**0,12**	**0,6**	**0,13**	**0,65**	**5**
X15: Spädbarnsdödlighetsfrekvens per 1000 levande födda	**-0,07**	-11,41	**-0,02**	**-3,26**	**-0,02**	**-3,26**	**163**
X16:Födelsefrekvens per 1000 människor	**0,03**	1,56	**-0,07**	**-3,64**	**-0,07**	**-3,64**	**52**
X17: Dödlighetsfrekvens per 1000 människor	**-1**	-22	**-0,2**	**-4,4**	**-0,14**	**-3,08**	**22**
X42: Inkomstkällor i procent Jordbruk	**-0,04**	-2,08	**-0,05**	**-2,6**	**-0,05**	**-2,6**	**52**
X43: Inkomstkällor i procent Industri	**-0,05**	-1,65	**0,04**	**1,32**	**0,04**	**1,32**	**33**
X44: Inkomstkällor i procent Tjänstesektorn	**0**	0	**0,03**	**0,45**	**0,03**	**0,45**	**15**
X45: Andelen människor sysselsatta med jordbruk i procent av folkmängden	**0,05**	3,05	**-0,03**	**-1,83**	**-0,03**	**-1,83**	**61**
X46: Andelen sysselsatta inom industrin i procent	**0,07**	0,98	**0,07**	**0,98**	**0,07**	**0,98**	**14**
X47: Andelen sysselsatta med service i procent	**0,12**	3	**0,04**	**1**	**0,04**	**1**	**25**
X50: Vuxnas läs- och skrivkunnighet i procent	**-0,03**	-0,87	**0,04**	**1,16**	**0,03**	**0,87**	**29**
X11: Skattade medellivslängd		41,419		**48,388**		**48,98**	**43,5***

* observerat värde för X11.
De gul markerade kolumnerna visar skattade beta koefficienterna från tre olika regressionsmetoder MLREG, PLSREG och PCREG. Både PLSREG och PCREG är skattade med den första komponenten. Blåmarkerade celler visar variabler som med hjälp av PLS och PCREG har fått tillbaka sina betydelse i regressionssammanhang.
B(X) kolumnerna visar multiplikationen med observerade värde för X-variabler, vilka summeras i sista raden till Skattade X11.

Från Tabell-1 går att avläsa att variabler som X3, X13, X16, X43, X44 och X50 har fått tillbaka sina betydelse vid både PLSREG och PCREG. Till exempel ser man att X50 ökar medellivslängden och inte tvärtom. En sånt tolkning önskas vid de flesta studier inom biostatistik och epidemiologi. En sammanställning av skattade koefficienterna från olika valmöjligheter jämfört med de sedvanliga regressionskoefficienterna redovisas i Tabell-2a.

I PCREG, jämfört med PLSREG är man mer fri att välja komponenter som är mer relevanta ur studiens perspektiv att ingå i modellen. Vid PLSREG kan man bara välja komponenterna i ordning från 1 och uppåt. Regressionskoefficienter från PCREG och PLSREG är identiska när man använder den första PC. Koefficienterna från multipel regression är identisk med RRR. De flesta mjukvaror förutom SAS och MiniTab kräver ett särskild tillägg för att utföra PLS och SIMPLS. För att utföra analysen med SAS måste man skriva egna kommandon (PROC IML) för att skatta regressionskoefficienterna för PCREG. Just nu finns ingen mjukvaror som utför PCREG.

Att reducera onödig variation (medelkvadratfel) i modellen med RREG jämfört med PCREG och PLSREG var inte så användbart. Metoden fungerar bara när några av det totala antalet variabler är multikollinjära. PLSREG, med tre olika val, ger inte heller ett bättre svar än PCREG. Med PCREG får vi välja vilka komponenter som ska vara med vid skattning av regressionskoefficienter. Regressionsdiagnostiska metoder utelämnas i denna uppsats utan jämförelser och bedömningskriterier är R2 och grafiska illustrationer.

Tabell-1a visar förklarad varians för X och Y vid val av antal komponenter. Vid val av bara en komponent i modellen är resultaten nästan identiska med varandra. Men när man ökar antalet komponenter i modellen (g) lägger PCREG mer vikt på X-variablerna än de andra metoderna. Tabellen ska vara avgörande i fall man vill betona vikten på Y-respons[1] eller på X-variabler för att välja någon av PLS- metoderna. En annan avgörande faktor är valet av antal komponenter, då man ser hur vikten förändras med ökande antal komponenter. I detta fall är tre komponenter optimalt och PCREG bedöms vara acceptabel jämfört med de övriga metoderna. Observera att detta inte betyder att de andra metoderna ger dåliga skattningar, utan vårt syfte är att använda PCREG.

Tabell-1a Procent av variation förklarad varians för olika regressionsmodeller jämfört med PCREG baserat på första dataset.

	W-Y g=1	**W-X_i g=1**	**W-Y g=3**	**W-X_i g=3**
PLSREG	88,3342	61,7806	95,4843	75,9696
SIMPLE	88,3342	61,7806	95,4843	75,9696
RRR	96,4598	56,1530		
PCREG	86,7281	61,8456	89,9063	80,9063

g = antal komponenter, Y= X11, Xi = X2-X50, W = procent av variation vikt i modellen (se sidan 9)

Tabell-1b visar hur de olika metoderna förklarad varians för X eller Y vid val av antal komponenter för Longley dataset. Siffrorna skiljer sig i decimaler i detta fall vilka beror på antal observationer i data, men ändå ser man hur PCREG har med lite decimal lagt mer vikt eller variation förklaring på X-variabler.

Tabell-1b Procent av variation förklarad varians för olika regressionsmodeller jämfört med PCREG baserat Longley dataset.

	W-X g=1	**W-Y g=1**	**W-X g=3**	**W-Y g=3**
PLSREG	76,6545	92,5743	99,7022	98,6238
SIMPLE	76,6545	92,5743	99,7022	98,6238
RRR	70,9455	99,5479		
PCREG	76,7230	91,4253	99,7024	98,5967

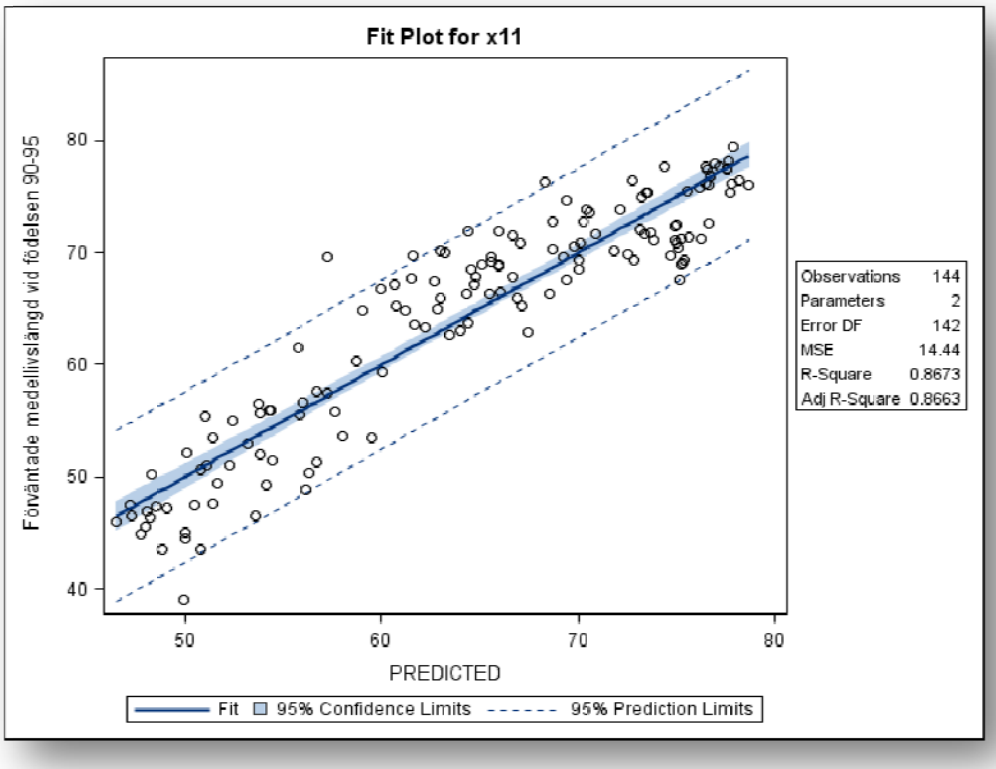
g = antal komponenter

En jämförelse mellan dessa två tabeller visar att variation förklaring (W) vid det första dataset med 16 variabler och 144 observationer skiljer sig från modell till modell beroende på antal komponenter som används i regression modeller, men analys med den andra dataset som har mindre observationer lägger stora vikt på W-Y vid modellen med bara en komponent. När vi ökar antal komponenter i modellen, ökar variationsförklaringen för W-X. En annan anmärkning vad gäller Tabell-1b är att när antalet observationer är få ger alla regressionsmodeller ungefär samma svar oavsett antal komponenter i modellen.

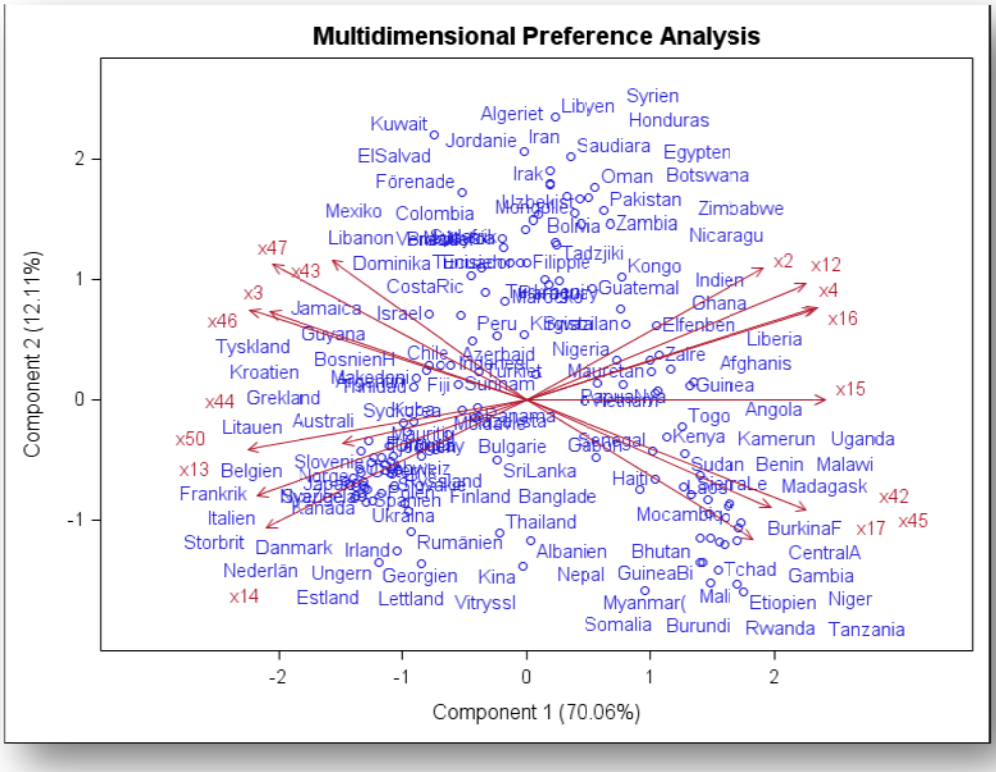
Figure-1 visar förhållanden mellan responsvariabel och skattade respons variabel från PCREG (PREDICTED) med hjälp av den första komponenten (PC1). Modell med den första PC visar tillräcklig stryka med R2 för en godtycklig statistisk slutsats. Modellkoefficienternas betydelse i förhållande till informationen som man har från korrelationerna mellan förklaringsvariabler och X11 stämmer också väl (se Tbell-2a).

[1] Dessa kriterier gäller också för multivariat data analys då antal responsvariabler är många.

Figur-1. Skattade responsvariabel med PCREG och observerade värde för X11.



Fiugur-1a. ”Multidimentional preference analys” som motsvarar Gabirel’s biplot.



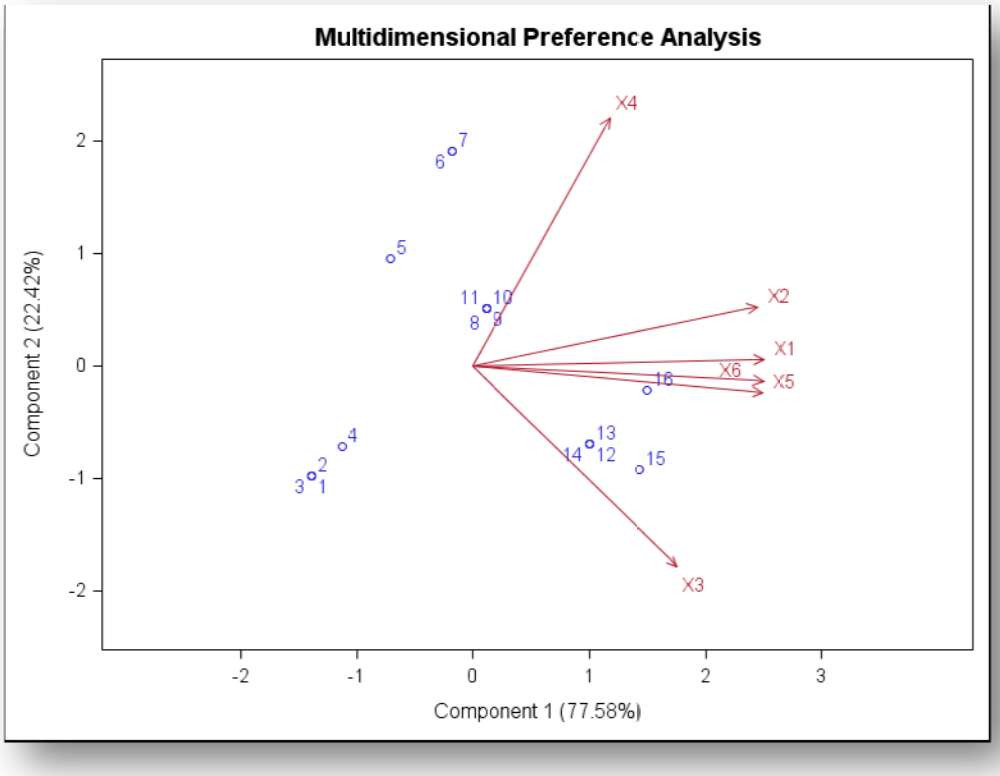
Multidimentional preference analys i SAS 9.2 ger motsvarande graf som Gabriel’s biplot från PCA (Carroll 1972).

