Acta Oeconomica Kaposváriensis (2007) Vol 1 No 1-2, 113-119
Kaposvári Egyetem, Gazdaságtudományi Kar, Kaposvár
Kaposvár University, Faculty of Economic Science, Kaposvár

A PLS (Partial Least Squares) regresszió és alkalmazása

Kövér[1] György, Bázár[2] György
[1]Kaposvári Egyetem, Gazdaságtudományi Kar, Matematika és Fizika Tanszék, 7400 Kaposvár, Guba Sándor u. 40.
[2] Kaposvári Egyetem, Állattudományi Kar, Sertés- és Kisállattenyésztési Tanszék,7400 Kaposvár, Guba Sándor u. 40.

ÖSSZEFOGLALÁS

A 29 mangalica húsmintából származó NIR spektrumok valamint a szárazanyag-, zsír- és fehérjetartalmat szolgáltató kémiai analízis eredményeit feldolgozva egy- és többváltozós kalibrációs modelleket készítettünk PLS és PCR regresszió segítségével, hogy a húsminták kémiai összetételét megbecsülhessük. Megállapítottuk, hogy a végső modell igen magas arányban (93.50…99.04%) magyarázza a függő változók varianciáját. A keresztvalidációs eredmények vizsgálata azt mutatja, hogy a 29 minta alapján négy komponensre alapozva robusztus kalibrációs egyenletek készíthetők. A statisztikai elemzéshez felhasznált „R" programcsomag megfelelő grafikus és numerikus outputot szolgáltatott a következtetések levonásához.
(Kulcsszavak: PLS regresszió, PCR, mangalica, „R")

A PLS (Partial Least Squares) regression and an application

György Kövér[1], György Bázár[2]
[1]Kaposvár University, Faculty of Economic Science, Department of Mathematics and Physics, H-7400 Kaposvár Guba S. u. 40
[2]Kaposvár University, Faculty of Economic Science, Department of Pig and Small Animal Breeding, H-7400 Kaposvár Guba S. u. 40

ABSTRACT

The aim of this study was to develop calibration equations to predict the chemical compositon of 29 mangalitza meat samples by means of near infrared spectroscoy (NIRS). Several different uni- and multivariate PLS and PCR were created. It was found that the variance of dry matter, ether extract and protein concentration was determined by the final prediction equations in 93.5% …99.04%. After the crossvalidation process, a 4 component robust prediction equation was concluded. The pls package of CRAN R is designed such that it provides the necessary procedures and plots to create sufficient prediction models for NIR spectroscopy.
(Keywords: PLS regression, PCR, mangalicza, "R")

BEVEZETÉS

Jelen közleményben a liofilizált minták szárazanyag-, zsír- és fehérjetartalmának becslését lehetővé tevő kalibrációs egyenletek létrehozására szolgáló módszerek közül a PLS regresszió (részleges legkisebb négyzetek) alkalmazására helyezzük a hangsúlyt.

Annak ellenére, hogy a PLS regresszió viszonylag új keletű eszköz a kísérleti adatok feldolgozásában, az elméleti megalapozásnak máris bőséges szakirodalma található, az egyes változatok, alkalmazások száma egyre gyarapszik (*Siesler és mtsai*., 2002). A módszer kedveltségére, használhatóságára utal az is, hogy *Wald* (2001) már azt javasolja a

szakmai közönségnek, hogy a PLS rövidítést „Projection to Latent Structures" (Vetítés látens struktúrákra) jelentéssel töltsék meg, ami jobban utal a módszer lényegére.

Saját vizsgálataink célkitűzése, hogy a szabadon hozzáférhető „R" programcsomag szolgáltatásait felhasználva egy- és többváltozós PLS regresszióra alapozott kalibrációs egyenleteket dolgozzunk ki. Az eredményeinket össze kívánjuk hasonlítani a PLS regresszióval nagyfokú rokonságot mutató PCR (főkomponens) regresszió által szolgáltatott modellekkel is.

ANYAG ÉS MÓDSZER

Vizsgálataink alapját 29 mangalica sertésből származó húsminta reflexiós spektruma és kémiai elemzés eredményeként kapott beltartalmi értékek (szárazanyag-, zsír- és fehérjetartalom) képezik.

