

JTI-rapport

Lantbruk & Industri

359

Transferkalibrering med ridge regression

Lennart Norell
Mikael Gilbertsson



JTI - Institutet för jordbruks- och miljöteknik

2007

Transferkalibrering med ridge regression

Lennart Norell¹
Mikael Gilbertsson²

¹ Enheten för tillämpad statistik och matematik, SLU

² JTI – Institutet för jordbruks- och miljöteknik

Innehåll

Förord.....	5
Sammanfattning.....	7
Bakgrund.....	7
Material och metoder.....	8
Tillgängliga data.....	8
Preliminär statistisk analys.....	8
Multikollinearitet.....	11
Ridge regression.....	14
Resultat.....	18
USA-instrumenten – Ett masterinstrument för kalibrering av flera slavinstrument.....	18
USA-instrumenten – Separata kalibreringar.....	19
Svenska instrumenten.....	20
Diskussion.....	23
Litteratur.....	24

Förord

Omkring år 2000 utvecklade JTI – Institutet för jordbruks- och miljöteknik ett NIT-instrument för skördetröskor. Instrumentet används tillsammans med en GPS för att samla in positionsbestämd data om spannmålens protein- och vattenhalt. Ganska snart insåg man att variationerna är stora över en åker och att det skulle vara lönsamt att sortera ut olika fraktioner. Frågan var i vilket led i kedjan som en sortering skulle ske. Under de senaste åren har JTI arbetat med olika samarbetspartners med on-line sortering av spannmål. Att analysera prover on-line har vi antagligen bara sett början på. Ett ökat värde på spannmålen, ökade krav på spårbarhet, ett ökat behov av att veta vad som finns i lager samt en ökad efterfrågan på olika kvaliteter kommer med all säkerhet att öka efterfrågan på on-line analysering och sortering.

Med ett ökat antal instrument så ökar behovet av att slippa kalibrera varje instrument för sig eftersom detta är tids- och kostnadskrävande. I denna rapport redovisas möjligheten till så kallad transferkalibrering av Zeltex NIT-instrument.

Projektet har genomförts som ett samarbete mellan SLU och JTI, med en projektgrupp bestående av Lennart Norell, Enheten för tillämpad statistik och matematik vid SLU och Mikael Gilbertsson, JTI. Mikael Gilbertsson har varit projektledare och ansvarat för datainsamling och teknik. Lennart Norell har gjort databehandlingar och statistiska analyser.

Medel för projektets genomförande har erhållits från Stiftelsen Lantbruksforskning (SLF). Zeltex har bidragit med viktig information och data om instrumenten och Lantmännen har bidragit med data samt varit en viktig diskussionspartner i projektet. Ett stort tack till alla som bidragit till projektets genomförande.

Uppsala i oktober 2007

Lennart Nelson

VD för JTI – Institutet för jordbruks och miljöteknik

Sammanfattning

I huvuddrag kan resultaten från projektet sammanfattas enligt följande:

- Att använda ridge regression för att skatta transferfunktioner är lovande. Jämfört med ordinär multipel regression blir skattningarna mer robusta. Detta illustreras bl a i figur 3a vs 8b samt i figur 14 där det framgår att skattningarna baserade på ridge regression har bättre precision.
- Att helt låta ett masterinstrument styra kalibreringar för ett antal slavinstrument verkar inte ge ett bra resultat. Varje slavinstrument bör få genomgå en separat nivåkalibrering. Om detta inte görs kan effekter som i figur 13 uppträda. Ridge regression förbättrar visserligen precisionen men det systematiska felet kvarstår, se figur 13b. I figur 14 har nivån kalibrerats.
- I andra sammanhang än detta projekt har PLS (Partial Least Squares) använts för transferkalibrering. Ridge regression är ett mellanting av denna teknik och ordinär multipel regression. Eftersom antalet prover är större än antalet uppmätta våglängder kan fördelarna vägas samman, vilket inte är möjligt då antalet våglängder överstiger antalet prover så att multipel regression inte kan användas.
- Valet av parametern q i uttrycket (15) kan inverka på resultatet. Det finns inget patentsvar utan här har $q = 1/2$ fungerat väl, men med mer informationsunderlag kanske detta val kan förbättras.
- Vid de transferkalibreringar som för närvarande utförs med ett referensinstrument utgår man ifrån att resultaten har en standardavvikelse 0.4 %-enheter för proteinhalter hos vete. Detta svarar mot den rent slumpmässiga variationen hos ett individuellt värde. Utöver denna osäkerhet tillkommer den som är kopplad till sambandet gentemot uppmätta temperaturer och transmittanser hos de 14 våglängderna. I de flesta fall är den senare osäkerheten mindre, men då temperaturerna och transmittanserna är olika jämfört med kalibreringsdata kan den dominera. Ridge regression innehåller möjligheter till att skapa en skattning med en mindre osäkerhet av detta slag.
- De numeriska beräkningarna för projektet har gjorts med en matrisherande procedur (Proc IML) i statistikpaketet SAS (Statistical Analysis System). De använda funktionerna är av standardtyp och programmering i ett annat datorspråk som inkluderar matrisberäkningar ska inte behöva vålla några väsentliga svårigheter.

Bakgrund

För att uppskatta protein- och vattenhalter i spannmål finns transmittansinstrument av filtertyp, vilka är förhållandevis billiga i inköp. För att mätresultaten ska bli tillförlitliga behöver instrumenten kalibreras vilket görs genom att prover som analyserats med en dyrare referensmetod även används för transmittanssensorerna. Kostnaden för denna kalibrering är ganska stor. Karaktären hos våglängdsdata leder också till statistiska svårigheter med effekten att skattningarna är mycket känsliga för avvikelser från kalibreringsdata. Målet med projektet är att skapa transferfunktioner för kalibrering av

denna typ av instrument. För detta studeras statistiska metoder som syftar till en robust kalibrering i meningen att den ska fungera väl även för spannmålsprover med måttliga avvikelser från kalibreringsområdet.

Material och metoder

Tillgängliga data

För utvärdering av statistiska metoder finns referensdata från 24 Zeltexinstrument i USA och 2 i Sverige, ett hos Lantmännen och ett vid JTI. De värden som registreras är transmittansen för 14 våglängder, i form av det negativa värdet av logaritmen, samt 2 temperaturer, en för provet och en för instrumentet.

Varje USA-instrument har testats med ca 200 referensprover. Totalt finns ca 300 prover som kan delas in i tre nästan lika stora grupper. En av grupperna används till alla 24 instrument, en annan grupp till enbart instrumenten 1 till 5 och den tredje till instrumenten 6 till 24. Inom den första gruppen av instrument har nr 2 något färre prover gemensamma med de övriga, 160 i stället för drygt 180. Mätningar finns gjorda i temperaturintervallet från 0 till 46 °C.

För de svenska instrumenten finns för Lantmännens ca 120 prover från Mälarvetetävlingen år 2005, ca 120 prover av annat vetematerial insamlat av Lantmännen år 2005, 50 prover från 2006 samt ca 200 amerikanska referensprover. Alla prover har avlästs i temperaturintervallet ca 18-25 °C. Dessutom har de från Mälarvetetävlingen och Lantmännen för 2005 avlästs för lägre temperaturer, ca 1-10 °C. För instrumentet vid JTI har enbart provserierna från 2005 använts, men i gengäld har mätningar gjorts i såväl rums- som lägre temperatur.

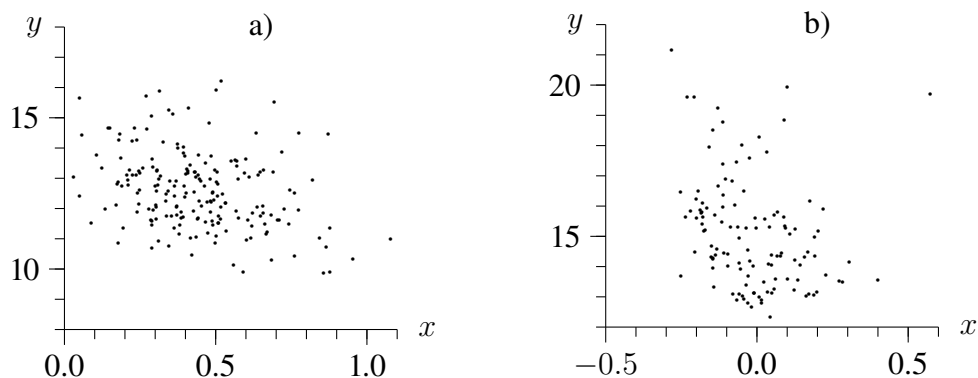
Preliminär statistisk analys

Som modell för hur en svarsvariabel (t ex protein- eller vattenhalt) beror av transmittansvärden och temperaturer görs först en ansats med multipel regressionsanalys på formen

$$y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{14} x_{i,14} + \beta_{15} x_{i,15} + \beta_{16} x_{i,16} + \varepsilon_i, \quad i = 1, \dots, n$$

där n anger antalet olika prover som används för kalibreringen. För prov nr i anger y_i uppmätt värde av protein- eller vattenhalt med ett referensinstrument (Foss Infratec), $x_{i,1}, \dots, x_{i,14}$ är uppmätta transmittanser för de 14 våglängderna i Zeltexinstrumentet, $x_{i,15}, x_{i,16}$ anger temperaturer för instrument respektive prov, samt ε_i betecknar den rent slumpmässiga avvikelsen för y_i från den teoretiskt förväntade givet aktuella värden på x -variablerna. En vanlig förutsättning för olika ε_i är att de är statistiskt oberoende och att de var för sig följer en normalfördelning med väntevärdet 0 och variansen σ^2 ; för korthets skull skrivs detta $\varepsilon_i \sim N(0, \sigma^2)$. Koefficienterna $\beta_0, \beta_1, \dots, \beta_{16}$ är okända och skattas ur en observationsserie. För att även σ^2 ska kunna skattas är i princip 18 observationer tillräckligt men resultatet blir inte tillförlitligt. För de datamängder som är aktuella här är antalet observationer, 50 till ca 200, betydligt fler per serie. Å andra sidan är det inte enbart antalet observationer som inverkar på precisionen i skattningarna av β -parametrarna. Mer om detta tas upp i avsnittet om multikollinearitet.