Figur-1a visar tydligt att X15 har positivt värde vid den första komponenten och 0 vid den andra. Variablerna X42, X45 och X17 klustras i eget grupp. Vidare X2, X4, X12 och X16 i en grupp, X3, X43, X46 och X47 i en annan grupp och X13, X14, X44 och X50 i en tredje grupp. Om syftet var att gruppera data efter variabler ser man vilka observationer som bidrar till vilka grupperingar.

Dessa variabler är en blandning av både socioekonomisk och demografisk information. Men spädbarnsdödlighet (X15) är den enda variabel som definitivt har sin egen riktning, där de afrikanska länderna, länder med låg medellivslängd och Afghanistan bidrar till detta. En annan tolkning av figuren är att rika länder har identiska förklaringsvariabler (X14, X13, X50, X44, X3, X46, X47 och X43) medan de övriga har andra variabler. Det vill säga att typiska informationskällor för dessa länder kan vara dessa variabler. Observera att X11 inte är involverad i detta sammanhang. En tredje tolkning är att länder som har ett negativt värde för den första och andra principalkomponentens ekvation (Z_i) motsvarar högre medellivslängd (Alinaghizadeh 2009).

Slutlig tolkning av grafen blir att förklaringsvariabler i förhållande till varandra har både negativa och positiva samband. Detta kan vi jämföra med nästa figur med Longley dataset där alla variabler är stark korrelerad mot varandra förutom X4 och X3 som är negativ korrelerad med varandra.

Fiugur-1b. ”Multidimensional preference analysis” med Longley dataset.

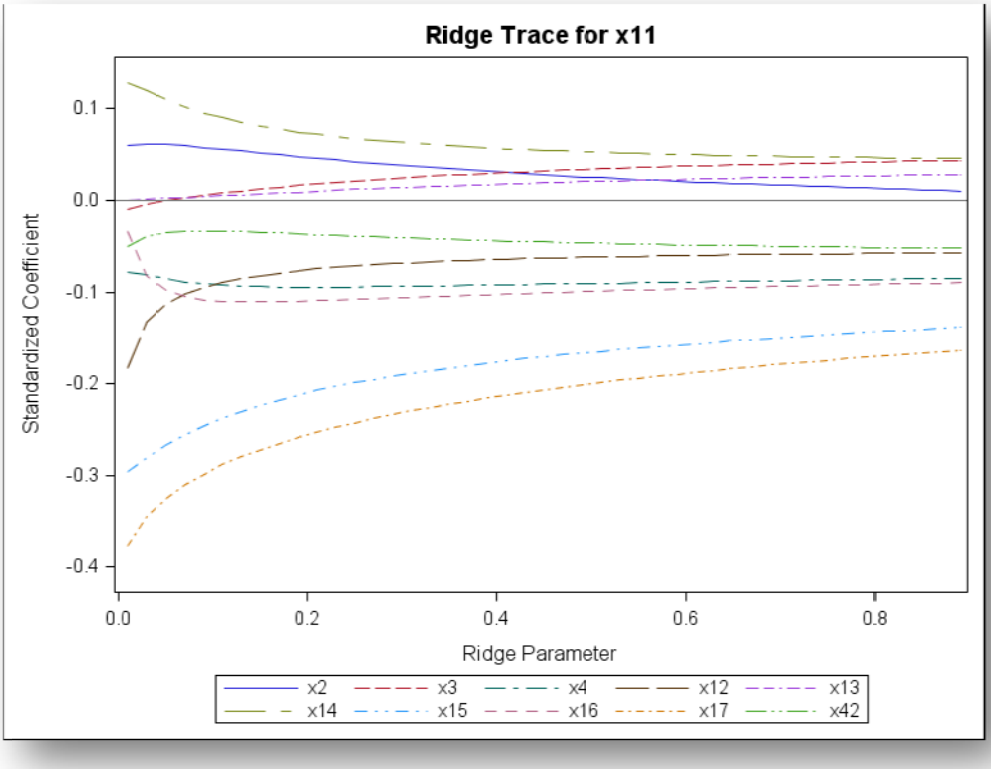


Det som framgår från Figur-1b är att alla variabler har en positiv laddning vid de två första komponenterna förutom X3 som har ett negativ laddning vid den andra komponenten. Båda variablerna X3 och X4 har höga laddningar i den andra komponenten. Den första komponenten förklarar 77,58 % av variansen i data och den andra förklarar 22,42 % av variansen vilket betyder att med två komponent kan vi förklara 100 % av variansen i data. Variablerna X3 och X4 korrelerar lågt med alla variabler (se appendix för korrelationsmatris).

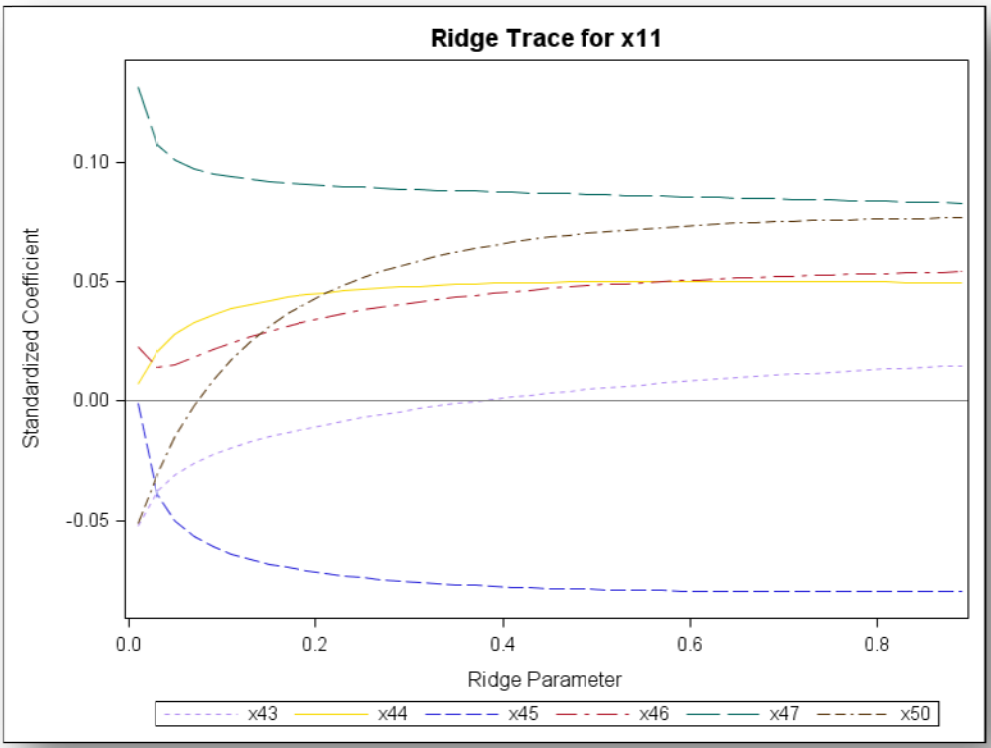
Vidare kan man avläsa att de första observationerna (1-7) inte har så stort inflytande på principalkomponenterna. Observationerna 12, 13, 14 och 15 som förklaras av PC1 och PC2 har störst inflytande på X3. Observation 8, 9, 10 och 11 ger sitt bidrag till X4. Observation 16 är enda observation som har sin inflyttande på X1, X2, X5 och X6.

Figur-1b visualiserar lättare observationernas inflytande på principalkomponenterna samt deras bidrag i samband med förklaringsvariabler till komponentsskattningar.

Figur-2a Ridge spar för X11 med 10 förklaringsvariabler.



Figur-2b Ridge spår för X11 med rest av variabler.

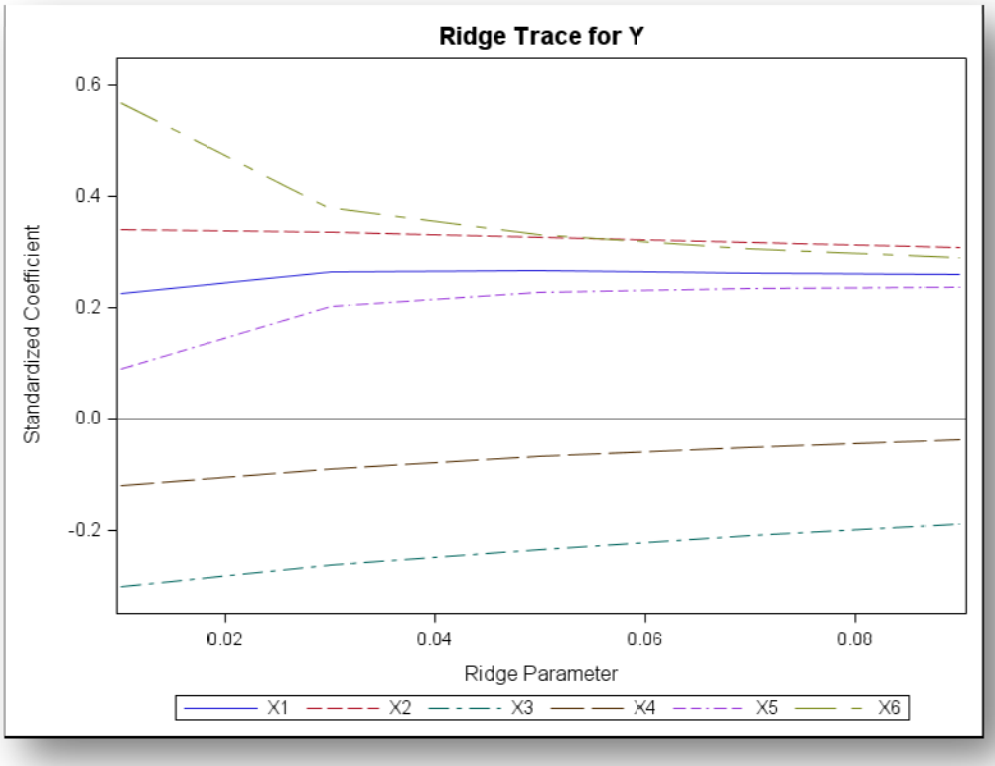


Observera att båda figurerna skapas med en och samma analys, men eftersom antal variabler är för många då programmet delar upp grafen i variablernas ordning i modell.

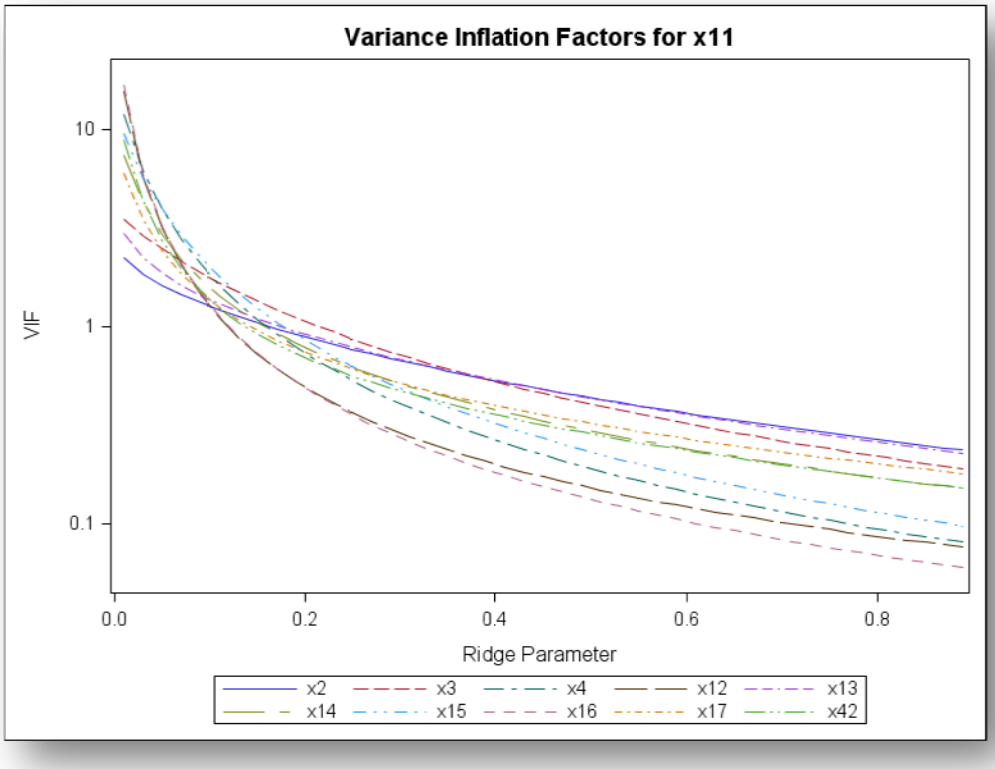
Figurerna 2a och 2b ger underlag vid valet av ”ridge” koefficienten (δ), som ligger i intervallet 0,10-0,30. Enligt ekvation (32-33) blir det beräknade resultatet för δ=0,30. Då har vi information från alla variabler som visas i Tabell-2a. Variabler som skiljer sig mycket från andra kan identifieras från grafen genom att markera dem i ett intervall på högre sidan av grafen, förutsätter

ett approximationsantagande. Variabler som bryter kurvan vid 0,05-0,10 ger en lättare bedömningsmöjlighet i vårt fall.

Figur-2c Ridge spår för Longley dataset med responsvariabeln Employment.



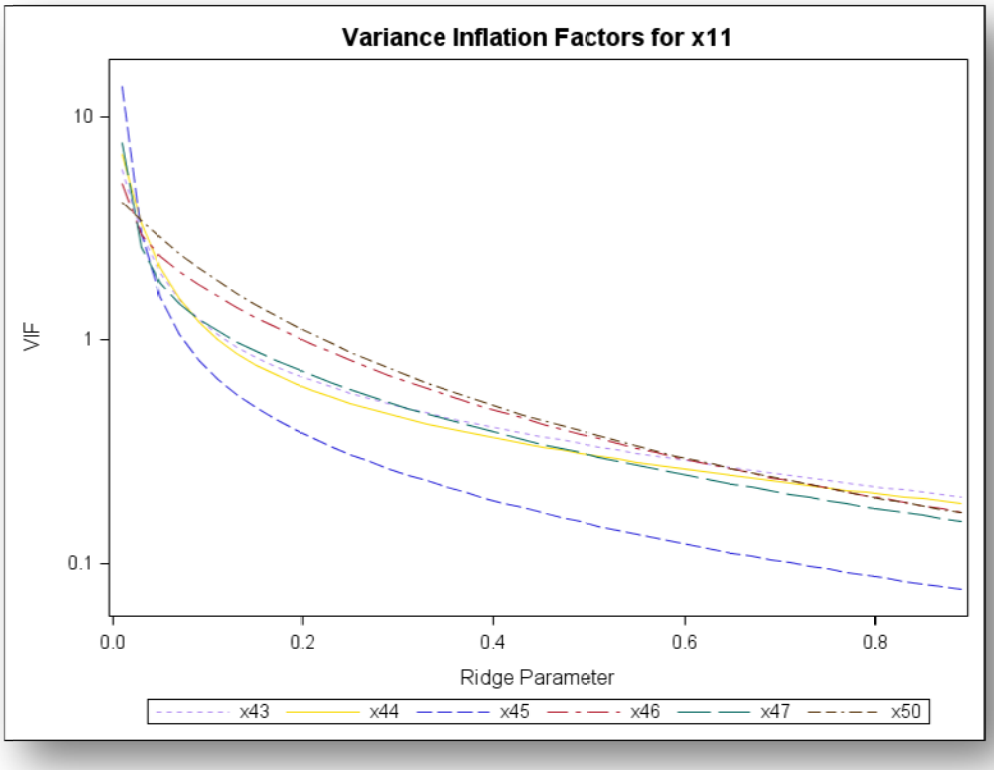
Figur-3a Förändring av VIF värde vid olika värde av ridge koefficient.