Az állatokat hagyományos takarmányozási és tartási körülmények között hizlalták, a vágáskori átlagos testtömeg 157 kg volt. 24 órás hűtést követően a bal oldali hosszú hátizom (*m. longissimus dorsi*) utolsó bordatájékról származó szelete (kb. 100 g) került vizsgálatra. Minden mintát gondosan megtisztítottunk a kötőszövettől, hogy csak az intramuszkuláris zsírtartalommal kelljen számolni. Az egyes mintákat IKA A11 basic berendezéssel homogenizáltuk majd Christ Alpha fagyasztva szárítóval liofilizáltuk.

A fagyasztva szárított (liofilizált) mangalica húsminták közeli infravörös vizsgálatát NIRSystem 6500 (Foss NIRSystem, Silver Spring, MD, USA) spektrométerrel végeztük el. A reflexiós spektrumokat az 1100-2500 nm-es tartományban rögzítettük (log 1/R), 2 nm-es lépésközzel. „Small ring cup" mintatartó küvettát (IH-0307) és „Sample transport" egységet használtunk a vizsgálat során. A műszer üzemeltetéséhez és az elsődleges adatkezeléshez a WinISI II version 1.5 szoftvert alkalmaztuk (InfraSoft International, Port Matilda, PA, USA). A küvettákat minden minta után elmostuk, majd szárazra töröltük.

A kémiai analízis során a liofilizált minták szárazanyag-tartalmát az MSZ ISO 1442 szabvány, a zsírtartalmat *Folch és mtsai* (1957) szerint határoztuk meg. Sósavas emésztést és Kjel-Foss Fast Nitrogen Analyzer készüléket alkalmazva a nitrogén tartalom meghatározására; a nitrogén tartalmat 6,25-dal szorozva fejeztük ki a fehérjetartalmat. A beltartalmi értékeket (zsír- és fehérjetartalom) 100% szárazanyagra vonatkoztatva adtuk meg.

A PLS regresszió matematikai-statisztikai modelljét *Siesler és mtsai* (2002) és *Mevik és Wehrens* (2007) nyomán foglaljuk össze. A statisztikai modell a 29 elemű minta három függő változója ($Y_{(29\times3)}$) és a spektrumonként 700 reflexiós értéket jelentő független változók ($X_{(29\times700)}$) között teremt kapcsolatot a következő formában:

$$Y = XB + \varepsilon \text{, ahol } \varepsilon \text{ a véletlen hibák mátrixa.} \quad (1)$$

A legkisebb négyzetek elvén alapuló lineáris regresszió módszerével az ismeretlen B általában meghatározható:

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

Sajnos (2)-ben szereplő $X^T X$ a NIR spektroszkópia esetében rendszerint nem invertálható a szinte mindig fellépő multikollinearitási problémák miatt.

A PCR és PLS regresszió úgy kerüli meg ezt a problémát, hogy mátrixok szorzatává bontja fel X-et (3). T ortogonális oszlopvektorokból álló, úgynevezett látens komponensek mátrixa, P pedig az ún „loading" mátrix. Másképpen felírva (4) a komponensek oszlopait megkaphatjuk az X és a W súlymátrix szorzataként.

$$X = TP \tag{3}$$

$$T = XW \tag{4}$$

A mennyiben T meghatározásra kerül, az első néhány oszlopa alkalmas arra, hogy Y függő változóra regressziós egyenletet határozzunk meg (4).

$$Y = TQ + \mathrm{E}\text{, ahol E a véletlen hibák mátrixa.} \tag{5}$$

T meghatározásához a PCR és PLS regresszió egymástól eltérő további követelményt támaszt. A főkomponens regresszió (PCR) a T varianciáját maximalizálja (6).

$$\mathrm{var}(T) = \frac{1}{n} W^T X^T XW \tag{6}$$

Ugyanakkor a PLS regresszió olyan T komponenseket állít elő, melyek a $Y^T T$ kovarianciája maximális (7), vagyis a PLS regresszió a komponensek meghatározásakor figyelembe veszi a regressziós egyenlettel közelítendő függő változó tulajdonságait is.

$$\mathrm{cov}(Y^T T) = \frac{1}{n} W^T X^T YY^T XW \tag{7}$$

A PLS regresszió általában kedvezőbb tulajdonságokat mutat, mint a PCR. Szélsőséges esetben elképzelhető, hogy főkomponens regresszió végzése közben a T első néhány komponensének megtartása mellett olyanokat is elhagyunk, melyek elsődlegesek Y meghatározásában.