Det är inte möjligt att plotta sambandet mellan y och alla x -variabler samtidigt. En serie av plottar av y mot en x -variabel i taget visar på linjära samband. Figur 1 visar resultat för USA-instrument nr 1 och Lantmännens instrument och data från 2005.



Figur 1. Linjära samband mellan a) y =proteinhalt och x =transmittans vid våglängden 990 i Zeltex instrument 1 och b) y =vattenhalt och x =transmittans vid våglängden 1045 i Lantmännens instrument.

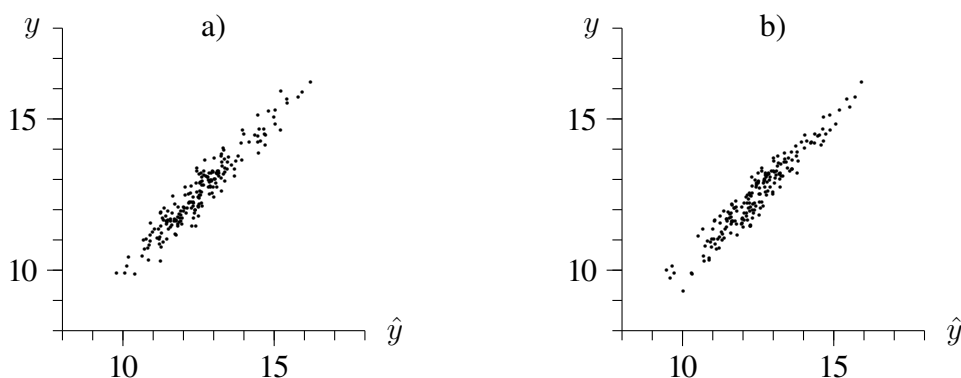
Sambanden är inte starka, värdena på R^2 (den andel av variationen hos y som förklaras av x -variabeln) är 0.10 respektive 0.06. Om det avvikande värdet i övre hörnet till höger av figur 1b utesluts blir $R^2 = 0.14$.

Däremot blir sambanden betydligt starkare om alla 14 våglängder, x_1, \dots, x_{14} och temperaturerna x_{15} och x_{16} tas med i modellen. I de flesta fall blir $R^2 = 0.90$ eller mer. I sammanhanget ska det nämnas att till en given serie av y -värden blir alltid R^2 minst lika stor då fler x -variabler får ingå i modellen. För att korrigera för detta kan ett modifierat mått R_a^2 (adjusted R^2) användas. Detta mått har en tendens att tona ned inkluderandet av överflödiga x -variabler. I exemplen ovan är R_a^2 ungefär lika med R^2 eftersom antalet β -parametrar är betydligt mindre än antalet observationer. Det enda fall där skillnaden märks annat än marginellt är för 2006 års proteinvärden i Lantmännens instrument, där $R^2 = 0.80$ minskar till $R_a^2 = 0.72$ i modellen med x_1, \dots, x_{14} och från $R^2 = 0.82$ till $R_a^2 = 0.73$ då även temperaturerna x_{15} och x_{16} inkluderas. Detta beror på att endast 50 värden används för att skatta 14 respektive 16 regressionsparametrar. Eftersom höga värden på både R^2 och R_a^2 nästan alltid förekommer finns det skäl att använda modeller med många x -variabler.

Låt skattningarna av $\beta_1, \dots, \beta_{16}$ betecknas med $\hat{\beta}_1, \dots, \hat{\beta}_{16}$. Med hjälp av dessa kan det skattade värdet av ett y skrivas

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_{16} x_{16}$$

Genom att sätta in faktiskt observerade x -värden eller tänkta sådana kan motsvarande y skattas. Differensen $y - \hat{y}$ anger hur pass väl skattningen \hat{y} överensstämmer med y . Exempelen nedan med USA-instrument 1 och 2 visar på goda överensstämmelser.



Figur 2. Samband mellan skattad och verklig proteinhalt, \hat{y} resp y , för a) USA-instrument 1 och b) USA-instrument 2.

De skattade samband som används i figuren ovan är för instrument 1

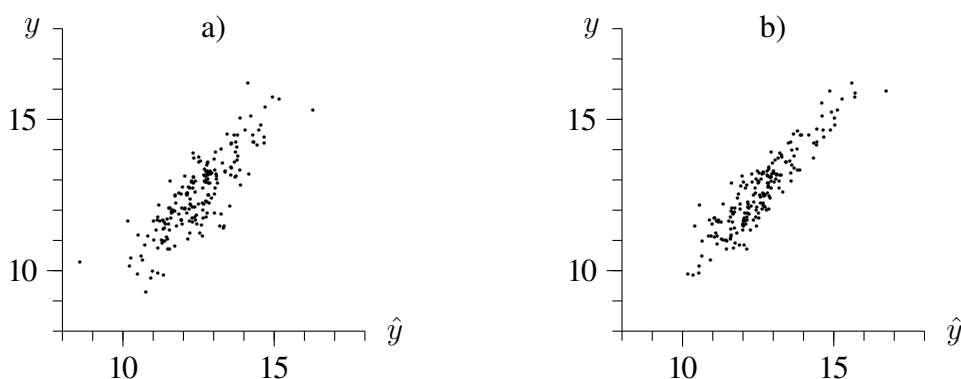
$$\begin{aligned} \hat{y} = & 17.4 - 254.3x_1 + 1129.5x_2 - 952.1x_3 + 74.7x_4 - 136.6x_5 + 127.5x_6 \\ & - 13.5x_7 - 102.9x_8 + 494.4x_9 - 404.0x_{10} - 312.4x_{11} + 264.3x_{12} \\ & + 199.8x_{13} - 112.6x_{14} + 8.3x_{15} - 14.9x_{16} \end{aligned} \quad (1)$$

och för instrument 2

$$\begin{aligned} \hat{y} = & 16.5 - 87.7x_1 + 690.1x_2 - 613.0x_3 - 92.8x_4 - 257.6x_5 + 503.0x_6 \\ & - 33.9x_7 - 531.0x_8 + 1192.7x_9 - 1060.5x_{10} - 194.2x_{11} + 601.7x_{12} \\ & + 117.8x_{13} - 232.7x_{14} + 10.9x_{15} - 19.1x_{16} \end{aligned} \quad (2)$$

I båda fallen ovan används data och skattningarna enbart inom respektive instrument. I meningen att verklig proteinhalt y och skattningen \hat{y} är ganska lika fungerar regressionsmodellen väl i båda fallen, trots att de numeriska värdena på $\hat{\beta}_1, \hat{\beta}_2, \dots$ är mycket olika, speciellt inom instrumenten.

Sett över alla instrument finns värdena på våglängdsvariablerna x_1, \dots, x_{14} i intervallet -0.25 till 1.60 och temperaturerna i intervallet ca 0 till 0.46 (enhet: $^{\circ}\text{C}/100$), vilket gör att uttryck som i (1) och (2) är mycket känsliga ifall x -värdena avviker marginellt från kalibreringsområdet. De stora värdena på koefficienterna har en stor genomslagskraft och kan ge helt orimliga haltangivelser. Detta gäller i synnerhet då t ex instrument 1 betraktas som ett masterinstrument och instrument 2 kalibreras på samma sätt enligt (1). För instrument 2 kan \hat{y} beräknas på detta sätt genom att i stället sätta in instrumentets x -värden i (1). Resultatet återges i följande figur, där också motsvarande utfall visas för instrument 3:



Figur 3. Skattning av proteinhalter med a) instrument 2 och b) instrument 3 kalibrerade enligt instrument 1.

Sambandet är inte lika tydligt i betydelsen att följsamheten till en rät linje i figur 3a är sämre för instrument 2 jämfört med sambandet i figur 2b. För instrument 3 i figur 3b har däremot kalibreringen fungerat ganska väl. Orsaken är att skattningarna $\hat{\beta}_1, \hat{\beta}_2, \dots$ avviker mellan instrumenten 1 och 2. En annan del av förklaringen ligger i att instrument 1 och 3 har 185 av 195 respektive 194 prover gemensamma, medan instrument 2 med 192 prover har 160 av dem gemensamma med instrument 1.