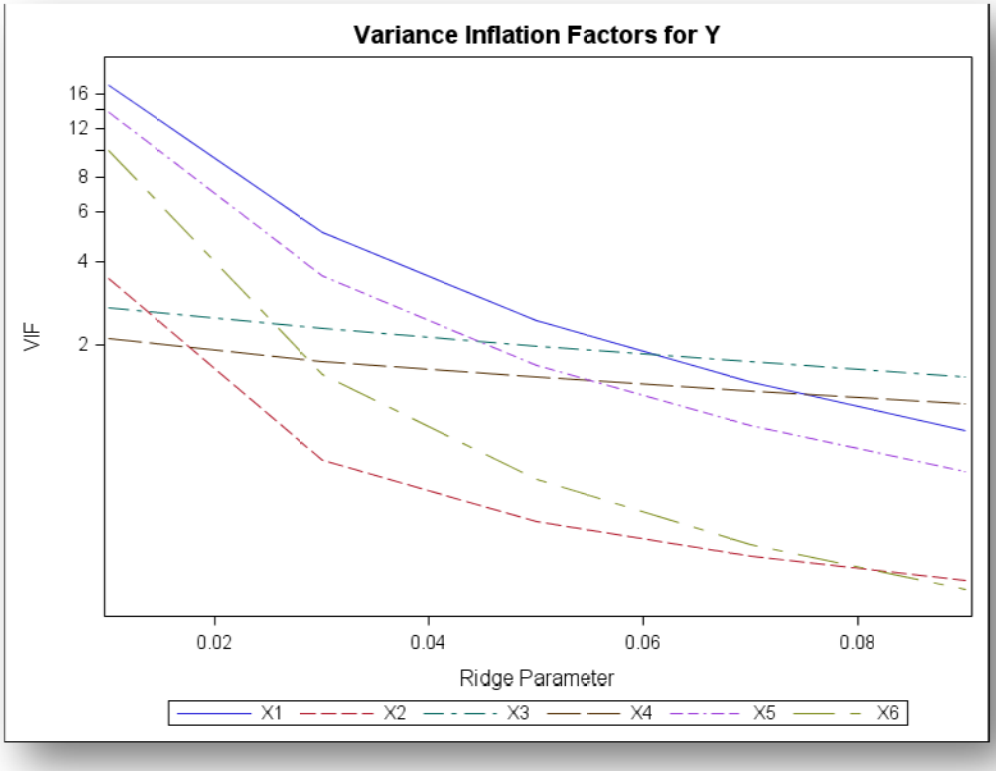
Figur-2c ger underlag för att välja ridge koefficienten (δ) för Longley dataset vilka är lättare och den ligger inom intervallet 0,02-0,03. Enligt ekvation (32-33) resultatet blir δ= 0,03 vilka vissas i Tabell-2b. De två variabler X3 och X4 som skiljer sig från de andra ligger under 0 och svårt att

välja deras brytningspunkt inför valet av att ”ridge koefficient” verkligen är 0,03, då ett approximativantagande måste göras. Variabler som ger en bra bedömningsmöjligheter är X1, X5 och X6.

Figur-3b Förändring av VIF värde vid olika värde av ridge koefficient.



Figur-3c Förändring av VIF värde vid olika värde av ”ridge” koefficient för Longley dataset.

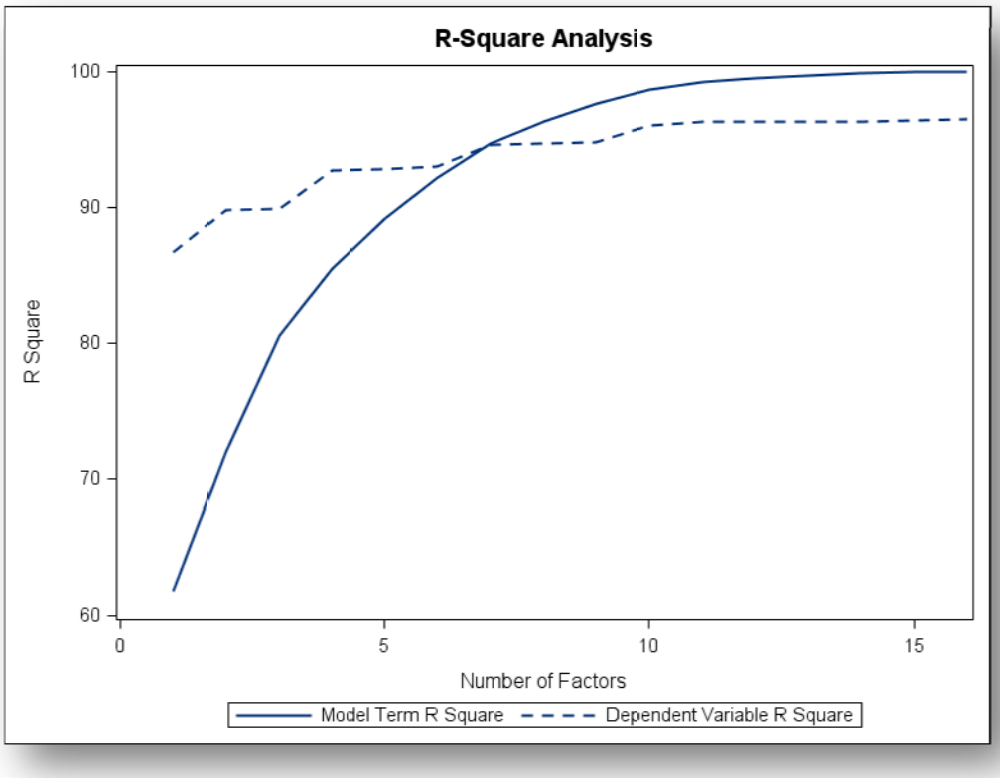


Figurerna 3a och 3b säger samma sak som Figurerna 2a och 2b men med utgångspunkt från VIF, dvs gränsvärdet för ”ridge” koefficienten är mellan 0,10 och 0,30. Men linjernas lutning är

fortfarande avtagande även efter 0,80 vilka gör det svårt att välja ett gränsvärde för ”ridge” regression. Det är önskvärt att dessa linjer ska minska sin lutning vid ett viss gränsvärde precis som i Figurerna 2a och 2b, men eftersom antalet variabler är stort och korrelationerna höga kommer dessa lutningar att fortsätta vidare. Alltså, det blir svårt att kombinera VIF och ”ridge” koefficienten i en sådan situation. Gränsvärdet 0,10 - 0,30 är igen en optimal approximation för ”ridge” koefficienten.

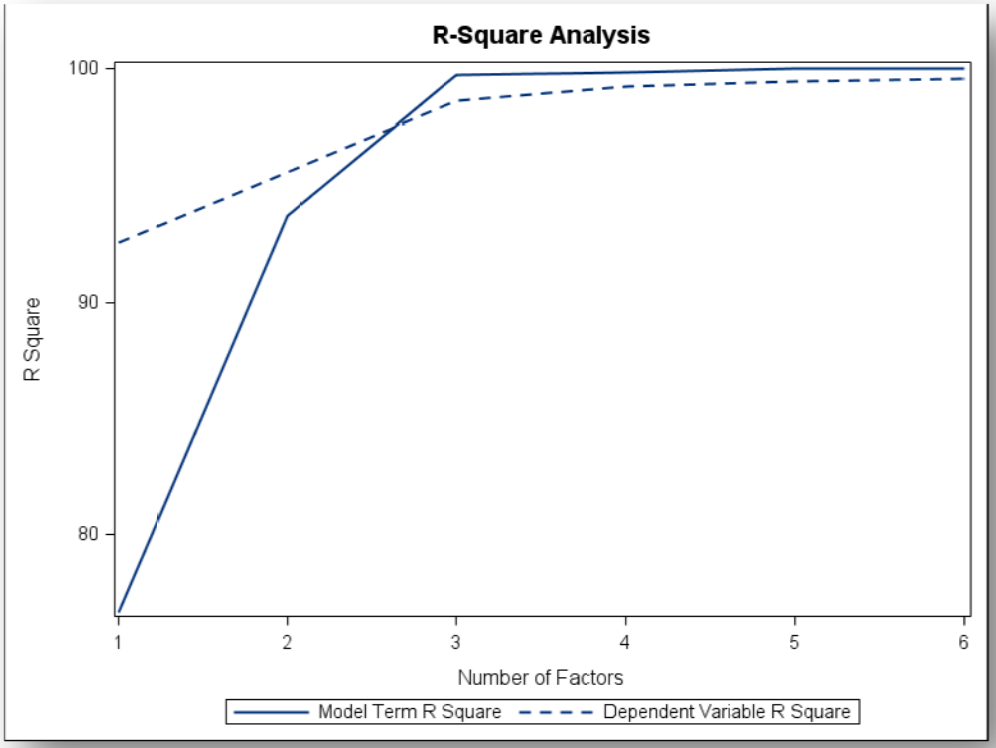
Figur 3c visar att ”ridge” koefficient är mellan 0,02 och 0,03 gentemot VIF=(1/1-R2). Men linjernas lutning är fortfarande avtagande även efter 0,04 vilka gör det svårt att välja ett gränsvärde inför ”ridge regression”. Önskvärt är att dessa linjer ska minska sin lutning vid ett viss gränsvärde. Det blir svårt att kombinera VIF och ”ridge” koefficient även i denna situation.

Figur-4a Determinationskoefficients analys ”R-Square Analysis” från PLSREG med *världs datasetet*.



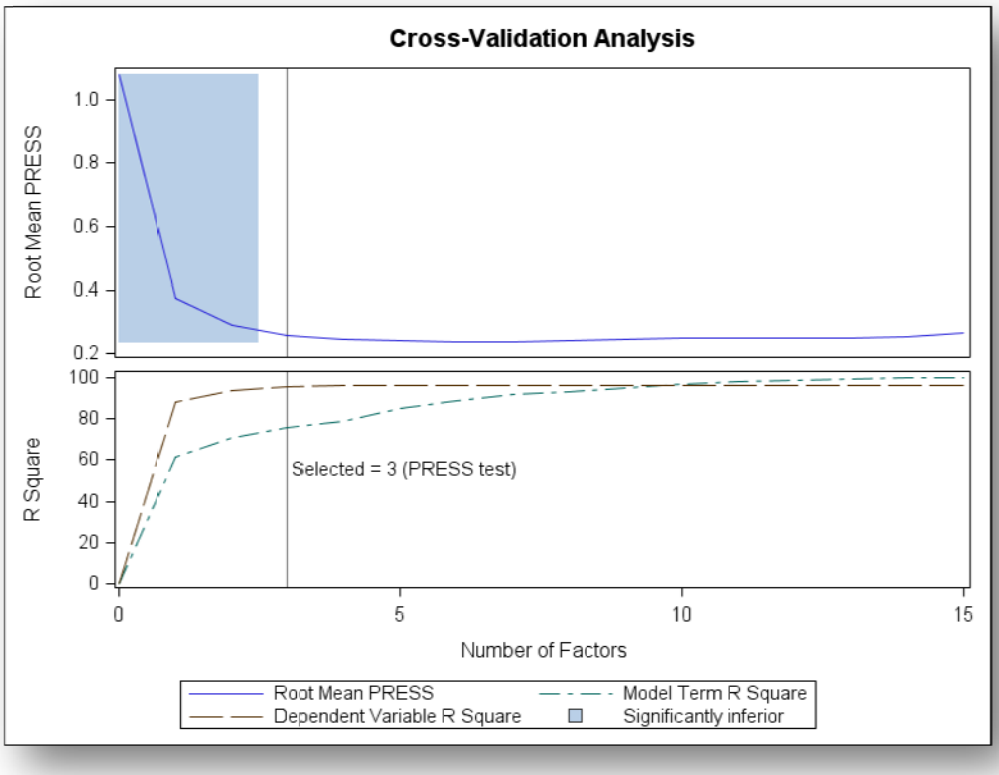
Figur-4a visar att PLSREG proceduren föreslår att 7-8 komponenter ska räcka för att få en bra modell med en bra determinationskoefficient på över 90 % med världsdata. "Model term R Squar" motsvarar procentuella variationsförklaring med antal komponenter och "Dependent Variable R Square" motsvarar procentuella variationsförklaring med Y i samma modell. Dessa siffror presenterades i tabell-1a.

Figur-4b Determinationskoefficients analys ”R-Square Analysis” för *Longley dataset*.



Figur-4b visar att PLS proceduren också föreslår att 2-3 komponenter ska räcka för att få en bra modell med en bra determinationskoefficient på ungefär 98 % (se Tabell-1b). I båda fallen krävs det minst 3 skattade komponenter för att nå det högsta determinationskoefficienten R2.