Az „R" nyílt forráskódú statisztikai szoftvercsomagot alkalmaztuk a számítások elvégzésére. Az „R" moduláris felépítésű, független szerzők járulnak hozzá a fejlesztéséhez. A PSL, PCR regressziót tartalmazó csomag *Ron Wehrens* és *Bjørn-Helge Mevik* munkája (*Mevik és Wehrens*, 2007). A pls csomag egyaránt alkalmas a téma szakirodalmában gyakran PLS1 és PLS2 elnevezéssel illetett modellek paramétereinek meghatározására. A PLS1 és PLS2 modell között az alapvető különbség az, hogy a T komponenseinek meghatározását csak egy függő változó, vagy egy időben az összes függő változó figyelembe vételével végezzük.

EREDMÉNY ÉS ÉRTÉKELÉS

A multi-kollinearitás mértékének szemléltetésére a 29 spektrum (*1. ábra*) esetére meghatároztuk a kétváltozós lineáris korrelációs együttható értékét $X_{(29\times700)}$ oszlop-szomszédai között. A *2. ábrán* feltüntettük a kiszámított együtthatókat. A rendkívül magas értékek ($r > 0.9990$) igazolják számunkra, hogy jogosan vetettük el a legkisebb négyzetek módszerét használó többváltozós lineáris regressziós modellt.

Mivel a kémiai analízisből származó függő változók között szoros a korrelációs kapcsolat (*1. táblázat*), elsőként a három tulajdonság egyidejű becslésére alkalmas PLS2 modellt állítjuk elő.

A PLS2 modell által magyarázott variancia mértéke egyre nagyobb attól függően, hogy NIR spektrumokból kivont komponensek (T első néhány oszlopvektora) közül hány kerül a kalibrációs egyenletbe (*2. táblázat*). A 3. táblázatban PCR modellel magyarázott variancia mértékeket tüntettünk fel. Érdemes összevetni a PLS2 és a PCR modell által szolgáltatott adatokat. Az előzetes várakozásnak megfelelően a NIR spektrumokat tartalmazó X varianciáját a PCR minden esetben jobban becsülte, mint a PLS2. A függő változók varianciáját viszont a *3. táblázatban* kiemelt három esettől eltekintve mindenhol a PLS2 magyarázta magasabb mértékben.

1. ábra

29 mangalica sertésből vett húsminta NIR spektruma

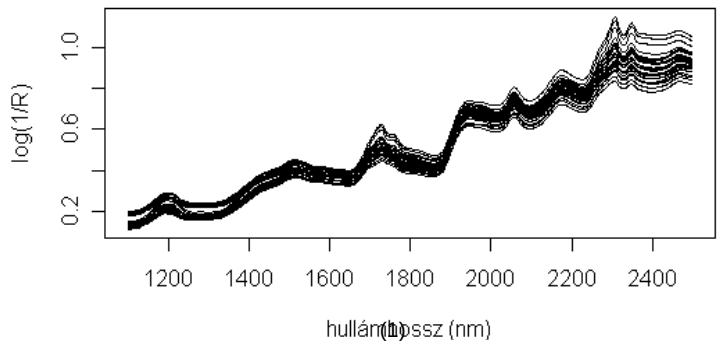


Figure 1: NIR spectra of 29 mangalitza pig meat sample.

Wavelength(1)

2. ábra

Korreláció az egyes NIR spektrumok szomszédos értékei között

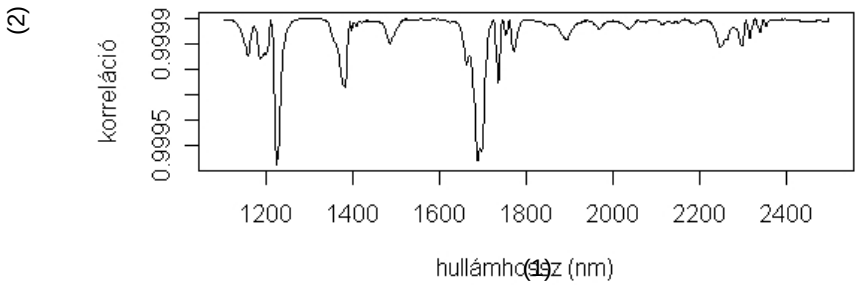


Figure 2: Correlation coefficients between the neighbouring values of NIR spectra

Wavelength(1), Correlation coefficients(2)