Allmänt sett kan multipel regression i fall liknande denna tillämpning ge helt urartade värden. Detta kan indikeras genom att för varje skattat värde \hat{y} även ange variansen $\text{Var}(y - \hat{y})$ som ett mått på skattningens tillförlitlighet. Om variansen visar sig bli för stor kan en mindre eller ingen tilltro alls sättas till \hat{y} . Variansen kan skattas ganska enkelt med hjälp av en skattning av den observationsvisa variansen σ^2 , de x -värden för de prover som använts vid kalibreringen och de x -värden som är aktuella för \hat{y} . En ytterligare aspekt är att ställa sig frågan om skattningen \hat{y} kan förbättras så att den inte blir så känslig för annorlunda x -värden som illustreras ovan med instrument 1 som master för instrument 2.

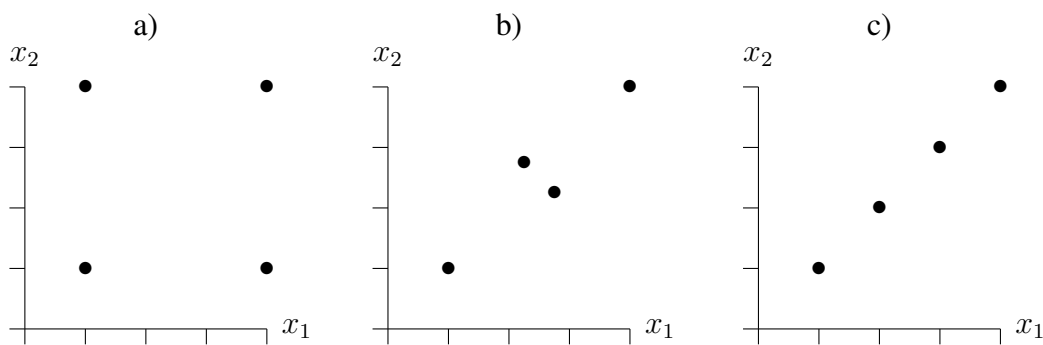
Multikollinearitet

En stor del av förklaringen till att x -värden utanför kalibreringarna leder till osäkra skattningar brukar inom statistisk teori kallas multikollinearitet.

Anta först för enkelhets skull att endast två x -variabler svarande mot två våglängder ska anpassas med hjälp av modellen

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \varepsilon_i, \quad i = 1, \dots, n \quad (3)$$

där n är antalet prover. Geometriskt svarar modellen mot ett plan i en tredimensionell rymd. Betrakta följande tre uppsättningar av värden på (x_1, x_2) som var för sig ska användas till att skatta regressionsplanet.



Figur 4. Olika uppsättningar av (x_1, x_2) för skattning av regressionskoefficienter.

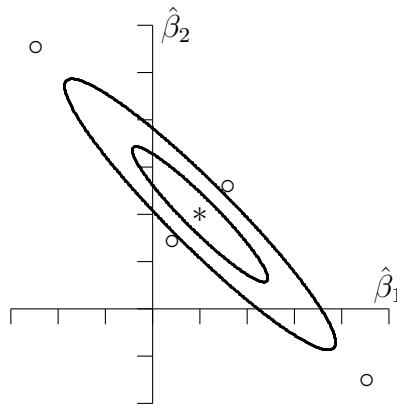
För vilken av de tre uppsättningarna kan man förvänta sig den bästa precisionen hos skattningarna $\hat{\beta}_1$ och $\hat{\beta}_2$? Bildligt sett svarar skattningarnas precision mot att en lutande bordsskiva läggs på ben som placerats enligt respektive delfigur. Ju stadigare placering, desto bättre precision hos skattningen. Utifrån detta är det klart att alternativ c) inte fungerar eftersom man inte kan veta något alls om skivans höjd utanför den linje som bildas av punkterna i (x_1, x_2) -planet. Det är också tydligt att alternativ b) är mindre stabilt än a).

Inom teorin för regression brukar varianserna studeras för att utvärdera tillförlitligheten hos skattade lutningskoefficienter. I detta förenklade fall ska $\text{Var}(\hat{\beta}_j)$ studeras för $j = 1, 2$. Som mått på hur osäker en skattning blir i en modell med två x -variabler jämfört med en brukar man ange

$$\text{VIF}_1 = \text{Variance Inflation Factor}_1 = \frac{\text{Var}(\hat{\beta}_1 \text{ i modell med både } x_1 \text{ och } x_2)}{\text{Var}(\hat{\beta}_1 \text{ i modell med bara } x_1)}$$

och analogt VIF_2 för $\hat{\beta}_2$. Tolkningen är att ju större VIF, desto större försämring av precisionen. I fall a) ovan är $\text{VIF}=1$, i b) är VIF ca 9.5, och i c) är VIF ej definierad pga division med 0 eftersom uppsättningen med x_1 och x_2 längs en linje inte gör det möjligt att skatta både β_1 och β_2 . Allmänt gäller att $1 \leq \text{VIF} < \infty$ förutsatt att skattningar är möjliga att beräkna. En vanlig rekommendation är att $\text{VIF} \leq 10$, se t ex Graybill & Iyer (1994, kap. 5). För transmittansdata är värdena högre, i storleksordningen 500 till 2000 om x_1 och x_2 svarar mot närliggande våglängder och ca 15 till 50 då våglängderna är mer åtskilda.

Effekten på skattningarna $(\hat{\beta}_1, \hat{\beta}_2)$ vid multikollinearitet och höga VIF-värden kan illustreras med följande figur:



Figur 5. Sannolikhetsfördelning för skattningar av regressionskoefficienter vid högt VIF. Sanna värden för β_1 och β_2 vid * och fyra tänkta observationer markerade med o.

Mittpunkten (β_1, β_2) anger de sanna men vid praktiska tillämpningar okända koefficienterna för regressionsplanet (3). Ellipserna anger nivåkurvor för täthetsfunktionen till sannolikhetsfördelningen för skattningarna $\hat{\beta}_1$ och $\hat{\beta}_2$. Ifall observationsserierna kunde upprepas oändligt många gånger skulle man i långa loppet få korrekta skattningar för β_1 och β_2 . Denna egenskap kallas inom statistisk teori väntevärdesriktighet (unbiasedness), se t ex Sen & Srivastava (1990, kap. 2). I figuren framgår att den slumpmässiga variationen är mycket stor i nordväst-sydöstlig riktning. Resultatet blir att $\hat{\beta}_1$ och $\hat{\beta}_2$ ofta är av motsatt tecken samt att deras absolutbelopp är stora, vilket är ekvivalent med att kvadratsumman $\hat{\beta}_1^2 + \hat{\beta}_2^2$ är stor. Mycket långa observationsserier krävs för att ellipserna i denna riktning någorlunda säkert ska kunna närma sig (β_1, β_2) . Däremot är variationen i sydväst-nordöstlig riktning inte så stor, vilket betyder att om x_1 och x_2 för ett prov är ungefär lika så blir variansen för $\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ ganska liten, medan den ökar om x_1 och x_2 är mer åtskilda.

Detta svarar i förenklad form mot vad som illustreras i figurerna 2 och 3 i föregående avsnitt. Om en kalibrering med enbart x_1 och x_2 skulle baseras på instrument 1 så skulle

skattningen $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ bli ganska säker så länge som x_1 och x_2 liknar dem som använts vid kalibreringen. Om andra avläsningar svarande mot ett annat instrument är lite annorlunda än de i instrument 1, så blir $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ osäker eftersom motsvarande hos $\hat{\beta}_1$ och $\hat{\beta}_2$ kan få större effekt.

I situationer med fler än två x -variabler, t ex 16 st, är det inte möjligt att illustrera alla i samma diagram. Däremot kan VIF-värden definieras för varje $\hat{\beta}_j$ enligt:

$$\text{VIF}_j = \frac{\text{Var}(\hat{\beta}_j \text{ i modell med övriga } x\text{-variabler})}{\text{Var}(\hat{\beta}_j \text{ i modell med bara } x_j)} \quad (4)$$

Stora VIF_j -värden tyder på att skattningarna är osäkra så att $\hat{\beta}_1^2 + \dots + \hat{\beta}_{16}^2$ blir stor och teckenbyten är vanliga mellan olika $\hat{\beta}_j$. I denna tillämpning blir också VIF-värdena mycket höga. Det är inte ovanligt med $\text{VIF} > 10^6$, jfr rekommendationen $\text{VIF} \leq 10$. Trots detta är R^2 och R_a^2 höga så att sammantaget har transmittansavläsningarna en hög förklaringsgrad.

Allmänt inom statistisk försöksplanering ska situationer med höga VIF undvikas. I fallet med avlästa transmittanser för olika våglängder är det inte möjligt att helt styra värdena på x -variablerna. Att ha större avstånd på våglängderna skulle nog ge lägre VIF-värden men informationen om halter skulle inte bli bättre med ovidkommande våglängder. Det är här fullt naturligt att x -värdena följs åt i de mätningar som görs. Konsekvensen är att ordinär multipel regression lätt kan ge osäkra skattningar \hat{y} , vilket påkallar modifieringar av regressionsmodellen.