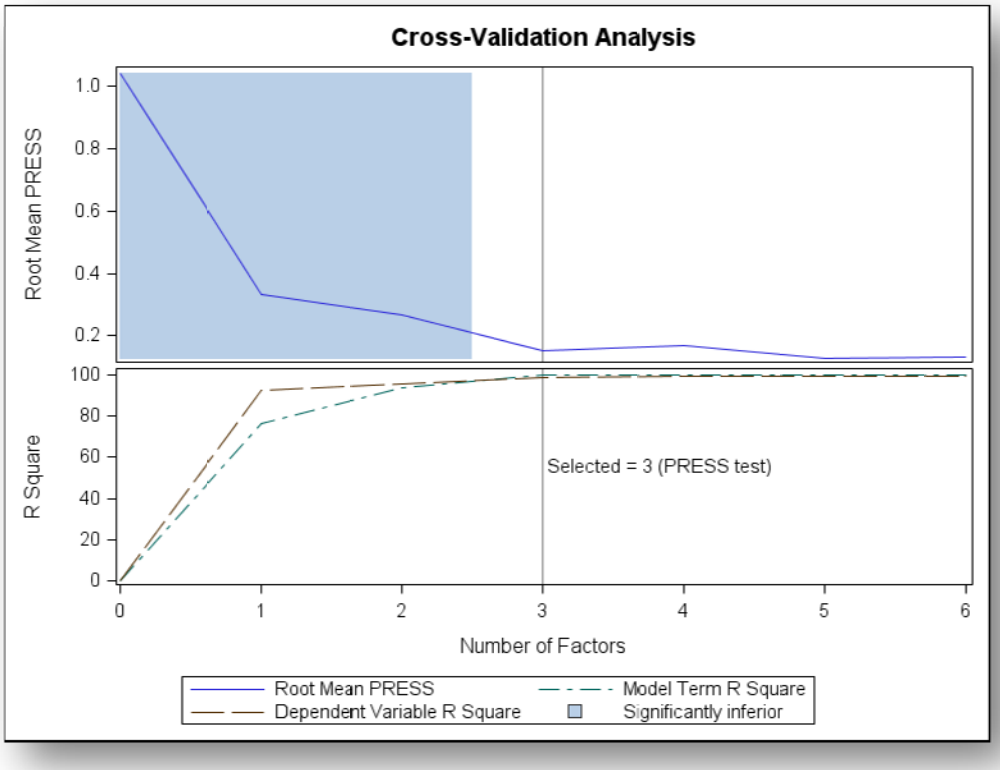
Figur-5a ”Korsvalideringsanalys” från PLS regression med *världsdataset*.



Figur-5a föreslår att tre komponenter ska räcka för att skatta den bästa modellen (PRESS=0,2383), dvs. minsta antal komponenter som ger ett värde för $p < 0,25$ är tre. Det andra

kritiska området är determinationskoefficienten (R2) som ligger nära 90 % vid val av tre komponenter. Men som föregående analys antyder, kvaliteten på observerade data kan inte användas för att välja antal extraherade faktorer, utan dessa skall väljas på grundval av hur väl modellen passar observationerna som inte är inblandade i själva modelleringsförfarandet.

Figur-5b ”Korsvalidering analys” från PLS regression med *Longley dataset*.



Figur-5b föreslår att tre komponenter ska räcka för att skatta den bästa modellen med minsta värde på PRESS=0,1274. Minsta antal komponenter som ger p<0,10 är tre. I detta fall har vi mindre PRESS–värde vilka tyder på att färre observationer i data och mindre antal variabler underlättar approximationer.

Tabell-2a Resultat med första dataset. Multipel regression, Ridge regression (δ =0,10 δ=0,30), Principalkomponentsregression (g=1, g=3, g=4 och g=8). PROC PLS med tre olika val alternativ med första komponent. Medelvärdet (μ) av X11=64,4 SD=10,4. N=144

	Multipel regression	Deskriptiv analys		PROC PLS med en komponent			Principalkomponents regression				Ridge regression	
Variabler/koefficienter	BETA (SE)	Means (SD)	X11 Corr.	SIMPLS	RRR	PLS	g=1 B (SE)	g=3 B (SE)	g=4 B (SE)	g=8 B (SE)	δ=0,30 B (SE)	δ=0,10 B (SE)
Intercept	83,22 (10,27)			60,75	83,22	60,75	59,93	59,33	70,84	70,83	72,06 (4,77)	77,73 (5,51)
X2: Befolkningsförändring	0,33 (0,16)	2,04 (1,77)	**-0,39**	-0,24	0,33	-0,24	-0,30 (0,0055)	0,08 (0,0369)	0,54 (0,1146)	0,60 (0,1284)	0,22 (0,1735)	0,33 (0,1676)
X3:Stadsbefolkning i procent av befolkningen	**0,00 (0,01)**	50,19 (23,22)	0,73	0,03	-0,00	0,03	**0,03 (0,0006)**	0,04 (0,0015)	0,009 (0,0030)	-0,004 **NS** (0,0059)	0,01 (0,0154)	0,003 (0,0157)
X4: Antal födda barn per kvinna	-0,46 (0,41)	3,98 (1,91)	**-0,87**	-0,48	-0,46	-0,48	-0,48 (0,0086)	-0,37 (0,0137)	-0,50 (0,0210)	-0,56 (0,0313)	-0,51 (0,2300)	-0,50 (0,2805)
X12:Andel av befolkningen i åldern 0 – 14	-0,26 (0,10)	35,40 (10,50)	**-0,76**	-0,08	-0,26	-0,08	-0,08 (0,0018)	-0,04 (0,0043)	-0,05 (0,0073)	-0,06 (0,0086)	-0,07 (0,0420)	-0,091 (0,0536)
X13: Andel av befolkningen i åldern 15 – 59	*-0,01 (0,05)*	55,59 (6,97)	0,60	0,09	-0,01	0,09	**0,10 (0,0018)**	0,06 (0,0046)	-0,03 (0,0106)	-0,015 **NS** (0,0300)	0,02 (0,0465)	0,006 (0,0465)
X14:Andel av befolkningen i åldern 60+	0,24 (0,11)	8,80 (5,52)	0,63	0,12	0,24	0,12	0,13 (0,0024)	0,04 (0,0102)	0,10 (0,0221)	0,10 (0,0290)	0,12 (0,0694)	0,174 (0,0801)
X15: Spädbarnsdödlighetsfrekvens per 1000 levande födda	-0,07 (0,02)	52,06 (42,33)	**-0,96**	-0,02	-0,07	-0,02	-0,02 (0,0004)	-0,03 (0,0005)	-0,04 (0,0017)	-0,05 (0,0020)	-0,05 (0,0101)	-0,06 (0,0119)
X16:Födelsefrekveper 1000 människor	0,03 (0,09)	29,90 (13,48)	**-0,88**	-0,07	0,03	-0,07	-0,07 (0,0012)	-0,05 (0,0022)	-0,08 (0,0043)	-0,10 (0,0050)	-0,08 (0,0337)	-0,08 (0,0433)
X17: Dödlighetsfrekvens per 1000 människor	**-1,00 (0,13)**	10,37 (4,26)	**-0,78**	-0,20	-1,00	-0,20	**-0,14 (0,0025)**	-0,29 (0,0150)	-0,57 (0,0288)	-0,60 (0,0315)	-0,57 (0,0833)	-0,72 (0,0944)
X42: Inkomstkällor i procent Jordbruk	-0,04 (0,04)	21,16 (15,70)	**-0,72**	-0,05	-0,04	-0,05	-0,05 (0,0009)	-0,08 (0,0029)	0,008 **NS** (0,0069)	-0,013 **NS** (0,0097)	-0,03 (0,0245)	-0,02 (0,0291)
X43: Inkomstkällor i procent Industri	-0,05 (0,04)	31,85 (12,48)	0,44	0,04	-0,05	0,04	0,04 (0,0007)	0,05 (0,0075)	-0,03 (0,0078)	-0,024 **NS** (0,0136)	-0,003 (0,0270)	-0,02 (0,0304)
X44: Inkomstkällor i procent Tjänstesektorn	**0,00 (0,04)**	46,79 (14,21)	0,46	0,03	-0,00	0,03	**0,03 (0,0006)**	0,06 (0,0029)	0,02 (0,0024)	0,044 (0,0097)	0,03 (0,0242)	0,03 (0,0282)
X45: Andelen människor sysselsatta med jordbruk i procent av folkmängden	0,05 (0,07)	38,75 (29,20)	**-0,88**	-0,03	0,05	-0,03	-0,03 (0,0006)	-0,04 (0,0010)	-0,04 (0,0010)	-0,03 (0,0035)	-0,03 (0,0152)	-0,02 (0,0199)
X46: Andelen sysselsatta inom industrin i procent	0,07 (0,07)	20,55 (12,12)	0,77	0,07	0,07	0,07	0,07 (0,0012)	0,06 (0,0029)	0,09 (0,0059)	0,06 (0,0226)	0,03 (0,0310)	0,02 (0,0337)
X47: Andelen sysselsatta med service i procent	0,12 (0,07)	40,53 (20,32)	0,80	0,04	0,12	0,04	0,04 (0,0007)	0,06 (0,0019)	0,05 (0,0027)	0,033 (0,008)	0,05 (0,0190)	0,05 (0,0227)
X50: Vuxnas läs- och skrivkunnighet i procent	-0,03 (0,02)	74,67 (24,60)	0,83	0,04	-0,03	0,04	0,03 (0,0006)	0,04 (0,0006)	0,07 (0,0029)	0,076 (0,0058)	0,02 (0,0153)	0,006 (0,0160)
R2	96 %			88 %	96 %	88 %	88 %	95 %	96 %	96,3%	89 %	93 %

PCREG med (g=1) innehåller PC1, PCREG med (g=3) innehåller PC1, PC2, PC4, PCREG med (g=4) innehåller PC1, PC2, PC4, PC7, PCREG med (g=8) innehåller PC1-PC8. Markerade koefficienter är signifikant på 5 % nivå, vid OLS och ridge regression. Alla variabler vid PCREG är signifikant förutom de som är markerad med NS. Tolkning av koefficienter görs precis som vid OLS regression. Variablerna X2, X13, X16, X43, X45 och X50 får tillbaka sin betydelse i PCREG jämfört med multipel regressionsmodell.

Tabell-2b Resultat med Longely dataset. Multipel regression, Ridge regression (δ =0,03 δ=0,10), Principalkomponentsregression (g=1, g=3) och PROC PLS med tre olika val alternativ (SIMPLE, RRR och PLS). Medelvärdet av Employment (Y)=65317 SD=3511,97 N=16 .

	Multipel regression	**Deskriptiv analys**		**PROC PLS med tre komponent**			**Principalkomponents regression**			**Ridge regression**	
Variabler/koefficienter	BETA (SE)	Means (SD)	Korre. (Y)	SIMPLS 3 factors	RRR 1 faktor	PLS 3 factors	g=1	g=3* B (SE)	g=3 B (SE)	δ=0,03 B (SE)	δ=0,10 B (SE)
Intercept	-3482258,635 (890420,384)			-389001,3526	-3482258,635	-389001,3526	-258158,4	-2083193	-358712,8	-500076,087 (378805,688)	-367980,643 (299255,1)
X1: PRICES	15,062 (84,915)	101,68 (10,79)	0,97	94,4148	15,062	94,4148	66,98 (1,5700)	-124,1277 * (48,819329)	94,79 * (2,90)	85,655 (76,092)	83,656 (65,668)
X2: GNP	-0,036 (0,033)	387698,44 (99394,94)	0,98	0,0127	-0,036	0,0127	0,007267 (0,0001703)	0,0167614 * (0,0023512)	0,0127 * (0,00052)	0,012 (0,009)	0,011 (0,007)
X3: JOBLESS	-2,020 * (0,488)	3193,31 (934,46)	0,50	-1,1674	-2,020	-1,1674	0,54 (0,0126148)	0,2275083 * (0,0503681)	-1,1615 * (0,1442)	-0,986 * (0,335)	-0,680 (0,425)
X4: MILITARY	-1,033 * (0,214)	2606,69 (695,92)	0,46	-0,6097	-1,033	-0,6097	0,45 (0,0106229)	0,6875878 * (0,0999799)	-0,5987 * (0,16396)	-0,452 (0,96)	-0,160 (0,492)
X5: POPSIZE	-0,051 (0,226)	117424,00 (6956,10)	0,96	0,1446	-0,051	0,1446	0,10 (0,0024381)	-0,385124 * (0,12290289	0,15386 * (0,0055)	0,102 (0,125)	0,120 (0,103)
X6: YEAR	1829,151 * (455,478)	1954,50 (4,76)	0,97	219,0581	1829,151	219,0581	152,84 (3,5827083)	1124,245 * (245,32928)	202,9575 * (5,90684)	278,552 (198,284)	209,340 (156,603)
R2	99 %			98,62 %	99,55 %	98,62	97,13 %	93,7 %	98,60 %	95,5 %	91,13 %

Markerade koefficienter med (*) är signifikant på 5 % nivå. Principalkomponent med (g=1) innehåller PC1. Principalkomponent med (g=3*) innehåller PC1,PC2,PC5, Principalkomponent med (g=3) innehåller PC1, PC2, PC3. PROC PLS använder de tre första faktorer (komponenter) förutom RRR som använder bara den första.

X2, X3, X4 och X5 som har en negativ tecken vid den sedvanliga regressionsmodellen har fått tillbaka sina rätta tecken i PCREG som stämmer med korrelation sambandet. Variabler som har störst betydelse i detta sammanhang är X1 och X6 som ökar (Y) kraftigare än de andra, medan de andra variabler X2, X3, X4 och X5 ökar (Y) med decimaler. En enhet ökning för X1 och X5 ökar Y vilka speglar den naturliga ökningen gentemot multipel regression som ger fel svar att X5 har negativ påverkan. Variabeln som X2 brukar inte öka med en enhet och det normal stigningen enligt data är 100 000 under 5 år och detta betyder att en positiv ökning i Y (Employment) är ett naturlig tolkning av data och självklart denna ökning måste justeras för de övriga variablerna för att tolkningen ska vara korrekt.

Om vi antar vid PCREG att PC1, PC2 och vidare PC5 ska involveras vid parameterskattningar får vi ett annorlunda svar än de övriga "biased" regressionsmodellerna, fast antal komponenter är lika. PLS proceduren tillåter inte val av PC och de väljas ut i ordning. Denna ordning av användning av komponenter är diskuterad av Jolliffe (1982), som kritiserar feltolkning av komponenternas betydelse. I verkligheten inom demografi, medicin och socioekonomiska data, där multipla faktorer används för att skatta regression modellen med förklaringsvariabler ger olika vikt vid olika komponentsladdning. Dessa skattade komponenter får sin karaktär (indexering) efter sina laddningar och efter den första komponenten vill man helst använda komponenter som har en stor laddning av en viss karaktär (som finns hos förklaringsvariabler), oavsett deras variationsförklaring.

Slutsatser:

Ett antal procedurer diskuterades som alternativ till den vanliga minstakvadratmetoden (OLS regression) för att förbättra tolkningen av förhållandet mellan förklaringsvariabler och beroende variabeln och / eller för att lösa problemet med multikollinjäritet bland förklaringsvariabler. Även om det mest populära alternativet historiskt sett har varit PCREG, har det vissa brister och ersätts av metoder som PLSREG och "maximal redundans". Eftersom PLSREG och "maximal redundans" har utvecklats mycket i olika områden, har det ännu inte fastställts vad de relativa fördelarna är med PLSREG förutom tiden som diskuteras av de Jong, S. (1993). En enhetlig behandling av dessa metoder som diskuterades i denna uppsats tillsammans med information om beräkningsmetoder kan hittas i (Jackson 1991).

