Least squares means multiple comparison testing of reference versus predicted residuals for evaluation of partial least squares spectral calibrations

J.B. Reeves, III,[a] S.R. Delwiche[b] and V.B. Reeves[c]

[a]*USDA, ARS, Beltsville Agricultural Research Center, Environmental Management and By-Product Utilization Laboratory, Bldg 306, BARC East, Beltsville, MD 20705, USA*

[b]*USDA, ARS, Instrumentation and Sensing Laboratory, Bldg 303, BARC East, Beltsville, MD 20705, USA*

[c]*CVM, FDA, HHS, 7500 Standish Place, Rockville, MD 20855, USA*

It is common to use data pre-treatments such as scatter correction, derivatives, mean centring and variance scaling prior to the development of near- and mid-infrared spectral calibrations. As a result, it is possible to generate a multitude of calibrations, many of which will have similar statistical properties, as measured by the coefficient of determination (R^2) and the residual error. With respect to validation data sets, calibration equations have a tendency to provide the most optimistic modelling statistics on the set of data to which they were developed; however, the pre-treatment that was optimal for one set of samples may not be the best for future samples and, therefore, not the most robust calibration. If several calibrations are found to be statistically the same, then other criteria could be used to determine which one to use (for example, one with fewest partial least squares (PLS) factors or based on past experience) or further investigations could be carried out on the more limited set of calibrations deemed to represent the best of all those originally developed. However, there has been no single accepted statistical procedure for determining which calibrations are statistically the same and which are not from a large group of calibrations. This study describes the use of least squares means multiple comparisons testing of squared reference-versus-predicted residuals for determining the statistical similarity of multiple PLS calibrations. A program has been developed using common commercial statistical software (SAS, Mixed procedure) which computes and summarises comparisons of PLS calibrations. This method is also applicable to other multivariate regression methods as it only requires a list of reference and predicted values from each calibration as input.

Keywords: **AUTHOR PLEASE SUPPLY UP TO TEN KEYWORDS**

Introduction

Traditionally, in multivariate regression, the residual modelled error [for example, the standard error of calibration (*SEC*); the standard error of cross-validation (*SECV*); the standard error of performance (*SEP*)] is used to select the "best" model. Comparisons may be among different numbers of latent variables, different pre-treatments, or different model types altogether [partial least squares (PLS) or principal component regression (PCR), artificial neural networks (ANN), etc.]. In the first case, commercial multivariate analysis programs often utilise a heuristic test to balance the choice between the lowest standard error with a model that possesses a slightly larger error, but with fewer factors. Recent efforts have been made to codify these selection procedures in a statistical basis. Indahl and Næs[1] developed a two-way analysis of variance (ANOVA) procedure that uses the squared or absolute values of the residuals, with sample number (a random effect) as one level and method type (six model types examined) as the second level. This procedure was used to evaluate six different transformations that reduced two-dimensional image data to one-dimensional spectra. Non-parametric test evaluation of models was suggested by Thomas,[2] in which the signs of the differences of corresponding sample residuals of a pair of models were used, with and without ranking of the difference, to compare models with a varying number of latent variables. Another procedure evaluates cross-validation error by

The authors wish to state the use or mention of any commercial software does not imply any endorsement of that product or any better suitability than other similar products by either the authors, the US Department of Agriculture or NIR Publications.

© NIR Publications 2006, ISSN 0967-0335

determining an acceptance region for which errors of models that are simpler than a model of minimum error are evaluated by a chi-square distribution.[3] In a recently published study, the ANOVA and Wilcoxon signed-rank tests, along with another non-parametric test, the Friedman test, were compared using four unrelated spectral datasets, with the general conclusion that the ANOVA test on the absolute value of residuals was the most suitable method for evaluating models.[4]

The present report demonstrates the application of the ANOVA test in the evaluation of spectral pre-treatments in near infrared PLS modelling. Spectral pre-treatments can greatly improve spectral calibration results by removing artifacts such as baseline shifts and tilts caused by scatter variation due to particle size effects,[5,6] or by accentuating the differences between spectra as in the use of derivatives.[7] While such spectral pre-treatments can be very helpful, it is generally difficult or impossible to know *a priori* what specific pre-treatment will be the most helpful unless a data set is very similar to other data sets previously investigated and for which the best pre-treatments have been determined by experience. In addition, different spectral pre-treatments often produce very similar results statistically, despite being based on unrelated mathematical transformations or using dissimilar numbers of PLS or PCR factors. These events are demonstrated by the data presented in Table 1 for the determination of lignin[8] in a set of 241 forage samples.[9,10] The calibrations were run twice using a random division of samples into calibration and validation sets. The results are presented by increasing R^2 for the

Table 1. Partial least squares regression results for 22 spectral pre-treatments for lignin for two random runs, calibration set = 161 randomly selected samples out of 241, remaining 80 used as test or validation set.