För att lindra effekten av multikollinearitet finns metoder föreslagna. De mest kända är, jfr Draper & Smith (1981, kap. 6),

- a) PCR, principalkomponentregression. Välj ut de linjärkombinationer av x_1, \dots, x_{16} som varierar mest och använd dem som är signifikanta i en regressionsmodell med linjärkombinationerna som förklarande variabler. Sambandet med svarsvariabeln kommer i viss mån i andra hand eftersom de första principalkomponenterna har de bästa chanserna att ge signifikanta resultat utan att för den skull ha de praktiskt mest betydelsefulla sambanden.
- b) PCA, principalkomponenter med y inkluderad bland x_1, \dots, x_{16} . Metoden kan ge uppslag till modeller, men blir svårhanterlig om resultaten skiljer mellan olika instrument.
- c) PLS, Partial Least Squares (svensk benämning förekommer inte). Speciellt användbar då antalet mätningar är färre än antalet våglängder. I denna tillämpning är inte detta aktuellt eftersom det i de flesta fall finns nästan 200 prover till de 14 våglängderna och 2 temperaturerna.
- d) Stegvis regression. Välj successivt ut de våglängder som ger goda förklaringar till svarsvariabeln. Eftersom metoden bygger på upprepade signifikanstest är det svårt att sätta gränser för när en x -variabel ska inkluderas eller exkluderas.
- e) Ridge Regression (svensk benämning förekommer inte). Denna metod kan ses som ett mellanting av vanlig multipel regression och PLS. En konsekvens med ridge regression är att koefficienterna $|\hat{\beta}_j|$ dämpas genom att kvadratsumman $\hat{\beta}_1^2 + \dots + \hat{\beta}_{16}^2$ begränsas på lämpligt sätt.

En egenskap som förekommer i stor utsträckning hos många av metoderna är att absolutvärdena på lutningskoefficienterna blir stora eftersom de var för sig har stora varianser. Den sistnämnda metoden e) innehåller en möjlighet till att begränsa variationen hos skattningarna $\hat{\beta}_j$ genom att kvadratsumman ska anta ett visst värde. I litteraturen anges också detta som ett starkt skäl för ridge regression, speciellt då x -variablerna har samma variation, jfr Draper & Smith (1981, kap. 6). Speciellt gäller detta här för de 14 våglängdsvariablerna. I det följande undersöks därför ridge regression.

Ridge regression

För att presentera metoden studeras först modell (3) med två x -variabler. Modellen skrivs om enligt

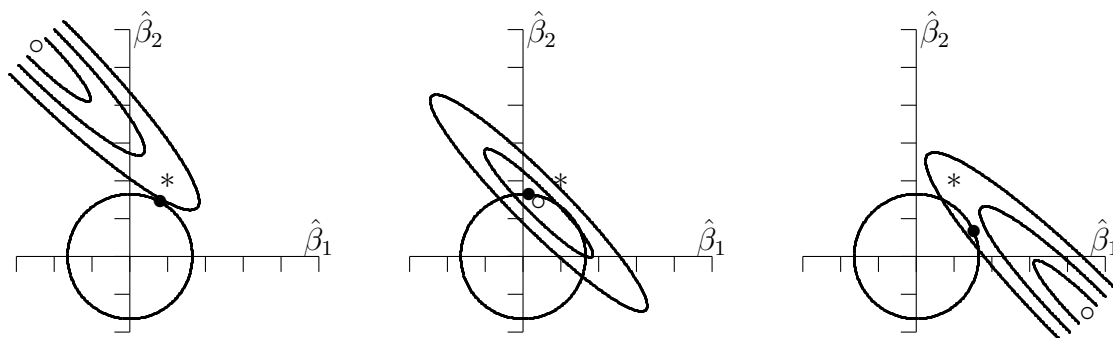
$$y_i = \alpha + \beta_1(x_{i,1} - \bar{x}_1) + \beta_2(x_{i,2} - \bar{x}_2) + \varepsilon_i \quad (5)$$

där \bar{x}_1 och \bar{x}_2 är medelvärdena av de uppmätta x_1 - och x_2 -värdena. Omskrivningen innebär att β_0 i modell (3) ersätts med $\alpha - \beta_1\bar{x}_1 - \beta_2\bar{x}_2$ men att β_1 och β_2 har oförändrade betydelser. Om skattningarna av α , β_1 och β_2 är kända så kan man enkelt beräkna motsvarande för den ursprungliga parametern β_0 .

Minsta-kvadratmetoden innebär att som parameterskattningar $\hat{\alpha}$, $\hat{\beta}_1$ och $\hat{\beta}_2$ tas de värden på α , β_1 och β_2 som minimerar

$$\sum_{i=1}^n (y_i - \alpha - \beta_1(x_{i,1} - \bar{x}_1) - \beta_2(x_{i,2} - \bar{x}_2))^2$$

Det kan visas att $\hat{\alpha} = \bar{y} = (y_1 + \dots + y_n)/n$. De explicita formlerna för $\hat{\beta}_1$ och $\hat{\beta}_2$ är av underordnat intresse eftersom de numeriska resultaten beräknas med hjälp av dator. Figur 5 ovan visar sannolikhetsfördelningen för $(\hat{\beta}_1, \hat{\beta}_2)$ vid hög multikollinearitet. Geometriskt kan minsta-kvadratskattningen ses som en minimipunkt till en buktig yta ovanför planet. Positionen för den erhållna minimipunkten beror av hur mätningarna har utfallit. Figur 6 nedan visar tre tänkbara utfall, alla med samma x -värden men med olika y -värden. Positionen i och höjden ovanför planet varierar, men formen på den buktiga ytan är lika om x -värdena är desamma, vilket för enkelhets skull förutsätts här.



Figur 6. Tre möjliga utfall för minsta-kvadratskattningar. Sanna värden (β_1, β_2) markeras med *, observerad minsta-kvadratskattning $(\hat{\beta}_1, \hat{\beta}_2)$ med \circ , och skattning enligt ridge regression med \bullet .

Alla tre utfallen av minsta-kvadratskattningar har valts så att de är lika sannolika. I fallen a) och c) avviker de mycket från den korrekta lösningen (β_1, β_2) , även om genomsnittet av de tre tänkta utfallen är korrekt.

Metoden med ridge regression innebär att $\hat{\beta}_1$ och $\hat{\beta}_2$ också ska uppfylla bivillkoret $\hat{\beta}_1^2 + \hat{\beta}_2^2 = c^2$, dvs lösningen ska finnas på en cirkel med radien c och centrum i origo. Den punkt på cirkeln som minimerar den buktiga ytan är skattningen enligt ridge regression. I fallen a) och c) ses effekten att skattningen markerad med \bullet på cirkeln blir mer dämpad jämfört med den globala minimipunkten \circ . Resultatet beror i hög grad på radien c ; är den för liten blir skattningen intetsägande och är den för stor innebär ridge regression ingen förbättring jämfört med minsta-kvadratskattningen. Valet av c är inte uppenbart, men metoder finns föreslagna, se nedan.

I statistiska läroböcker om regression, se t ex Draper & Smith (1981) och Sen & Srivastava (1990), brukar teorin för ridge regression presenteras på flera sätt. Det mest vanliga är att beskriva metoden enligt nedan för att sedan visa på olika egenskaper, bl a den att kvadratsumman av koefficienterna begränsas. I detta sammanhang har denna egenskap använts som en motivering för metoden.

Den gängse beskrivningen av ridge regression är i korthet enligt följande: Låt först modellen ha två x -variabler. Observationerna för modellen (5) kan sättas upp i vektorer och matriser enligt följande:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} - \bar{x}_1 & x_{1,2} - \bar{x}_2 \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} - \bar{x}_1 & x_{n,2} - \bar{x}_2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

eller med ett mer komprimerat skrivsätt

$$y = \mathbf{1}\alpha + X\beta + \varepsilon \quad (6)$$

där

$$\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \quad \text{och} \quad X = \begin{bmatrix} x_{1,1} - \bar{x}_1 & x_{1,2} - \bar{x}_2 \\ \vdots & \vdots \\ x_{n,1} - \bar{x}_1 & x_{n,2} - \bar{x}_2 \end{bmatrix}$$

Minsta-kvadratskattningarna kan skrivas

$$\hat{\alpha} = \bar{y} \quad \text{och} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1}X'y \quad (7)$$

Det detaljerade räknandet överläts på dator. Inför införandet av skattningar enligt ridge regression standardiseras x -variablerna genom att de ersätts med z -variablerna

$$z_{i,1} = (x_{i,1} - \bar{x}_1) / \sqrt{\sum_{i=1}^n (x_{i,1} - \bar{x}_1)^2} \quad \text{och} \quad z_{i,2} = (x_{i,2} - \bar{x}_2) / \sqrt{\sum_{i=1}^n (x_{i,2} - \bar{x}_2)^2}$$

Följden blir att varje β_j byts mot $\beta_{z,j}$ så att modellen (5) i stället skrivs

$$y_i = \alpha + \beta_{z,1}z_{i,1} + \beta_{z,2}z_{i,2} + \varepsilon_i \quad (8)$$

med $\beta_{z,1} = \beta_1 \sqrt{\sum (x_{i,1} - \bar{x}_1)^2}$ och $\beta_{z,2} = \beta_2 \sqrt{\sum (x_{i,2} - \bar{x}_2)^2}$. Skalbytet innebär ingen inskränkning, det är lätt att återtransformera (ungefär som att byta mellan cm och tum vid längdmätningar). Analogt med (6) kan (8) också skrivas på matrisform enligt

$$y = \mathbf{1}\alpha + Z\beta_z + \varepsilon \quad (9)$$

där

$$Z = \begin{bmatrix} z_{1,1} & z_{1,2} \\ \vdots & \vdots \\ z_{n,1} & z_{n,2} \end{bmatrix}$$

Minsta-kvadrat-skattningen för β_z erhålls genom

$$\hat{\beta}_z = (Z'Z)^{-1}Z'y \quad (10)$$

En konsekvens av standardiseringen är att alla diagonalelement hos $Z'Z$ blir lika med 1.