PCREG definierar alla termer i form av ursprungliga antalet variabler (centrerade och skalade) dvs. prediktorer av X, liksom SIMPLS-metoden. Men till skillnad från både PLS- och SIMPLS-metoderna, väljer PCREG-metod X-weights/X-scores utan hänsyn till respons variabeln Y. X-värdena förklarar så mycket variation i X som möjligt, allt annat är lika. X-vikterna för PCREG-metoden är egenvektorer från kovariansmatrisen **X'X**. Återigen, X-och Y-laddningar definieras som i PLS algoritm, men som i SIMPLS är det lätt och mindre tidskrävande att beräkna de övergripande modellkoefficienterna för ursprungliga antalet variabler (centrerade och skalade) jämfört med de ursprungliga förklaringsvariablerna X.

Ingen av de regressionsmetoder som genomförs i PLSREG (SAS) passar observerade data bättre än PCREG; i själva verket föreslår alla metoder nästa samma svar som PCREG, när antal komponenter ökar, förutom RRR som ger identiskt svar med MLREG oavsett antalet komponenter. Den avgörande punkten är att när det finns många förklaringsvariabler, kan PCREG eller PLSREG anpassa observerade data i en regressionsmodell bättre än andra metoder. Däremot kan den sedvanliga regressionsmetoden med färre förklaringsvariabler förklara responsvariabeln bättre. Skattad regressionsekvation med sedvanlig regressionsmodell ger en bättre anpassning enligt determinationskoefficienten R2, men tveksamt att acceptera de skattade koefficienterna till exempel variabler som har negativ tecken vid parameterskattningar enligt regressions ekvationen. PLS-proceduren är användbar om man håller ordning på användningen av antalet komponenter, speciellt vid modeller som innehåller litet antal observationer (30-60). ”Ridge” regression ger bättre resultat vid mindre antal förklaringsvariabler och med tillräckligt många observationer i data.

Om syftet är att belysa informationen från data, krävs det att man visualiserar materialet på ett förnuftigt sätt. Dagens mjukvaror ger dessa möjligheter och SAS 9.2 ger en bra möjlighet att studera kritiska läggen för att göra en korrekt val vid RREG och tolkning av resultat vid PLSREG (Alinaghizadeh 2008).

Det är svårt att välja antal komponenter vid PLSREG och PCREG när antalet förklaringsvariabler är för många. Men i detta arbete presenterades några metoder för att underlätta detta. Om man vill utgå från variationsvikter antigen W-x eller W-y presenterades ett resultat i tabell-1a och tabell-1b. Ett annat kriterium för att välja antal komponenter som ofta sker vid medicinska och epidemilogiska studier är att man vill lägga vikt på några grupperade variabler som brukar kallas för ”mediator”. Då väljer man komponenter efter variablernas betydelse i samma gruppering eller grupperingar som sker efter laddningar i komponenter. I ett annat fall använder man dessa faktorer för att standardisera data som kommer från ett slumpmässigt urval från en population (förslagsvis kan man använda dessa metoder inför SF-36). Den första komponenten är given och kan inte utelämnas. Därefter väljer man komponenter beroende på betydelsen av variablerna i en modell och deras vikt.

Ämnet PCREG och PLSREG är mycket större än det material som omfattades i denna uppsats. Vissa frågor diskuterades inte alls eller inte i detalj, liksom metoder för att identifiera extrema observationer, behandling av uppgifter som saknas, F- och t statistik, klassificering, "leverage", val av variabler, data transformationer, utvidgningar av flera block, hierarkiska modeller och brist på modell anpassning ”lack of fit”. Det finns också andra PLS algoritmer som kan ha vissa fördelar vid modell tillämpningar. Den slutliga slutsatsen är att vid PCREG eller PLSREG får alla variabler tillbaka sina riktiga tecken och värden för regressionskoefficienterna vilka i sin tur gör det lättare för forskare att tolka de skattade koefficienterna gentemot förklaring av responsvariabeln med hjälp av förklaringsvariabler.

Hassan Alinaghizadeh
Farhad.Alinaghizadeh@ki.se
2009-09-30
Stockholm - Sweden

Erkännande:

Appendix-1 Longley dataset och specifikations resultat

Labor Data Longley (1967) visar sig vara i dåligt kondition. Datasetet innehåller en beroende variabel, Employment (totalt sekundär sysselsättning) och sex oberoende variabler: Prices (GNP implicita prisdeflatorn med 1954 = 100), GNP (bruttonationalprodukt BNP), Jobless (arbetslöshet), Military (storleken av de väpnade styrkorna) PopSize (icke-institutionella befolkningen 14 år och äldre) och Year (år).

Longley, J. W. (1967), "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User," *Journal of the American Statistical Association*, 62, 819 - 41.

Variabler/observationer i datasetet Longley med predicerade och skattade residualer från PCREG.

	EMPLOYME	PRICES	GNP	JOBLESS	MILITARY	POPSIZE	YEAR	Predicted Value	Residual
1	60323	83	234289	2356	1590	107608	1947	60445,105	-122,1047
2	61122	89	259426	2325	1456	108632	1948	61800,011	-678,0106
3	60171	88	258054	3682	1616	109773	1949	60731,495	-560.4947
4	61187	90	284599	3351	1650	110929	1950	61832.759	-645.7593
5	63221	96	328975	2099	3099	112075	1951	63011.864	209.13556
6	63639	98	346999	1932	3594	113270	1952	63435.907	203.09324
7	64989	99	365385	1870	3547	115094	1953	64342.521	646.47902
8	63761	100	363112	3578	3350	116219	1954	63434.914	326.08557
9	66019	101	397469	2904	3048	117388	1955	65291.235	727.76533
10	67857	105	419180	2822	2857	118734	1956	66565.408	1291.592
11	68169	108	442769	2936	2798	120445	1957	67640.096	528.90424
12	66513	111	444546	4681	2637	121950	1958	66900.706	-387.7063
13	68655	113	482704	3813	2552	123366	1959	68848.094	-193.0944
14	69564	114	502601	3931	2514	125368	1960	69692.713	-128.7128
15	69331	116	518173	4806	2572	127852	1961	69758.707	-427.7068
16	70551	117	554894	4007	2827	130081	1962	71340.465	-789.4654

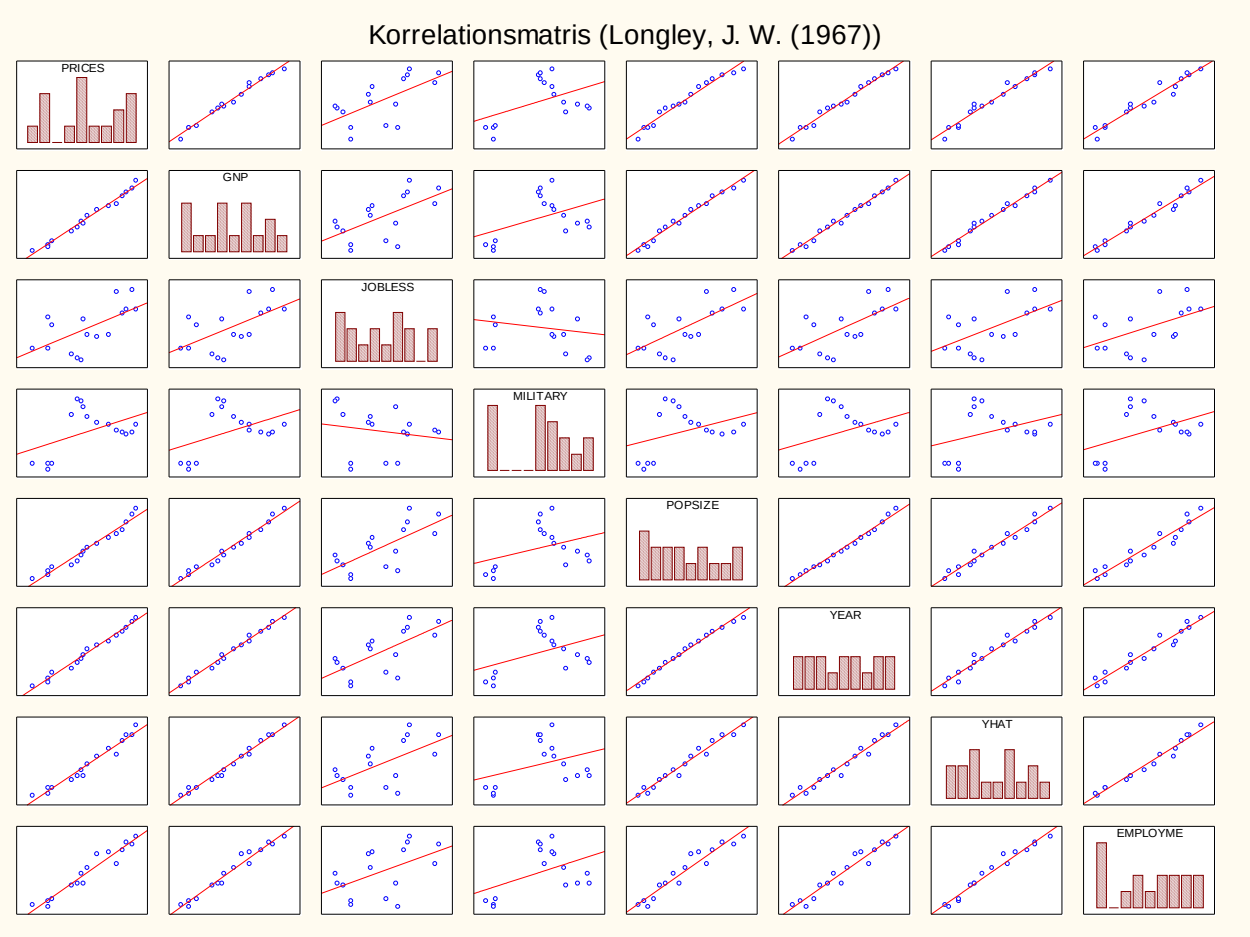
”Predicte Value” och ”Residual” är från PCREG med tre komponenter (PC1- PC3) med R2=97,13 %.

Observerade data från 1951-1954 har extrema låga värde för JOBLESS och höga för MILITARY

Deskriptivanalys och korrelations tabell Longley data N=16

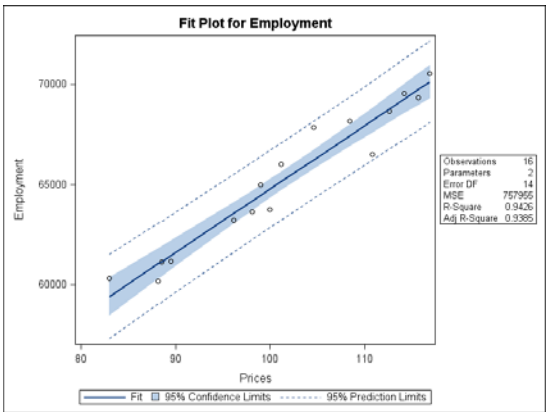
	PRICES	GNP	JOBLESS	MILITARY	POPSIZE	YEAR	EMPLOYME
PRICES	1,00000						
GNP	0,99159	1,00000					
JOBLESS	0,62063	0,60426	1,00000				
MILITARY	0,46474	0,44644	-0,17742	1,00000			
POPSIZE	0,97916	0,99109	0,68655	0,36442	1,00000		
YEAR	0,99115	0,99527	0,66826	0,41725	0,99395	1,00000	
EMPLOYME	0,97090	0,98355	0,50250	0,45731	0,96039	0,97133	1,00000
Means	101,68	387698,44	3193,31	2606,69	117424,00	1954,50	65317,00
Std.Dev.	10,79	99394,94	934,46	695,92	6956,10	4,76	3511,97

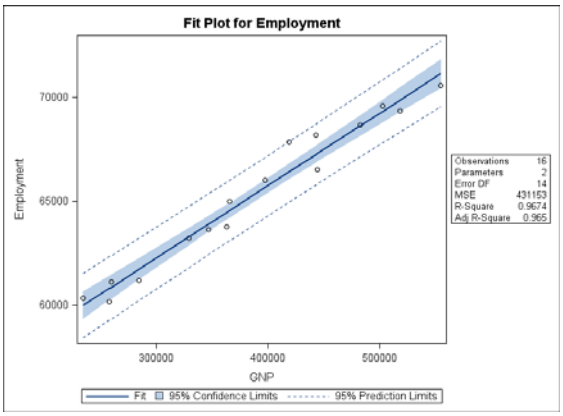
Korrelationsmatris för Longley data.

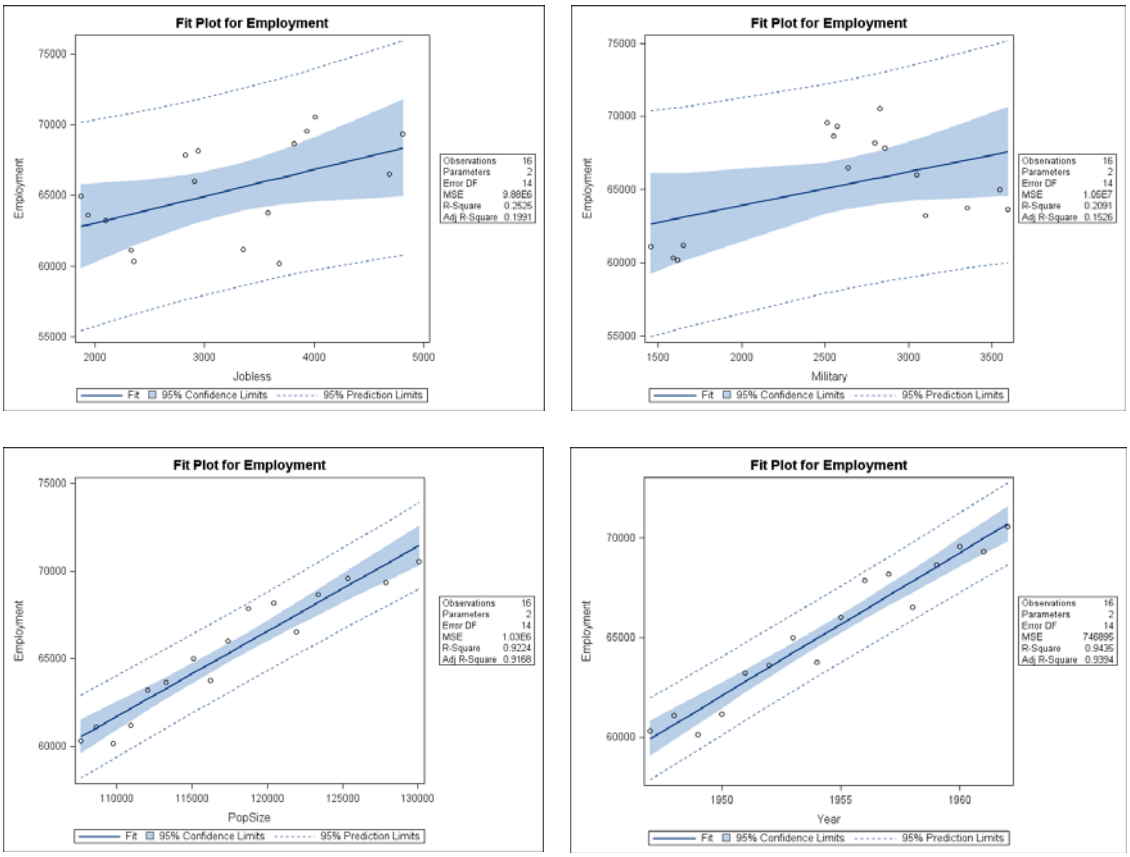


Variablerna X3 (JOBLESS) och X4 (MILITARY) har sämre korrelations förhållande med alla variabler i jämförd med de andra variabler. Korrelationen mellan dessa två variabler och responsvariabeln (EMPLOYMENT) är lite tveksam.