1. táblázat

Korreláció a szárazanyag-, zsír- és fehérjetartalom között

	szárazanyag-tartalom(1)	**zsírtartalom(2)**
zsírtartalom(2)	0,839	
fehérjetartalom(3)	-0,850	-0,999

Table 1: Correlation coefficients between the dry matter, fat and protein content of the meat samples

Dry matter(1), Fat content(2), Protein content(3)

2. táblázat

PLS2 modell által magyarázott variancia mértéke. A kalibrációs egyenlet egytől hatig növekvő számban tartalmazza a NIR spektrumokból származó komponenseket

	1 komp. (4)	**2 komp.**	**3 komp.**	**4 komp.**	**5 komp.**	**6 komp.**
X NIR	76.88	92.86	99.23	99.67	99.84	99.88
Y1 sz.a. (1)	77.26	90.13	91.64	93.58	93.64	93.68
Y2 zs. (2)	97.01	97.86	98.48	99.04	99.31	99.52
Y3 f. (3)	97.31	97.95	98.54	99.00	99.37	99.58

Table 2: Variance explained by fitted PLS2 model with 1 to 6 components

Dry matter(1), Fat content(2), Protein content(3), Number of component(4)

3. táblázat

PCR modell által magyarázott variancia mértéke. A kalibrációs egyenlet egytől hatig növekvő számban tartalmazza a NIR spektrumokból származó komponenseket

	1 komp. (4)	**2 komp.**	**3 komp.**	**4 komp.**	**5 komp.**	**6 komp.**
X NIR	76.89	95.18	99.26	99.71	99.85	99.90
Y1 sz.a. (1)	**78.22**	85.23	**92.09**	93.17	93.60	**93.70**
Y2 zs. (2)	96.71	97.17	98.35	98.70	99.12	99.17
Y3 f. (3)	97.05	97.37	98.44	98.67	99.16	99.24

Table 3: Variance explained by fitted PCR model with 1 to 6 components

Dry matter(1), Fat content(2), Protein content(3), Number of component(4)

Az egyváltozós PLS1 modell a szárazanyag varianciájának becslésében szembetűnően kedvezőbb eredményeket szolgáltat, mint a háromváltozós PLS2 (*2.* és *4. táblázat*). A zsír- és fehérjetartalom esetében ezt nem jelenthetjük ki.

A komponensek optimális számának megállapítására keresztvalidációt végeztünk. A háromváltozós PLS2 modell esetében meghatározott keresztvalidációs hibákat a *3. ábrán* találhatjuk. A keresztvalidációs hiba (CV) nem más, mint a becslési hibák négyzetösszegeinek átlagából vont négyzetgyök (RMSEP). Torzítatlan formában is (adjCV) megtalálható az ábrán. Mivel a keresztvalidációt jelen modellnél egy-egy minta figyelmen kívül hagyása jelenti a CV és adjCV megegyezik. Az ábrán megfigyelhetjük, hogy a keresztvalidációs hiba minimumát szárazanyagtartalom esetében a 4 komponenst tartalmazó modell szolgáltatja. A szárazanyagtartalom varianciájának magyarázata (*2. táblázat*) az 5 és 6 komponenst tartalmazó modellben már nem növekszik jelentősen, viszont a CV igen. A zsír- és fehérjetartalom esetében a feltüntetett hat komponens is csökkenő keresztvalidációs hibát találhatunk (*3. ábra*). A három függő változót egy logikai egységként kezelve kijelenthetjük, hogy a 4 komponenst tartalmazó modell megfelelő, különös tekintettel arra, hogy a zsír- és fehérjetartalom esetében a variancia magyarázat a 99.0% eléri illetve meghaladja.

4. táblázat

A három független változóra egyenként létrehozott PLS1 modell által magyarázott variancia mértéke. A kalibrációs egyenlet egytől hatig növekvő számban tartalmazza a NIR spektrumokból származó komponenseket