Ridge regression med parametern k definieras genom att modifiera skattningen (10) enligt

$$\hat{\beta}_z(k) = (Z'Z + kI)^{-1}Z'y \quad (11)$$

där $k \geq 0$ och I är identitetsmatrisen (diagonalelementen=1 och övriga=0). Valet $k = 0$ svarar mot den ursprungliga minsta-kvadratskattningen.

I föregående stycke har ridge regression presenterats i fallet med två regressionsvariabler. Generaliseringar till ett godtyckligt antal variabler, låt säga p , är omedelbara och modelluttrycket (9) är identiskt men med

$$Z = \begin{bmatrix} z_{1,1} & \dots & z_{1,p} \\ \vdots & \vdots & \vdots \\ z_{n,1} & \dots & z_{n,p} \end{bmatrix}, \quad \text{där} \quad z_{i,j} = (x_{i,j} - \bar{x}_j) / \sqrt{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2}$$

Speciellt gäller fortfarande (11). Räkningarna blir förstås mer omfattande men vållar inga problem för ett matrishanterande datorprogram.

Med hjälp av matrisalgebra följer att kvadratsumman för koefficienterna $\hat{\beta}_z(k)$ är lika med

$$\hat{\beta}_{z,1}^2(k) + \dots + \hat{\beta}_{z,p}^2(k) = \hat{\beta}_z(k)' \hat{\beta}_z(k) = y'Z(Z'Z + kI)^{-2}Z'y$$

Det finns ett samband mellan parametern k och cirkelradien c illustrerad i figur 6, även om k inte kan uttryckas enkelt i c . Det kan visas att ett växande k medför att $\hat{\beta}_z(k)' \hat{\beta}_z(k) = c^2$ avtar, vilket utnyttjas för att få skattningar med mindre variation.

Den förväntade avvikelser (väntevärdesfelet, biasen) för (11) från det sanna värdet β_z är

$$E(\hat{\beta}_z(k) - \beta_z) = ((Z'Z + kI)^{-1}Z'Z - I)\beta_z = -k(Z'Z + kI)^{-1}\beta_z$$

som är 0 endast då $k = 0$. Ett ökande k innebär således att ett väntevärdesfel uppträder. Varians-kovariansmatrisen för (11) ges av

$$\text{Var}(\hat{\beta}_z(k)) = (Z'Z + kI)^{-1}Z'Z(Z'Z + kI)^{-1}\sigma^2$$

vilken kan visas avta då k växer.

Det är alltså två egenskaper som förändras med k , en till det bättre och en till det sämre. Motsvarande gäller också då ett enskilt y_0 svarande mot ett nytt prov ska skattas med $\hat{y}_0(k)$ enligt

$$\hat{y}_0(k) = \bar{y} + z_0' \hat{\beta}_z(k) \quad (12)$$

där $z'_0 = [z_{0,1}, \dots, z_{0,p}]$ är vektorn av de transformerade x -värdena som har avlästs i instrumentet för det aktuella provet. Avvikelsen $y_0 - \hat{y}_0(k)$ har då väntevärdet

$$\begin{aligned} E(y_0 - \hat{y}_0(k)) &= E(y_0 - \bar{y} - z'_0 \hat{\beta}_z(k)) \\ &= z'_0 E(\beta_z - \hat{\beta}_z(k)) = kz'_0 (Z'Z + kI)^{-1} \beta_z \end{aligned}$$

och variansen är

$$\begin{aligned} \text{Var}(y_0 - \hat{y}_0(k)) &= \text{Var}(y_0 - \bar{y} - z'_0 \hat{\beta}_z(k)) \\ &= \left[\frac{n+1}{n} + z'_0 (Z'Z + kI)^{-1} Z'Z (Z'Z + kI)^{-1} z_0 \right] \sigma^2 \end{aligned}$$

För att sammanfatta väntevärdesfelet och variansen för $y_0 - \hat{y}_0(k)$ används det så kallade medelkvadratfelet (mean square error) definierat av

$$\begin{aligned} \text{MSE}(y_0 - \hat{y}_0(k)) &= \text{Var}(y_0 - \hat{y}_0(k)) + [E(y_0 - \hat{y}_0(k))]^2 \\ &= \left[\frac{n+1}{n} + z'_0 (Z'Z + kI)^{-1} Z'Z (Z'Z + kI)^{-1} z_0 \right] \sigma^2 + k^2 [z'_0 (Z'Z + kI)^{-1} \beta_z]^2 \end{aligned} \quad (13)$$

Uttrycket är komplicerat. Dessutom beror det av parametern σ^2 och koefficienterna $\beta_{z,1}, \dots, \beta_{z,p}$ i β_z som alla har okända värden. Matematiskt kan det visas att

$$\begin{aligned} \text{MSE} &\leq \left[\frac{n+1}{n} + z'_0 (Z'Z + kI)^{-1} Z'Z (Z'Z + kI)^{-1} z_0 \right] \sigma^2 + k^2 z'_0 (Z'Z + kI)^{-2} z_0 \cdot \beta'_z \beta_z \\ &= \left[\frac{n+1}{n} + z'_0 (Z'Z + kI)^{-1} \left(Z'Z + \frac{k^2 \beta'_z \beta_z}{\sigma^2} I \right) (Z'Z + kI)^{-1} z_0 \right] \sigma^2 \end{aligned} \quad (14)$$

Det kan visas att denna övre gräns minimeras då $k = k^* = \sigma^2 / \beta'_z \beta_z$. Denna kvot beror av okända storheter. Ett sätt att skatta k^* är att först skatta σ^2 och β_z med ordinär multipel regression. Skattningen för k^* kan successivt korrigeras genom att använda det senast uträknade värdet på k för att få fram ett uppdaterat $\hat{\beta}(k)$. Metoden har dock en tendens att ge alltför små värden på k^* . Ett ytterligare påpekande är att uttrycket för k^* inte beror av de aktuella värdena i z'_0 . Ett villkor för att k^* kopplat till den övre gränsen för MSE är rimligt är att uppskattningen (14) är någorlunda precis, vilket beror av vektorn z_0 .

En annan metod att finna ett värde på k för ridge regression är att lösa ekvationen

$$\frac{y'Z(Z'Z + kI)^{-2}Z'y}{y'Z(Z'Z)^{-2}Z'y} = \frac{\hat{\beta}_z(k)' \hat{\beta}_z(k)}{\hat{\beta}_z(0)' \hat{\beta}_z(0)} = q^2 \quad (15)$$

för något q , $0 < q < 1$. Talet q är ett övergripande mått på förhållandet mellan storleksordningen av komponenterna i $\hat{\beta}_z(k)$ jämfört med $\hat{\beta}_z(0)$. Denna metod svarar mot den som i litteraturen kallas ridge plots, där skattningarna $\hat{\beta}_{z,j}(k)$ plottas mot växande värden av k . Det k där skattningarna stabiliseras någorlunda kan tas som värde på parametern. Metoden är subjektiv eftersom skattningarna egentligen inte stabiliseras helt förrän $k \rightarrow \infty$, då skattningarna alla blir 0. I Draper & Smith (1981) och Sen & Srivastava (1990) finns olika förslag på andra metoder. Där nämns också att det inte finns någon metod som alltid är den bästa.

Resultat

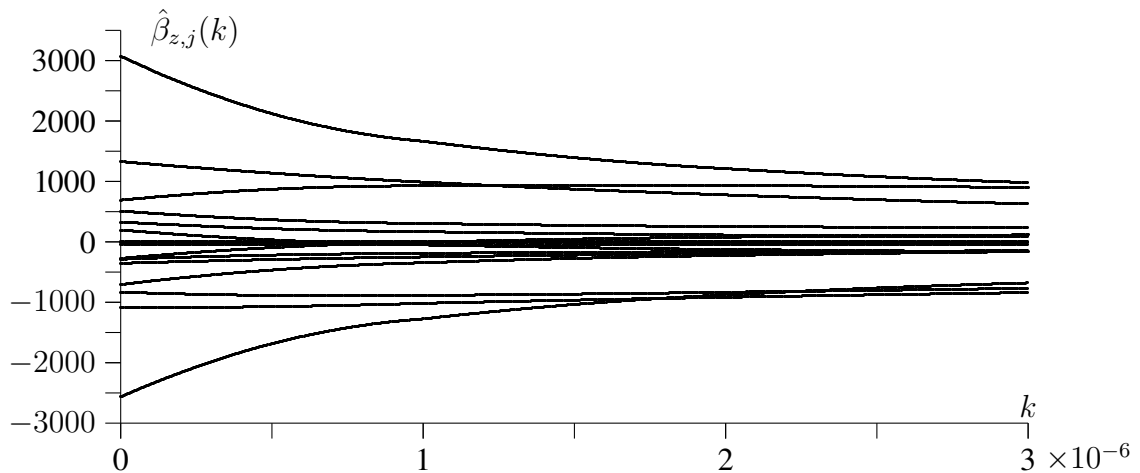
USA-instrumenten. Ett masterinstrument för kalibrering av flera slavinstrument.