Enkel regression modeller med all förklaringsvariabler i Longley dataset.







Alla figurer tyder på ett starkt positivt samband mellan responsvariabeln (EMPLOYMENT) och förklaringsvariabler, förutom JOBLESS och MILITARY som har en bred spridning. Variabeln har en misstänksam positiv lutning mot responsvariabeln (EMPLOYMENT), dvs. tar vi bort extremt låga värden blir sambandet negativ. Däremot är dessa variabler högt korrelerade med de andra förklaringsvariabler som i sin tur skapar multikollinjäritet.

Egenvektor av korrelationsmatrisen och proportion av variansförklaring av skattade komponenter.

	Eigenvalue	% Total - variance	Cumulative - Eigenvalue	Cumulative - %
1	4,603377	76,72295	4,603377	76,7230
2	1,175340	19,58901	5,778718	96,3120
3	0,203425	3,39042	5,982143	99,7024
4	0,014928	0,24880	5,997071	99,9512
5	0,002552	0,04253	5,999623	99,9937
6	0,000377	0,00628	6,000000	100,0000

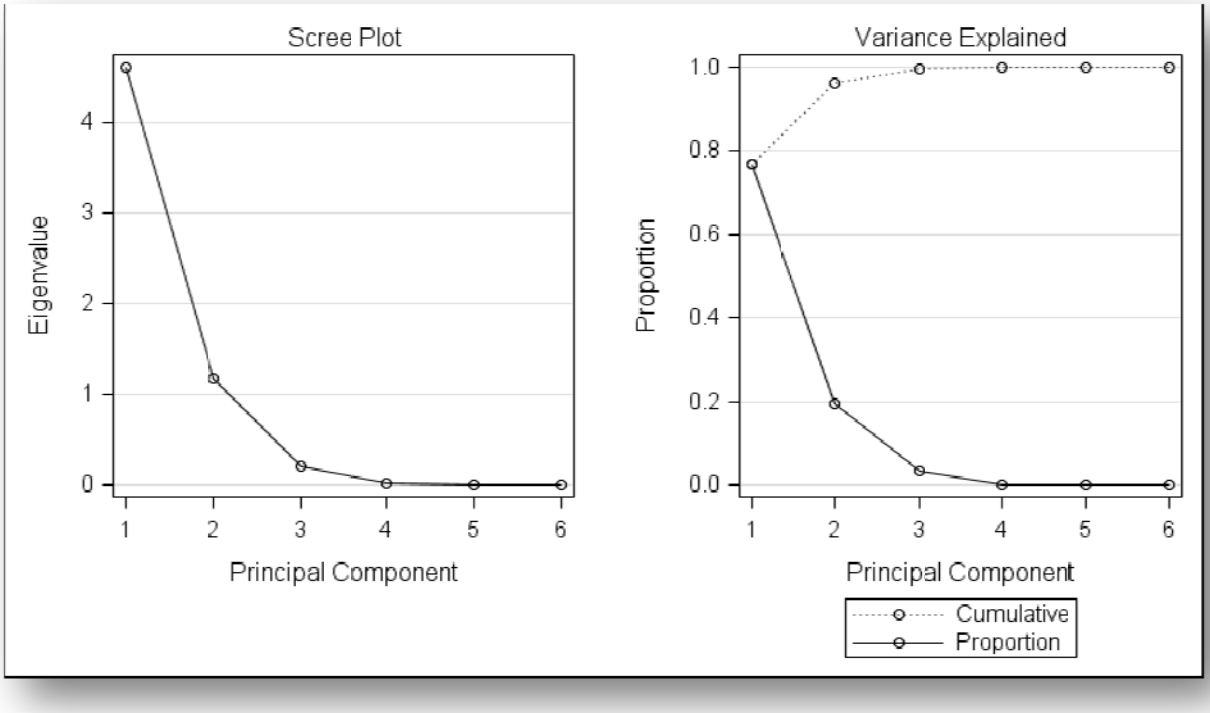
Skattade principalkomponents laddningar baserad på korrelationsmatris (Egenvektor)

	PC1	PC 2	PC 3	PC 4	PC 5	PC 6
PRICES	0,46	0,06	-0,15	-0,80	0,34	-0,14
GNP	0,46	0,05	-0,28	0,12	-0,15	0,82
JOBLESS	0,32	-0,60	0,73	-0,01	0,01	0,11
MILITARY	0,20	0,80	0,56	0,08	0,02	0,02
POPSIZE	0,46	-0,5	-0,20	0,60	0,55	-0,31
YEAR	0,46	0,00	-0,13	0,05	-0,75	-0,45

Den första PC har höga variationsförklaringar för PRICES, GNP, POPSIZE och YEAR och den tredje PC har dessa variationsförklaringar för JOBLESS och MILITARY.

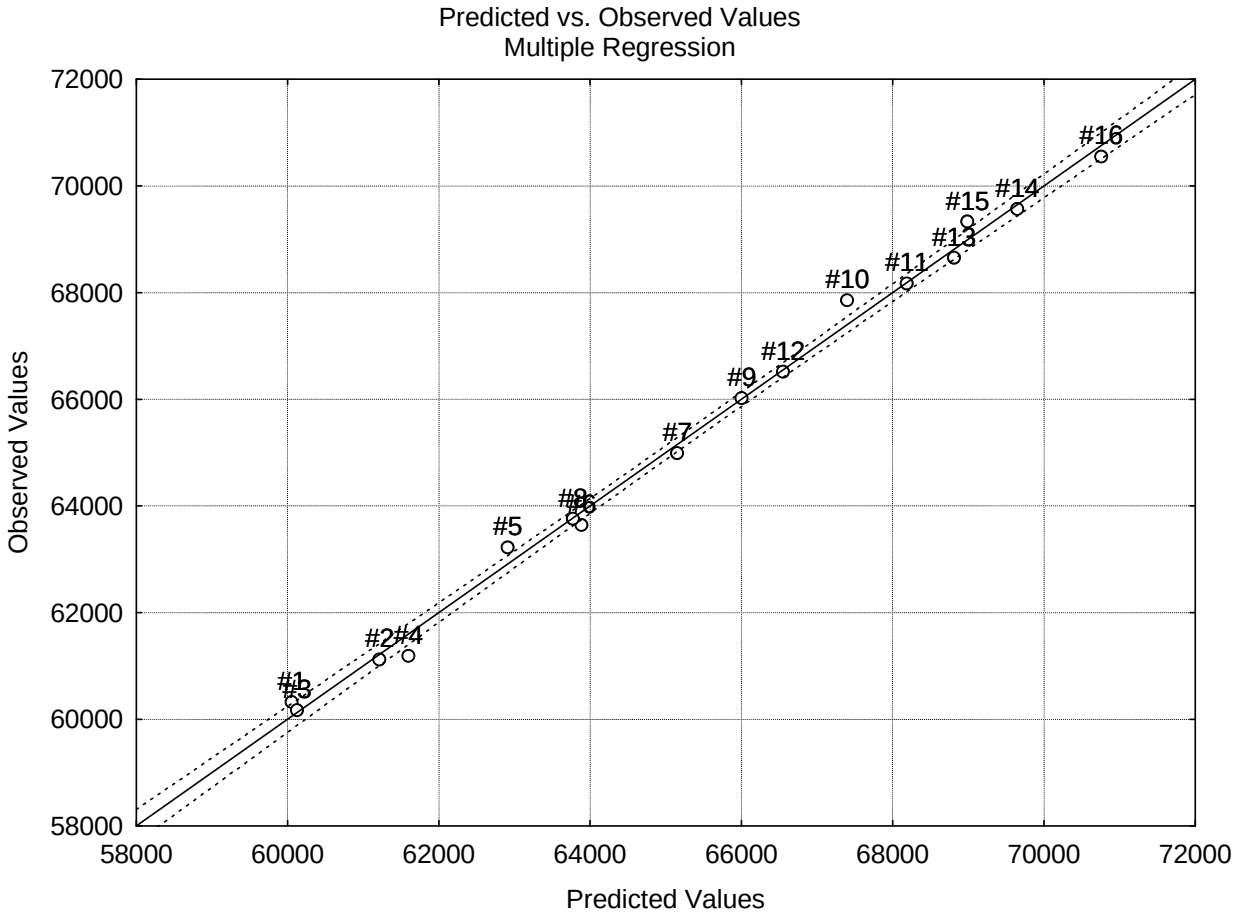
Dessa två komponenter blir valda i en stegviss regressionsmodell med EMPLOYMENT som responsvariabel. Vi kommer att acceptera de tre första skattade principalkomponenterna vid PCREG, men som har vi tidigare sett i tabell-2b ska det räcka att använda den första skattade komponenten.

”Scree plot” av egenvärdena och varians förklaring av principalkomponenter.



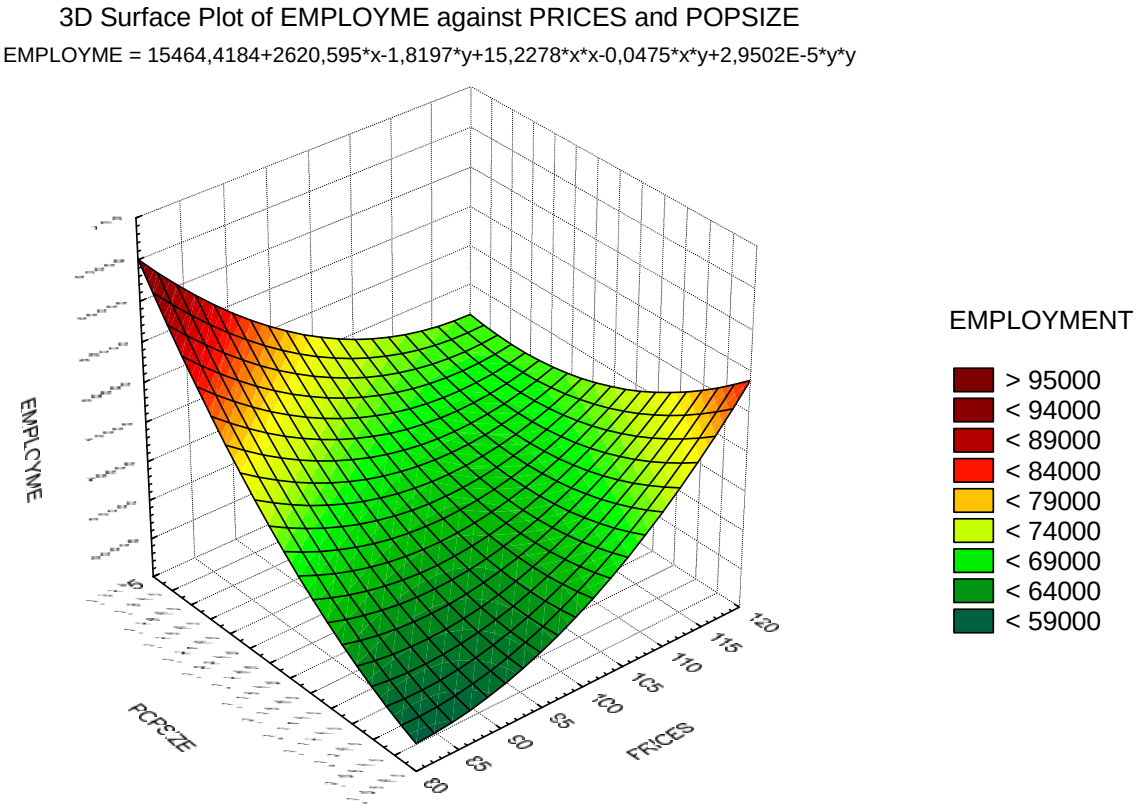
Figuren visar att den första principalkomponentens egenvärde och proportion av variationsförklaring av förklaringsvariabler. I detta fall med så få observationer förslår metoden att de tre första komponenterna kan förklara nästan 100% av variansen i data.

2 Dimension graf med observerade Y-värde mot skattade Y-hat från PCREG.



”Predicte Value” är från PCREG med tre komponenter (PC1 - PC3) med R2=97,13 % mot ”Observed Value” av Employment. PCREG har verkligen lyckat i detta fall att skatta alla värde så nära observerade värde som möjligt.