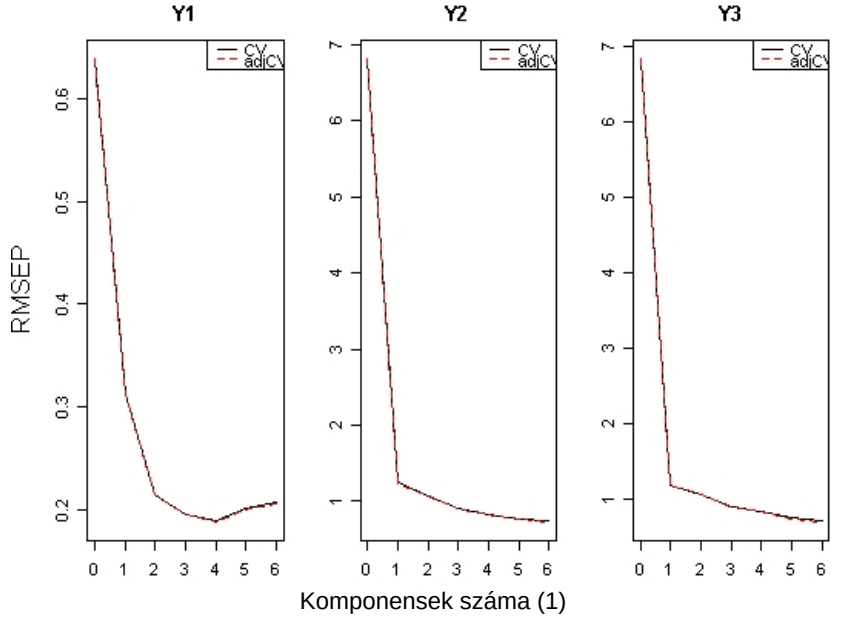
	1 komp. (4)	2 komp.	3 komp.	4 komp.	5 komp.	6 komp.
X NIR	76.58	94.18	99.24	99.70	99.83	99.87
Y1 sz.a. (1)	81.92	88.73	92.44	93.67	94.80	97.08
X NIR	76.87	92.85	99.23	99.68	99.84	99.88
Y2 zs. (2)	97.03	97.86	98.49	99.01	99.32	99.54
X NIR	76.88	92.29	99.23	99.65	99.84	99.89
Y3 f. (3)	97.30	98.02	98.55	99.05	99.37	99.57

Table 4: Variance explained by the three fitted univariate PLS1 model with 1 to 6 components

Dry matter(1), Fat content(2), Protein content(3), Number of component(4)

3. ábra

A háromváltozós PLS2 modell keresztvalidációs eredményei. Y1=szárazanyag-tartalom, Y2=zsírtartalom, Y3=fehérjetartalom



The cross validation results of the three-variable PLS2 model. Y1=dry matter, Y2=fat content, Y3=protein content the NIR spectra neighbouring values

Number of component(1)

KÖVETKEZTETÉSEK

A 29 mangalica húsmintából származó NIR spektrumok és a kémiai analízis eredményeit feldolgozva egy- és többváltozós kalibrációs modelleket készítettünk PLS és PCR regresszió segítségével. Megállapíthatjuk, hogy minden modell igen magas arányban (93.50 … 99.04%) magyarázza a függő változók varianciáját. A keresztvalidációs eredmények vizsgálata azt mutatja, hogy a 29 minta alapján, PLS2 modellel négy komponensre alapozva robusztus kalibrációs egyenletek készíthetők. A statisztikai elemzéshez felhasznált „R" programcsomag megfelelő grafikus és numerikus outputot szolgáltatott a következtetések levonásához.

IRODALOMJEGYZÉK

Folch, J.M., Leeas, M., Sloane-Stanley, G.H. (1957): A simple method for the isolation and purification of total lipids from animal tissues. In: J. Biol. Chem. 226. 495-509. p.

Mevik, B.H., Wehrens, R. (2007): The pls Package: Principal Component and Partial Least Squares Regression In: R. Journal of Statistical Software, 18. 2.

R (2007): A Language and Environment for Statistical Computing. [online] <http://www.R-project.org> [2007 dec. 10.]

Siesler, H.W., Ozaki, Y., Kawata, S., Heise, H.M. (2002): Near-Infrared Spectroscopy. Weinheim : Wiley-VCH GmbH, 132-136. p.

Wald. S., Sjöström, M., Eriksson, L. (2001): PLS-regression: a basic tool of chemometrics. In: Chemometrics and Intelligent Laboratory Systems 58. 109-130. p.

Levelezési cím (*Corresponding author*):

Kövér György
Kaposvári Egyetem, Gazdaságtudományi Kar
Matematika és Fizika Tanszék
7401, Kaposvár, Pf. 16.
Kaposvár University, Faculty of Economic Science
Department of Mathematics and Physics
H-7401, Kaposvár, POB 16.
Tel.: 36-82-505-956
e-mail: kovergy@ke.hu