Låt kalibreringen baseras på de prover som mätts med instrument 1. Ordinär multipel linjär regression ($k = 0$) ger skattningen

$$\begin{aligned} \hat{y}(0) = & 12.6 - 705.2z_1 + 3074.9z_2 - 2564.8z_3 + 197.1z_4 - 354.3z_5 + 329.2z_6 \\ & - 35.1z_7 - 273.3z_8 + 1328.8z_9 - 1084.8z_{10} - 832.3z_{11} + 691.2z_{12} \\ & + 511.2z_{13} - 286.2z_{14} + 5.5z_{15} - 14.7z_{16} \end{aligned} \quad (16)$$

Kvadratsumman $\hat{\beta}_z(0)' \hat{\beta}_z(0) = 21\,335\,670$. Regressionssambandet är ekvivalent med (1) så när som på att x -variablerna här har bytts mot z -variabler, vilket leder till att $\hat{\beta}_{z,j}(0)$ måste skalas upp eftersom z -värdena är mer koncentrerade än motsvarande x -värden. För våglängdsvariablerna varierar VIF, definierat enligt (4), från 63 100 (för z_1) till 610 999 (för z_{10}), medan VIF för temperaturerna är 27.5 och 51.6. Skattningen av observationsvariansen är $s^2 = 0.147$, ($s = 0.383$). Värdena på s^2 och VIF förändras ej av att x byts mot z .

En s_k ridge plot av hur regressionskoefficienterna förändras med k visas i figur 7. Eftersom koefficienterna i stort sett stabiliseras efter det att deras storleksordning har halverats sätts $q = 0.5$ i ekvationen (15), varur lösningen $k = 2.1 \cdot 10^{-6}$ beräknas.

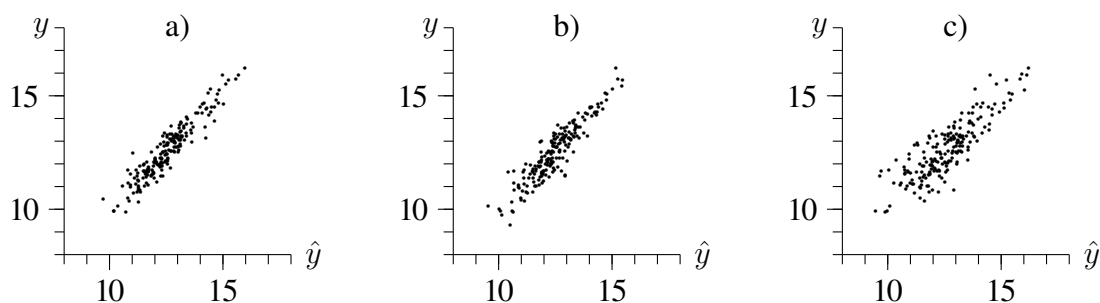


Figur 7. Ridge plot. Inflytande av k på skattningar av regressionskoefficienter.

Med hjälp av (11) och (12) erhålls med ridge regression följande skattning av sambandet mellan y och z -variablerna:

$$\begin{aligned} \hat{y}(2.1 \cdot 10^{-6}) = & 12.6 - 215.8z_1 + 1182.8z_2 - 850.7z_3 - 130.2z_4 - 188.2z_5 + 110.5z_6 \\ & - 45.8z_7 + 83.6z_8 + 760.7z_9 - 907.7z_{10} - 826.9z_{11} + 930.9z_{12} \\ & + 252.6z_{13} - 160.3z_{14} + 12.3z_{15} - 12.9z_{16} \end{aligned}$$

Figuren nedan illustrerar utfallet av $\hat{y}(2.1 \cdot 10^{-6})$ för instrumenten 1, 2 och 3.



Figur 8. Samband mellan skattad och verklig proteinhalt, $\hat{y}(2.1 \cdot 10^{-6})$ resp y , för a) USA-instrument 1 (kalibrator), b) instrument 2 och c) instrument 3.

En jämförelse med figurerna 2 och 3, baserade på minsta-kvadratskattningarna $\hat{y}(0)$, ger att resultaten för instrument 1 är i stort sett lika, se figurerna 2a och 8a. För instrument 2 innebär ridge regression en förbättring, jfr figurerna 3a och 8b, medan en liten försämring sker för instrument 3, se figurerna 3b och 8c. För att få överblick på alla 24 instrument studeras det genomsnittliga medelkvadratfelet för $\hat{y}(k)$ definierat av

$$a(k) = \frac{1}{n} \sum_1^n (y_i - \hat{y}_i(k))^2$$

som i fallet $k = 0$ svarar mot skattningen av σ^2 förutom att nämnaren för s^2 är $n - 17$. Numeriska resultat ges i följande tabell för $k = 0, 2.1 \cdot 10^{-6}$ och $4.2 \cdot 10^{-6}$. Det sista värdet svarar mot $q = 0.39$ i (15).

Tabell 1. Genomsnittliga medelkvadratfel vid skattningar kalibrerade enligt instrument 1.

Instrument:	1	2	3	4	5	6	7	8	9	10	11	12
$a(0)$	0.15	0.53	0.29	0.36	0.67	0.59	1.90	1.49	0.68	1.90	0.63	0.55
$a(2.1 \cdot 10^{-6})$	0.20	0.25	0.54	0.28	0.43	0.56	0.84	0.82	0.74	1.06	0.43	0.50
$a(4.2 \cdot 10^{-6})$	0.23	0.27	0.56	0.28	0.38	0.63	0.73	0.74	0.76	0.86	0.44	0.57
Instrument:	13	14	15	16	17	18	19	20	21	22	23	24
$a(0)$	0.42	3.90	0.55	0.92	0.66	0.91	2.47	0.83	1.59	0.56	0.70	0.27
$a(2.1 \cdot 10^{-6})$	0.43	1.26	0.45	0.46	0.49	0.30	0.84	0.57	0.64	0.47	0.62	0.37
$a(4.2 \cdot 10^{-6})$	0.54	1.01	0.52	0.44	0.52	0.32	0.66	0.63	0.56	0.55	0.63	0.44

Genom att inom varje instrument förse de tre värdena på k med placeringsnummer 1, 2, 3 efter storleken på $a(k)$ visar sig $k = 2.1 \cdot 10^{-6}$ ha placeringssumman 36.5, $k = 4.2 \cdot 10^{-6}$ ha 46.5 och $k = 0$ ha 61. (I vissa fall är det oavgjort så att summan av placeringsnumren delas jämnt.)

USA-instrumenten. Separata kalibreringar.

När ett instrument ska användas i en ny tillämpning kan antalet referensprover vara begränsat. För att studera hur ridge regression fungerar med ett relativt litet kalibreringsunderlag väljs här hälften av proverna för instrument 1. Övriga prover används för att kontrollera utfallet. Till instrument 1 finns 195 prover. Baserat på de 97 första erhålls

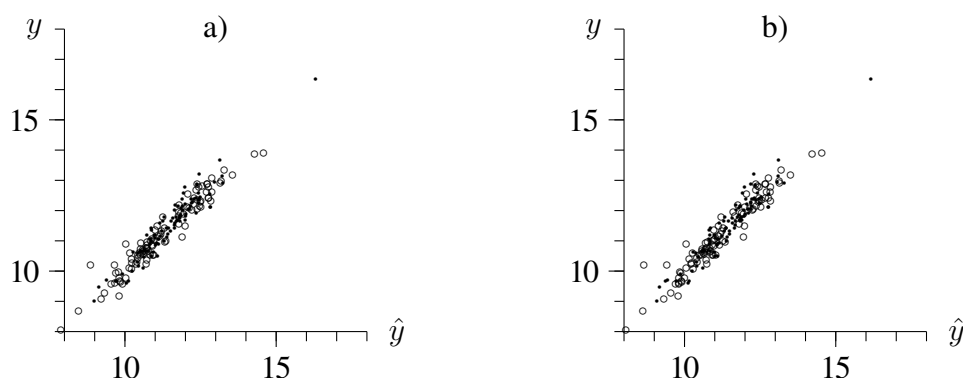
följande minsta-kvadratskattning (ridge regression med $k = 0$) för vattenhalten enligt modellen (9).

$$\begin{aligned}\hat{y}(0) = & 11.5 + 28.0z_1 + 243.9z_2 - 368.9z_3 + 221.4z_4 - 551.9z_5 + 88.8z_6 \\ & + 477.7z_7 - 111.9z_8 + 236.8z_9 - 572.1z_{10} + 457.5z_{11} - 161.6z_{12} \\ & + 53.8z_{13} - 40.6z_{14} - 4.6z_{15} - 2.3z_{16}\end{aligned}$$