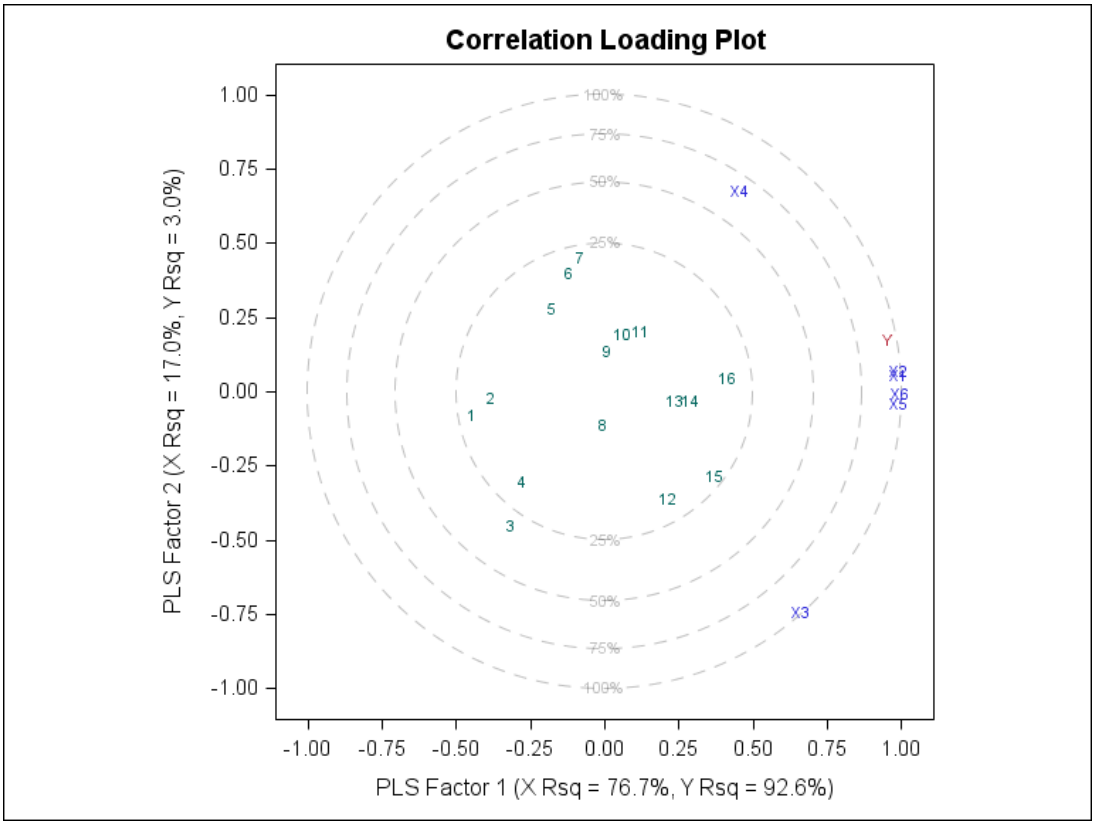
3Dimension diagram med PRICE och POPSIZE i relation till EMPLOYMENT.



Från ett ekonomiskt perspektiv (balansekonomi inom Nationalekonomi); tolkning av denna figur säger att ökning av priser och populationsstorlek ökar anställningen (detta samband är sinsemellan). Den centrala värdet eller balansen ligger i den gröna område (EMPLOYMENT 74000-69000).

Bilden kan lätt tolkas ur en nationalekonomisk synvinkel vad gäller de centrala begrepp inom ämnet för arbetslöshet och inflation i samband med den naturliga ökningen av population.

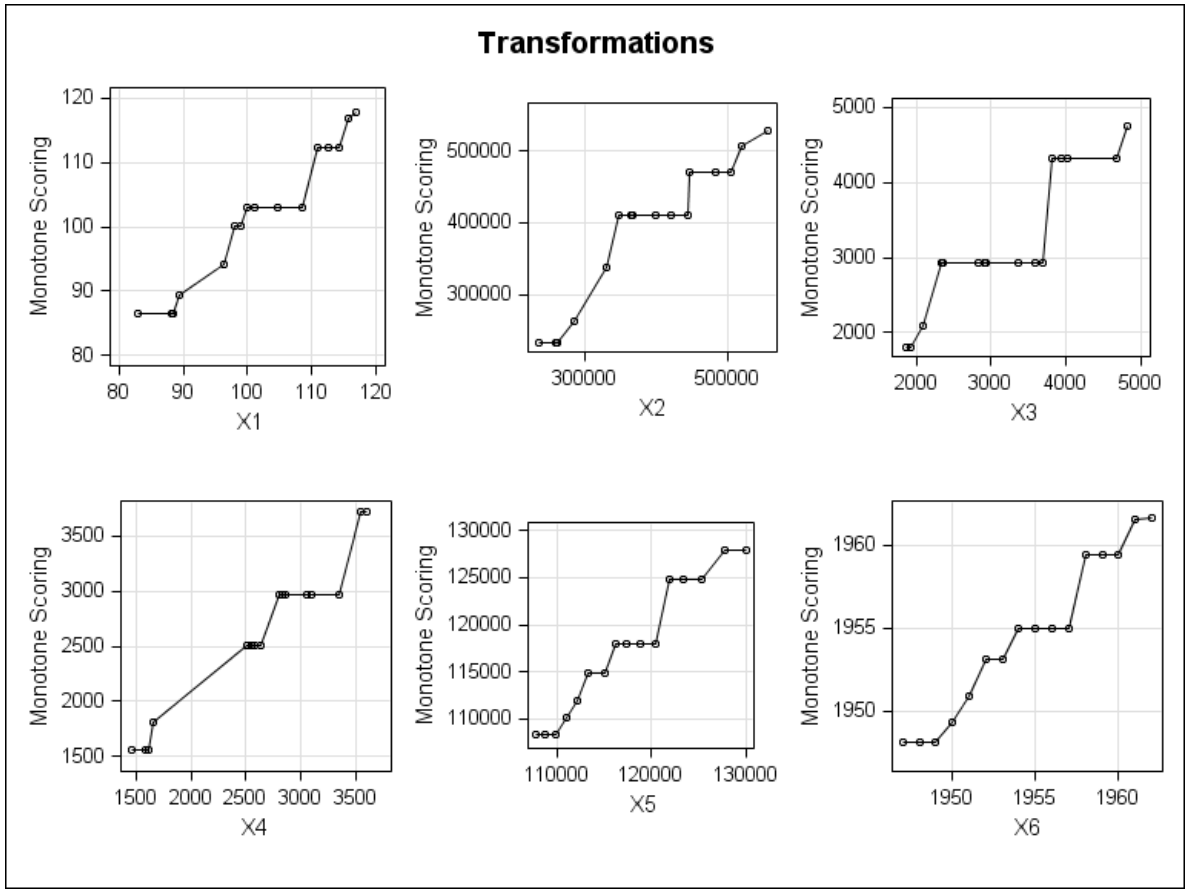
”Korrelations laddning graf” vid PLSREG



Korrelationsladdningsfigur är en kompakt sammanfattning av valmöjligheter som man kan utföra vid PLSREG. Här kan vi avläsa att första komponent (PLS Factor = PC1) är starkt positivt korrelerad med alla variabler och att den andra komponent har lägsta värde av alla variabler förutom X3 som har ett negativt samband och X4 som har positivt samband. Observationerna ligger ganska spridda (vilka presenteras med deras id nummer) och tyder på att data ger ganska bra information om de två första komponenterna.

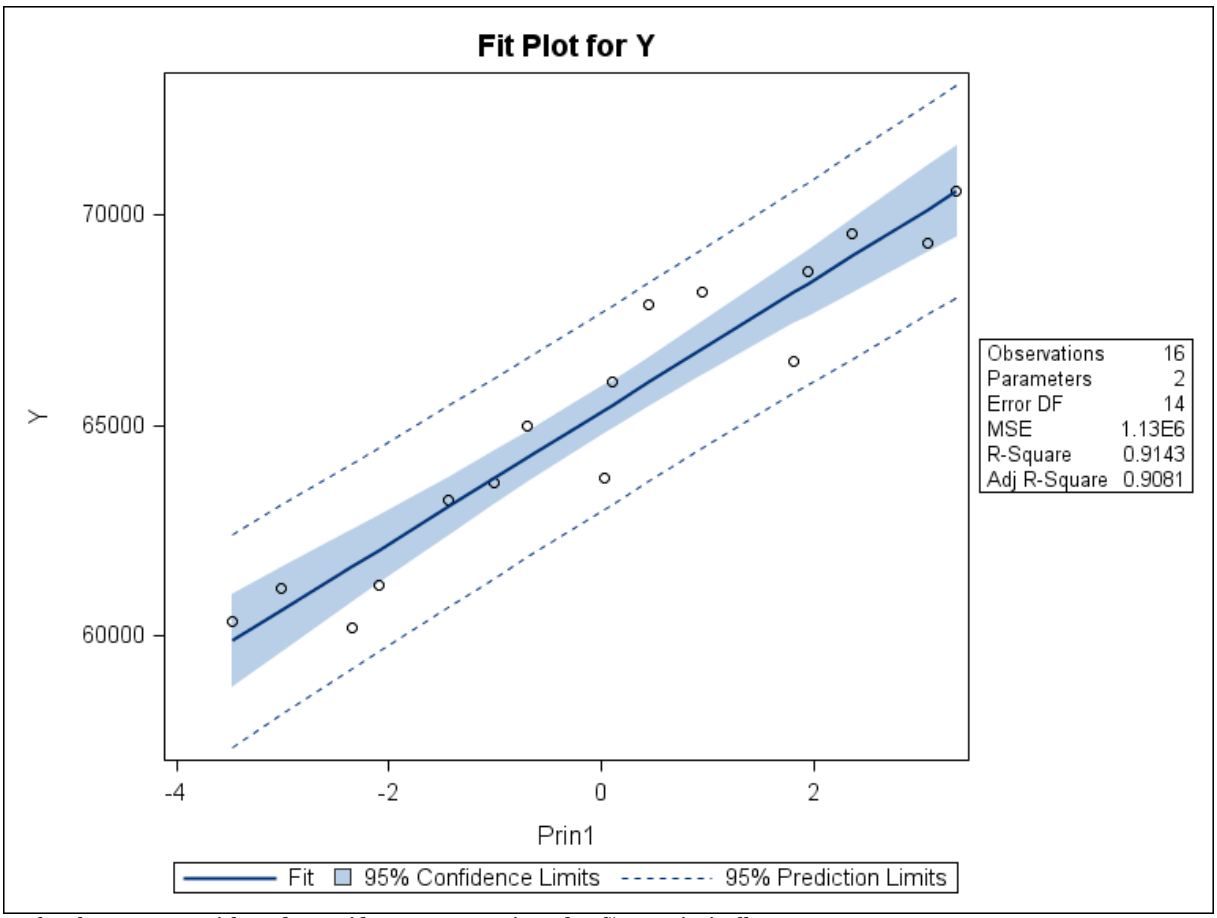
En kombination av dessa två faktorer ger största variationsförklaring från data. Om man utgår från figuren ovan den första PLS faktorn har den största variationsförklaringen för Y medan den andra faktorn ger större variationsförklaring för X. Denna karaktär hos PLSREG har vi redan diskuterat, men ju mer faktorer i modellskattningarna desto jämnare variationsförklaring för både Y och X-variabler.

Transformations grafer inför PLREG.



Inför PLSREG är nästan alla variabler vältransformerade. X3 har flera värden som ligger under och över 3000 som är transformerade till 3000 gentemot EPLOYMENT. För mer förklaring om grafen se SAS dokumentation. Om variablerna verkligen hade en bra korrelation med responsvariabeln då skulle dessa linjer ha en rak lutning uppåt.

Skattade enkel regression modell med den första principalkomponenten.



Grafen skapas automatisk med SAS vid PROC REG. Prin1= den första principalkomponent

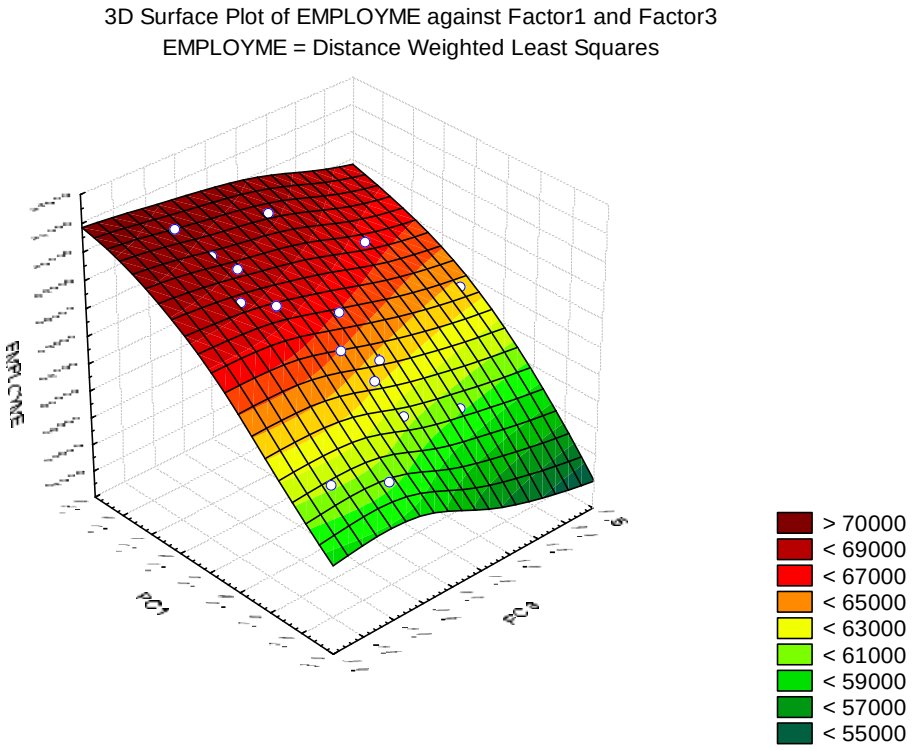
SAS output för enkel regression med den första principalkomponenten (Prin1). Den förts principalkomponenten kan förklara vår responsvariabel med ungefär 91% determinationskoefficient.

Resultat för stegvis regression med principalkomponenter

Step	Variable Entered	Partial R-Square	Model R-Square	C(p)	F Value	**P-värdet**
1	PC-1	**0,9143**	**0,9143**	**158,697**	**149,27**	**<,0001**
2	PC-3	**0,0571**	**0,9713**	**47,0604**	**25,89**	**0,0002**
3	PC-2	**0,0146**	**0,9860**	**19,9357**	**12,51**	**0,0041**
4	PC-5	**0,0079**	**0,9938**	**6,2558**	**14,07**	**0,0032**
5	PC-6	**0,0015**	**0,9953**	**5,3076**	**3,17**	**0,1055**

Stegvis regression med skattade principalkomponenter visar att den fjärde komponent kan utelämnas i en regression samband och den tredje komponent väljas som andra förklarings variabel i en multipel modell.

3D diagram med två komponenter i relation till EMPLOYMENT.



Figuren visar att den första PC är väl korrelerad mot responsvariabeln EMPLOYMENT. Figuren är skapad med PC1 och PC3, där PC3 har är mindre korrelerad med EPLOYMENT. Ju större negativa värden för båda PC1 och PC3 desto högre värde för EMPLOYMENT. Grafen visar också sambandet mellan PC1 och PC3 i en modell gentemot EMPLOYMENT och att problemet med multikollinjäritet är löst.

Appendix-2 The SIMPLS algorithm

de Jong, S. (1993), "**SIMPLS: An Alternative Approach to Partial Least Squares Regression**" Chemometrics and Intelligent Laboratory Systems, 18, 251–263.