Skattningen av observationsvariansen σ^2 är $s^2 = 0.109$ ($s = 0.330$), och förklaringsgraderna $R^2 = 0.92$ och $R_a^2 = 0.90$. VIF-värdena för de standardiserade våglängdsvariablerna varierar från 79 673 till 783 484 medan VIF för temperaturvariablerna är 45.5 och 49.5. En studie av ridge plots för $\hat{\beta}_{z,j}(k)$ ger att de i stort sett stabiliseras då deras storleksordning halveras. Insättning av $q = 0.5$ i (15) ger lösningen $k = 2.4 \cdot 10^{-6}$. Skattning enligt (11) ger

$$\begin{aligned}\hat{y}(2.4 \cdot 10^{-6}) = & 11.5 + 143.3z_1 + 6.6z_2 - 257.8z_3 + 96.5z_4 - 347.3z_5 + 105.2z_6 \\ & + 308.6z_7 + 9.6z_8 + 1.7z_9 - 145.5z_{10} + 81.6z_{11} + 18.7z_{12} \\ & + 11.2z_{13} - 31.4z_{14} - 5.7z_{15} - 0.9z_{16}\end{aligned}$$

Resultatet av skattningarna visas i figur 9 nedan:



Figur 9. Skattade vattenhalter med USA-instrument 1, a) $\hat{y}(0)$ och b) $\hat{y}(2.4 \cdot 10^{-6})$. Kalibreringsdata anges med \bullet , övriga med \circ .

Diagrammen är i stort sett identiska, vilket visar på att ridge regression har liten effekt. De 97 prover som används för kalibreringen har sina z -värden i samma område som de övriga proverna. Detta medför att skattningarna $\hat{y}(0)$ och $\hat{y}(k)$ med ett något så när litet k ger goda resultat. Överensstämmelsen mellan $\hat{y}(0)$ och $\hat{y}(k)$ är i huvudsak god även för övriga instrument. I ett fåtal fall finns prover med mycket avvikande skattningar då de inte ingår i kalibreringsmängden. Dessa kan inte korrigeras med annat än att k väljs stort med följden att skattningen för övriga prover blir utslätad och oprecis.

Svenska instrumenten

Ridge regression studeras här med kalibreringen baserad på Lantmännens instrument. De data som används för kalibreringen är Lantmännens referensprover från 2005 uppmätta i låg temperatur samt värdena från 2006. Dessa data har valts för att få en någorlunda stor variation mellan x -värdena och samtidigt ett begränsat antal värden, 122 respektive 50. De effekter med ridge regression som uppkommer för de svenska instrumenten framträder tydligast för vattenhalterna och för korthets skull presenteras endast denna variabel i detta avsnitt.

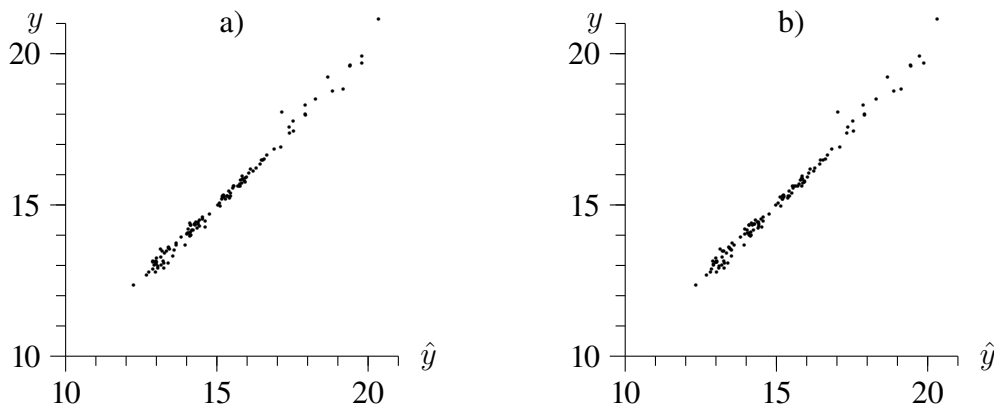
Minsta-kvadratskattningen för vattenhalten baserad på kalibreringsdata är

$$\begin{aligned}\hat{y}(0) = & 17.8 + 37.6x_1 - 66.7x_2 + 32.8x_3 + 8.7x_4 - 16.1x_5 - 269.7x_6 \\ & + 274.8x_7 + 59.4x_8 + 99.7x_9 - 99.8x_{10} - 15.3x_{11} - 90.2x_{12} \\ & + 69.9x_{13} - 25.3x_{14} - 24.7x_{15} - 0.27x_{16}\end{aligned}\quad (17)$$

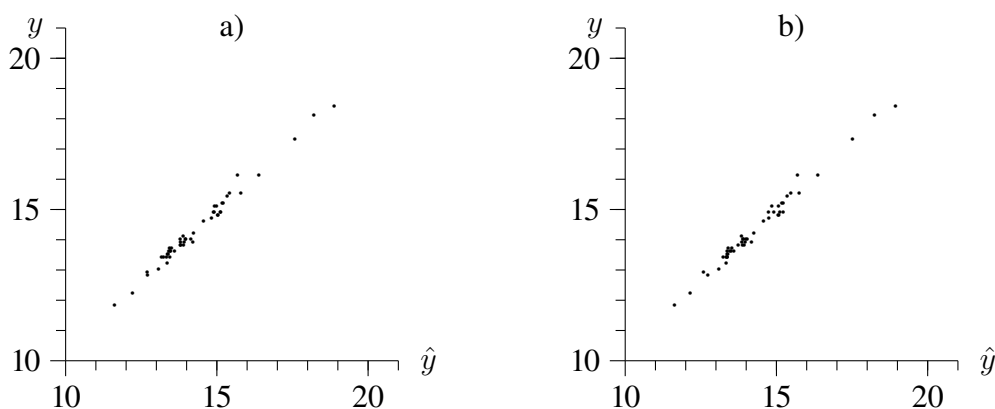
med $s^2 = 0.0376$ ($s = 0.194$), $R^2 = R_a^2 = 0.99$. VIF-värdena för våglängdsvariablerna varierar från 41 069 till 703 257 och för temperaturerna är de 145.3 och 77.2. För $q = 0.5$ i (15) erhålls lösningen $k = 3.6 \cdot 10^{-6}$ som parameter för ridge regression. Detta värde insatt ger skattningen

$$\begin{aligned}\hat{y}(3.6 \cdot 10^{-6}) = & 18.1 + 39.8x_1 - 31.3x_2 - 4.8x_3 - 31.4x_4 - 61.8x_5 - 56.5x_6 \\ & + 132.5x_7 + 83.0x_8 + 10.3x_9 - 24.0x_{10} - 7.4x_{11} - 52.5x_{12} \\ & + 13.6x_{13} - 9.6x_{14} - 23.9x_{15} - 2.0x_{16}\end{aligned}\quad (18)$$

Resultatet av uttrycken $\hat{y}(0)$ och $\hat{y}(3.6 \cdot 10^{-6})$ för kalibreringsdata visas i figur 10 och 11.

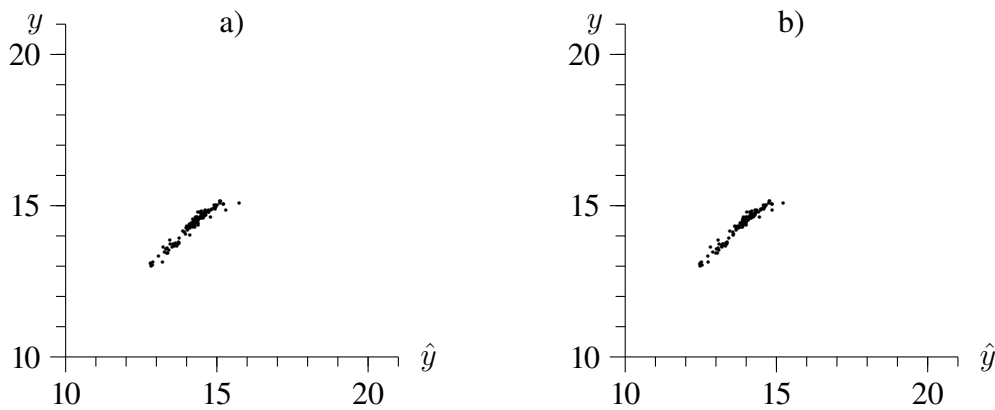


Figur 10. Skattade vattenhalter för Lantmännens referensdata från 2005 i låg temperatur uppmätta i Lantmännens instrument. Kalibrering enligt a) $\hat{y}(0)$ i (17) och b) $\hat{y}(k)$ i (18).



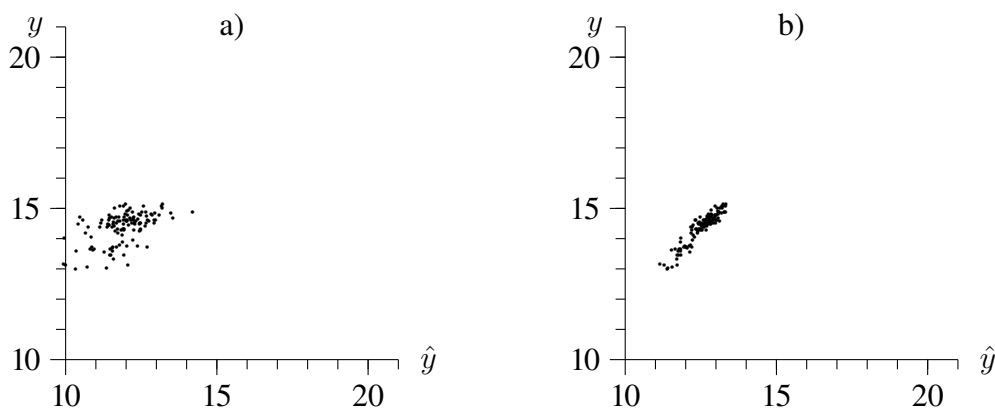
Figur 11. Skattade vattenhalter för Lantmännens referensdata från 2006 uppmätta i Lantmännens instrument. Kalibrering enligt a) $\hat{y}(0)$ i (17) och b) $\hat{y}(k)$ i (18).

För data från Mäljarvetetävlingen 2005 anges i figur 12 samband mellan skattad vattenhalt utifrån mätningar i Lantmännens instrument vid rumstemperatur men med kalibreringar med hjälp av de data som illustreras i figur 10 och 11.



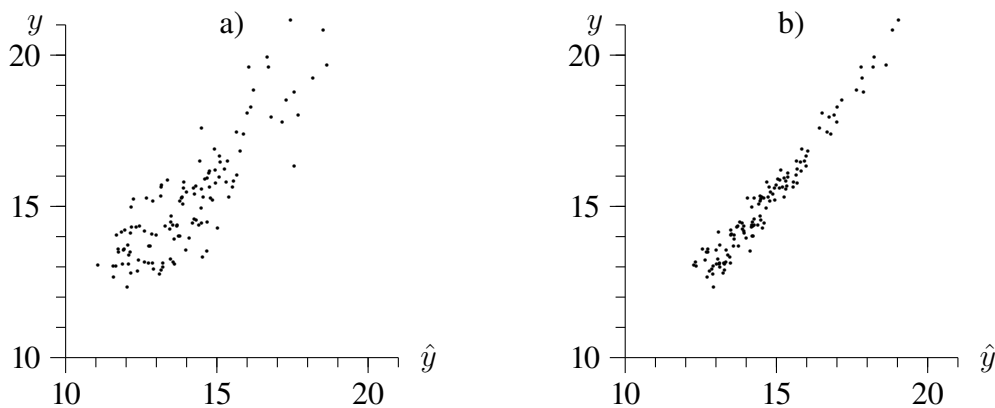
Figur 12. Skattade vattenhalter för prover från Mälärvetetävlingen 2005 uppmätta i rumstemperatur med Lantmännens instrument. Kalibrering enligt a) $\hat{y}(0)$ i (17) och b) $\hat{y}(k)$ i (18).

Motsvarande resultat för JTI:s instrument med x -värden från Mälärvetetävlingen 2005 avlästa i rumstemperatur illustreras i figur 13.



Figur 13. Skattade vattenhalter för referensdata från Mälärvetetävlingen 2005 uppmätta i rumstemperatur med JTI:s instrument. Kalibrering enligt a) $\hat{y}(0)$ i (17) och b) $\hat{y}(k)$ i (18) baserad på Lantmännens instrument.

I figur 14 visas resultat med x -värden avlästa vid låg temperatur i JTI:s instrument för Lantmännens referensdata från 2005, dvs en av de datamängder som används för kalibreringen med Lantmännens instrument.



Figur 14. Skattade vattenhalter för Lantmännens referensdata från 2005 uppmätta i låg temperatur med JTI:s instrument. Kalibrering enligt a) $\hat{y}(0)$ i (17) och b) $\hat{y}(k)$ i (18) baserad på Lantmännens instrument.

Resultaten för de svenska instrumenten är blandade. I vissa fall är skattningarna inte så känsliga, t ex i figurerna 10, 11 och 12 där resultaten är i stort sett inte visar någon skillnad mellan a-delarna med vanlig regression och i b-delarna med ridge regression. I figurerna 13 och 14 däremot syns en mycket klar skillnad mellan metoderna. De till vänster med multipel regression innehåller en större osäkerhet jämfört med dem till höger där skattningarna korrigerats med ridge regression. I figur 13 framträder också det nivåfel som kan förekomma vid såväl multipel som ridge regression genom att värdena på x - och y -axlarna inte överensstämmer även om följsamheten till en rät linje är någorlunda god i figur 13 b. I figur 14 har korrigerings gjorts för nivåfelet.

Diskussion

Möjligheten till användbara resultat med ridge regression beror i hög grad av om parametern k kan väljas lämpligt. Som omnämnts tidigare minimeras den övre gränsen (14) då $k = \sigma^2 / \beta_z' \beta_z$, men eftersom kvoten beror av de parametrar som ska skattas hamnar man i en slags återvändsgränd. Problemen med att välja ett lämpligt k finns behandlat, men det finns ingen formelbaserad metod föreslagen som fungerar väl i alla situationer. Snarare framhålls det att olika plottar ska göras för vägledning till val av k . I detta projekt har ridge plots som i figur 7 studerats varvid ett lämpligt q , vanligen $q = 0.5$, har satts in i (15). I de flesta fall har en förbättring erhållits jämfört med den ordinära regressionsskattningen baserad på minsta-kvadratmetoden svarande mot $k = 0$. I de fall då en försämring har skett är den mycket marginell.

Samtidigt ska det framhållas att ridge regression inte är en metod som kan lösa alla situationer. Det tal k som används är vanligtvis mycket litet, ofta är det fråga om miljondelar. Ett något för stort k kan resultera i utslätade och ej meningsfulla skattningar.

Ridge regression kan ge en bättre följsamhet till en rät linje jämfört med multipel regression, se t ex fig 8b, 13 och 14. I vissa fall behöver nivån på linjen korrigeras. Konsekvensen av utebliven korrigerings visas i fig 13. Denna kalibrering som gjorts för fig 14 blir mest tillförlitlig om den utförs med hjälp av referensprover separat för varje instrument.

I de data som studerats i projektet finns facit för protein- och vattenhalterna i meningen att de har mätts upp med det mer exakta men också dyrare referensinstrumentet (Foss Infratec). Hur ska olämpliga skattningar avslöjas i en situation då facit inte finns? Ett möjligt sätt är att till varje skattning $\hat{y}(k)$ även ange kvadratroten av dess medelkvadratfel MSE definierat i (13) och eventuellt också den övre gränsen i (14). För detta behövs skattningar av σ^2 och komponenterna i β_z . Den första är enkel att få fram med god precision genom ordinär multipel regression, men osäkerheten är större för β_z . Å andra sidan kan denna osäkerhet kompenseras av att k vanligen är ganska litet så att det väsentliga bidraget till $MSE(y - \hat{y}(k))$ kommer från $\text{Var}(y - \hat{y}(k))$. Om en observation med insatta x - eller likvärdigt z -värden ger ett högt MSE ska den kunna ifrågasättas med åtgärder som betingas av eventuella konsekvenser beroende på t ex felaktigt skattad proteinhalt. Frekventa förekomster av höga MSE bör föranleda nya kalibreringar.

Litteratur

Draper, N R & Smith, H (1981). *Applied Regression Analysis, Second Edition*. Wiley, New York.

Graybill, F A & Iyer, H K (1994). *Regression Analysis: Concepts and Applications*. Duxbury Press, Belmont.

Sen, A & Srivastava, M (1990). *Regression Analysis. Theory, Methods, and Applications*. Springer-Verlag, New York.

JTI – Institutet för jordbruks- och miljöteknik...

... är ett industriforskningsinstitut som forskar, utvecklar och informerar inom områdena jordbruks- och miljöteknik samt arbetsmaskiner. Vårt arbete ger dig bättre beslutsunderlag, stärkt konkurrenskraft och klokare hushållning med naturresurserna.

Vi publicerar regelbundet notiser på vår webbplats om aktuell forskning och utveckling vid JTI. Du får notiserna hemskickade gratis om du anmäler dig på www.jti.se

På webbplatsen finns även publikationer som kan läsas och laddas hem gratis, t.ex.:

JTI-informerar, som kortfattat beskriver ny teknik, nya rön och nya metoder inom jordbruk och miljö (4-5 teman/år).

JTI-rapporter, som är vetenskapliga sammanställningar över olika projekt.

Samtliga publikationer kan beställas i tryckt form. JTI-rapporterna och JTI-informerar kan beställas som lösnummer. Du kan också prenumerera på JTI-informerar.

*För trycksaksbeställningar, prenumerationsärenden m.m.,
kontakta vår publikationstjänst (SLU Service Publikationer):*

tfn 018 - 67 11 00, fax 018 - 67 35 00

e-post: bestallning@jti.slu.se



JTI – Institutet för jordbruks- och miljöteknik

JTI – Swedish Institute of Agricultural and Environmental Engineering

Box 7033, 750 07 UPPSALA

Telefon: 018 - 30 33 00

Besöksadress: Ultunaallén 4

Telefax: 018 - 30 09 56

Webbplats: www.jti.se