Input:
n * p matrix X
n * m matrix Y
Number of factor A

$Y_o = Y - Mean(Y)$	center Y
$S = X' * Y_o$	cross-product
For a = 1, …,A	per dimension
$q = dominant\ eigenvector\ of\ S'S$	Y block factor weights
$r = S * q$	X block factor weights
$t = X * r$	X block factor scores
$t = t - mena(t)$	center scores
$norm\ t = sqrt\ (t' * t)$	compute norm
$t = t / norm\ t$	normalize scores
$r = r / norm\ t$	adapt weights accordingly
$p = X' * t$	X block factor loadings
$q = Y'_o * t$	Y block factor loadings
$u = Y_o * q$	Y block factor scores
$v = p$	initialize orthogonal loadings
if a > 1 then	
$v = v - V\ (V' * p)$	make v to previous loading
$u = u - T\ (T' * u)$	make u to previous t´ values
$v = v / sqrt(v' * v)$	normalize orthogonal loadings
$S = S - v*(v'*S)$	deflate S with respect to current loadnings

Store r, t, p, q, u and v into R, T, P, Q, U and V

$B = R * Q'$	regression coefficients
$h = DIAG(T * T') + 1/n$	leverages of objects
$varX = DIAG\ (P' * P) / (n-1)$	variance explained for X variables
$varY = DIAG\ (Q'*Q) / (n-1)$	

Appendix-3 The PLS algorithm

Geladi, P. and Kowalski, B. (1986)," **Partial Least Tutorial**" Analystical Chimica Acta, 185, 1-17.

It is assumed that X and Y are mean centered and scaled:

For each component:
(1) take u_{start} = some y_j.

In the X block:
(2) $w' = u'X/u'u$
(3) $w'_{new} = w'_{old} / ||w'_{old}||$ (normalization)
(4) $t = Xw/w'w$

In the Y block:
(5) $q' = t'Y/t't$
(6) $q'_{new} = q'_{old} / ||q'_{old}||$ (normalization)
(7) $u = Y\,q / q'q$

Check convergence:
(8) Compare the t in step 4 with the one from the preceding iteration. If they are equal (within a certain rounding error) go to step 9, else go to step 2. (if the Y block has only one variable, step 5-8 can be omitted by putting $q = 1$, and no more iteration is necessary).

Calculate the X loading and rescale the scores and weights accordingly:

(9) $p' = t'X/ t't$
(10) $p'_{new} = p'_{old} / || p'_{old} ||$ (normalization)
(11) $t_{new} = t_{old} ||p'_{old}||$
(12) $w'_{new} = w'_{old} || p`_{old} ||$
(p', q' and w' should be saved for prediction; t and u can be saved for diagnostic and / or classification purposes).

Find the regression coefficient b for the inner relation:
(13) $b = u't/t't$

Calculation of the residuals. The general outer relation for the X block (for component h) is:
(14) $E_h = E_{h-1} - t_h p'_h$; $X=E_o$

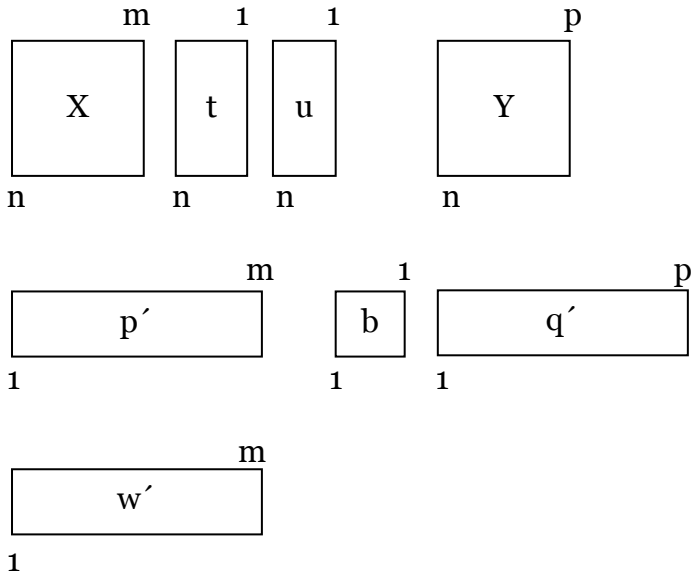
The mixed relation for the Y block (for component h) is:

(15) $F_h = F_{h-1} - b_h t_h q'_h$; $Y = F_o$

From here, one goes to step 1 to implement the procedure for the next component.

(Note: After the first component, X in step 2, 4 and 9 and Y in steps 5 and 7 are replaced by their corresponding residual matrices E_h and F_h).

Matrices and vector are shown graphically in Figure form



Referenser:

1. Alinaghizadeh Farhad H. (2008), "Multivariate analysis by SAS", compendium (3 days course), **SAS institute Stockholm,** Sweden.
2. Alinaghizadeh Farhad H. (2009), "En analys av förväntad medellivslängd i världens länder 1996–97", uppsats**, Lund Universitet,** Ekonomihögskolan, Statistiska Institutionen.
3. Allen D. M. (1971), the prediction sum of squares as a criterion for selection of predictor variables". Technical report 23, Department of statistics, **University of Kentucky**
4. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), "*Regression Diagnostics"*, New York: **John Wiley & Sons**, Inc.
5. de Jong, S. (1993), "SIMPLS: An Alternative Approach to Partial Least Squares Regression", **Chemometrics and Intelligent Laboratory Systems**, 18, 251–263.
6. de Jong, S. and Kiers, H. (1992), "Principal Covariates Regression", **Chemometrics and Intelligent Laboratory Systems**, 14, 155–164.
7. de Jong, S. (1993), "SIMPLS: An Alternative Approach to Partial Least Squares Regression", **Chemometrics and Intelligent Laboratory Systems**, 18, 251–263.
8. Dijkstra, T. (1983), "Some Comments on Maximum Likelihood and Partial Least Squares Methods", **Journal of Econometrics**, 22, 67–90.
9. Dijkstra, T. (1985), "Latent Variables in Linear Stochastic Models: Reflections on Maximum Likelihood and Partial Least Squares Methods", Second Edition, Amsterdam, The Netherlands: **Sociometric Research Foundation**.
10. Efron B., Tibshirani R. J. (1993), "An introduction to the bootstap", Monographs on statistics and applied probability 57, **Chapman & Hall** /CRC
11. Frank, I. and Friedman, J. (1993), "A Statistical View of Some Chemometrics Regression Tools", **Technometrics**, 35, 109–135.
12. Friedman J. H. and Popescu B. E. 2004, **Technical Report**, Statistics Department, Standford
13. Gabriel, K. R. (1971), "The biplot graphic display of matrics with application to principal component analysis", **Biometrika** 58:453-467.
14. 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**.
15. Geladi, P. and Kowalski, B. (1986), "Partial Least-Squares Regression: A Tutorial", **Analytica Chimica Acta**, 185, 1–17.
16. Gunst R. F. (1983), "Regression Analysis with Multicollinear Predictor Variables: Definition, Detection, and Effects", **Communication is Statistics - Theory and Mathods**, 12, 2217 -2260.
17. Gunst R. F. (1984), "Comment: Toward a balanced assessment of Collinearity diagnostics", **The American Statistician**, 38, 79-82
18. Gunst R. F., Mason R. L. (1980), "Regression Analysis and Its Application", **New York, Marcel Dekker, Inc**
19. Hawkins, S (1973), "On the investigation of alternative regression by principal component analysis", **Applied statistics**, 22, 275-286.
20. Haykin, S. (1994), Neural Networks: "A Comprehensive Foundation", New York: Macmillan.
21. Helland, I. (1988), "On the Structure of Partial Least Squares Regression", Communications in Statistics, **Simulation and Computation**, 17, 581–607.
22. Hocking, Speed and Lynn (1976), "A class of biased estimators in linear regression", **Technometrics**, 18, 425-437.
23. Hoerl and Kennard (1970), "Ridge regression: Biased estimation for nonorthogonal problem", **Technometrics**, 12, 55-67
24. Hoerl, Kennard and Baldwin (1975), "Ridge regression: Some simulations", **Communications in statistics**, 4, 105-124
25. Hotelling H. (1957), "Relation of the newer multivariate statistical methods to factor analysis", **British Journal of statistical psychology**, 10, 69-79
26. Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables into Principal Components," **Journal of Educational Psychology**, 24, 498 -520.
27. Jackson, J.E. (1991). "A User's Guide to Principal Components". **Wiley, New York**.
28. Jeffers J. (1967), "Two case studies in the application of principal component", **Applied statistics**, 16, 225-236
29. John, R.C. St. (1984). "*Experiments With Mixtures in Conditioning and Ridge Regression",* **Journal of Quality Technology** 16, pp.81-96.
30. Jolliffe I. T. (1982). "A note on the Use of Principal Components in Regression". **Journal of the Royal Statistical Society, Series C (Applied Statistics) 31** (3): 300–303.
31. Kendall M. G. (1957), "Studies in the history of probability and statistics", V. A note on playing cards, **Biometrika**, 44, 260-262
32. Lindberg, W., Persson, J.-A., and Wold, S. (1983), "Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate", **Analytical Chemistry**, 55, 643–648.
33. Lott W. F.,(1973), "The optimal set of principal component restriction on a least squares regression", **Communications in Statistics**, 2, 449-464
34. Mardia, K.V., Kent, J.T., Bibby, J.M. (1979), "Multivariate Analysis", **Academic Press, London**, p231.

35. Marquardt (1970), “Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation”, **Technometrics**, 12, 591-612
36. Marquart and Snee (1975), “Ridge regression in practices”, **The American Statistician**, 29, 3-19
37. McAvoy, T. J., Wang, N. S., Naidu, S., Bhat, N., Gunter, J., and Simmons, M. (1989), “Interpreting Biosensor Data via Backpropagation”, **International Joint Conference on Neural Networks**, 1, 227–233.
38. Mosteller R. and Tukey J. W. (1977), “Data Analysis and Regression: A second Course in Statistics", **Addison- Wesley**, Reading, Massachusetts
39. Naes, T. and Martens, H. (1985), “Comparison of Prediction Methods for Multicollinear Data”, **Communications in Statistics, Simulation and Computation**, 14, 545–576.
40. Park S. H.(1981), “Collinearity and optimal restrictions on regression parameters for estimating responses”, **Technometrics**, 23, 289-295
41. Ränner, S., Lindgren, F., Geladi, P., and Wold, S. (1994), “A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects”, **Journal of Chemometrics**, 8, 111–125.
42. Rawlings John O.(2002), “Applied regression analysis, A research tool”, **Wadsworth & Brooks/Cole**, Statistics/Probability series
43. Richard M. Heiberger, Burt Holland (2004). “Statistical analysis and data display” , **Springer**
44. Sarle, W. S. (1994), “Neural Networks and Statistical Models”, in Proceedings of the **Nineteenth Annual SAS Users Group International Conference**, Cary, NC: SAS Institute Inc.
45. Shao, J. (1993), “Linear Model Selection by Cross-Validation”, Journal of the **American Statistical Association**. 88, 486-494
46. Smith G., Campbell F.(1980), “A critique of some ridge regression methods”, **Journal of the American Statistical Association**, 75, 74-81
47. Stein C. M (1960), “Multiple regression. In Contributions to Probability and Statistics, Essays in Honor of Harold Hoteling”, **Stanford University Press**, Stanford, California
48. 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.
49. Tobias, R. (1995), “An Introduction to Partial Least Squares Regression”, in Proceedings of the **Twentieth Annual SAS Users Group International Conference**, Cary, NC: SAS Institute Inc. 1250-1257
50. Ufkes, J. G. R., Visser, B. J., Heuver, G., and Van Der Meer, C. (1978), “Structure-Activity Relationships of Bradykinin-Potentiating Peptides”, **European Journal of Pharmacology**, 50, 119.
51. Ufkes, J. G. R., Visser, B. J., Heuver, G., Wynne, H. J., and Van Der Meer, C. (1982), “Further Studies on the Structure-Activity Relationships of Bradykinin-Potentiating Peptides”, **European Journal of Pharmacology,** 79, 155.
52. van den Wollenberg, A. L. (1977), “Redundancy Analysis—An Alternative to Canonical Correlation Analysis”, **Psychometrika**, 42, 207–219.
53. van der Voet, H. (1994), “Comparing the Predictive Accuracy of Models Using a Simple Randomization Test”, **Chemometrics and Intelligent Laboratory Systems**, 25, 313–323.
54. Van Norstrand, R. C. (1980), “Comment: A critique of some ridge regression methods”, **Journal of the American Statistical Association**, 75, 92-94
55. Webster J. T, Gunst R. F, Mason R. L. (1974), “Latent root regression analysis”, **Technometrics**, 16, 513-522
56. Wold, H. (1966), “Estimation of Principal Components and Related Models by Iterative Least Squares”, in P.R. Krishnaiah, ed., Multivariate Analysis, New York: **Academic Press**.
57. Wold, S. (1994), “PLS for Multivariate Linear Modeling,” QSAR: **Chemometric Methods in Molecular Design. Methods and Principles in Medicinal Chemistry**. (Ed. H. van de Waterbeemd), Weinheim, Germany: Verlag-Chemie
58. Younger M. S. (1979), “*A Handbook for Linear Regression”*, USA: DUXBURY Press.

SAS References:

1. SAS/IML Software: Usage and Reference (1999), version 8. Cary, North Carolina, 1st edition, SAS Instituted Inc.
2. SAS/IML 9.1, User's Guide (2004), Volumes 1 and 2, Cary, North Carolina, SAS Instituted Inc.
3. SAS/IML 9.2, User's Guide (2008), Volumes 1 and 2, Cary, North Carolina, SAS Instituted Inc.
4. SAS/STAT 9.1, User's Guide (2004). SAS Publishing, Cary, North Carolina, 1st edition, SAS Instituted Inc.
5. SAS 9.1.3 Output Delivery System (2006), User's Guide, Volumes 1 and 2, Cary, North Carolina, SAS Instituted Inc.
6. Base SAS 9.1.3 Procedures Guide (2006), Second Edition, Volumes 1-4, Cary, North Carolina, SAS Instituted Inc.
7. Base SAS 9.2 Procedures Guide (2009), Statistical procedures, Second Edition, Cary, North Carolina, SAS Instituted Inc.
8. SAS/GRAPH 9.2: ODS Graphics Editor User's Guide (2009), Cary, North Carolina, SAS Instituted Inc.
9. SAS/GRAPH 9.2: Statistical Graphics Procedures Guide (2009), Cary, North Carolina, SAS Instituted Inc.
10. SAS/GRAPH 9.2: Reference (2009), Volumes 3, Cary, North Carolina, SAS Instituted Inc.