			Cal-R^2 Order		# Factors		R^2-Run1		R^2-Run2	
Pre-treatment			Run1	Run2	Run1	Run2	CAL	VAL	CAL	VAL
NON	MSC	0	1	12	8	13	0.901	0.877	0.940	0.828
1ST	MSC	64	2	2	12	11	0.904	0.897	0.915	0.836
1ST	STR	16	3	6	9	9	0.916	0.921	0.930	0.876
NON	STR	0	4	7	11	13	0.918	0.904	0.933	0.892
2ND	STR	32	5	10	11	12	0.923	0.915	0.936	0.860
2ND	STR	64	6	8	13	13	0.923	0.913	0.934	0.847
1ST	STR	32	7	5	14	11	0.929	0.920	0.926	0.856
1ST	STR	64	8	18	14	15	0.931	0.917	0.949	0.887
1ST	STR	8	9	15	10	11	0.937	0.925	0.946	0.895
2ND	MSC	32	10	11	14	13	0.939	0.919	0.937	0.857
2ND	MSC	64	11	13	15	14	0.941	0.906	0.941	0.836
1ST	MSC	32	12	4	15	10	0.945	0.915	0.919	0.850
2ND	MSC	4	13	21	7	7	0.946	0.886	0.957	0.794
1ST	MSC	8	14	3	12	8	0.946	0.909	0.919	0.871
1ST	MSC	16	15	1	15	8	0.951	0.901	0.910	0.872
1ST	STR	4	16	17	13	10	0.953	0.922	0.947	0.868
1ST	MSC	4	17	19	12	11	0.953	0.914	0.952	0.806
2ND	STR	4	18	22	8	9	0.955	0.891	0.969	0.833
2ND	STR	16	19	16	13	10	0.955	0.925	0.946	0.852
2ND	MSC	16	20	9	13	8	0.957	0.905	0.935	0.840
2ND	STR	8	21	20	14	9	0.971	0.924	0.953	0.832
2ND	MSC	8	22	14	14	8	0.971	0.900	0.945	0.787

Pre-treatment = gap derivative (1st or 2nd, scatter correction)
MSC = multiplicative scatter corrected
STR = none, gap in data points
CAL = calibration
VAL = validation

calibration set based on the first run. As demonstrated, simply choosing the calibration and validation sets randomly can easily result in radically different results. Of the best four calibration set results, only the pre-treatment 2ND STR 8 (second derivative, gap width = eight intervals) is found as one of the top four for both calibration sets. Running many different randomly chosen calibration/validation splits and comparing the results to find the pre-treatment which consistently gives good results is a possible approach for comparing a few pre-treatments. However, combining such random runs with the many thousands of possible spectral pre-treatment combinations[11–13] and such non-statistical measures as the number of PLS or PCR factors, and then trying to determine what pre-treatment is empirically the best, may well be prohibitive. Even running several hundred potential pre-treatments on many different randomly determined calibration/validation datasets can be impossible for a dataset of even modest size (200–300 samples) from the practical standpoint of computer time required.

Therefore, there is the need for a statistical method to at least select a limited number of spectral pre-treatments for further study or selection based on non-statistical parameters such as the fewest factors required. The objective of this work was to develop a statistical method based on least-squares means multiple comparisons testing of reference-versus-predicted residuals for the evaluation of PLS spectral calibrations using SAS.

Experimental

Samples

Three different datasets were utilised in the testing of the statistical methods. Set 1 consisted of 241 forage samples composed of 174 chlorite treated forages (16 different forages and by-products, derived from alfalfa, tall fescue, orchard grass, red clover, timothy, wheat straw, barley straw, corn cobs, corn stover and soybean stover) and 67 forages (derived from alfalfa, tall fescue, orchard grass and vegetative parts of corn and wheat) collected at different stages of growth. Further information on these samples may be found in the literature.[9,10] Approximately two-thirds of the samples ($n = 161$), drawn by random selection, were used for calibration development, while the remaining third ($n = 80$) was reserved for model validation. For the present study, lignin[8] was chosen as the analyte to model, but similar results were obtained with other analytes.

Set 2 consisted of 398 samples of hard red winter or hard white wheat from one season of the Nebraska Winter Wheat Variety Tests programme. These were either commercial releases or breeders' advanced lines grown in field-replicated plots throughout ten locations across the state of Nebraska. A constant set of ten red and ten white wheats was grown at each location. One field replicate ($n = 198$, two missing) was used for calibration development, while the other replicate ($n = 200$) was reserved for model testing. One analyte, protein content, was considered in this study. We (AUTHOR INSERT REFERENCE PLEASE) have previously reported on the implementation of a comparison of correlated variances procedure, using this set, with pair-wise tests of $RMSD_{\text{test set, no pretreatment}}$ vs $RMSD_{\text{test set, a pretreatment}}$.[11]

Set 3 consisted of durum wheat breeders' lines from two seasons. These lines were part of a programme for the development of amylose-free (i.e. waxy) durum wheat (recalling that starch is composed of two macromolecules: amylose, which is the linear chain of C1–C4 alpha-linked glucose units and amylopectin, which is the highly branched structure consisting of C1–C6 linked branchings in addition to the C1–C4 linkages). Forty-seven lines from one season constituted one sub-set, with the same lines, replanted and harvested the next season in field-replicated plots, constituting a second sub-set. Samples used for the NIR study were gathered from each field replicate of the second season, thus producing about twice the number of scanned samples ($n = 95$) compared to the first season. The larger sub-set was used for calibration equation development, while the smaller sub-set was used for model validation.

Statistical analysis

All regression analyses were performed by PLS, using the two PLS programs previously published in this journal[12] and discussed in *NIR news*.[13] The code for these programs is available on the NIR Publications journal website.[14]

Least squares means multiple comparison testing

Analysis of means was performed using a mixed models procedure (i.e. Proc Mixed) (two-way ANOVA without interaction) in SAS.[15] In addition, an SAS macro program,[16] which is available on the web[17] and used to produce a more readable output from the comparison testing, was incorporated with minor changes to fit the programs as used here. Results using PROC GLM were almost always identical to those using Proc Mixed, but the latter procedure is known to be better behaved for these types of analyses. All SAS codes for running the programs discussed in this paper, including an updated version of the SAS PLS, are available from the authors and will eventually be available at this journal's website. They are described below.

For each pre-treatment, the regression model's absolute or squared residuals [|predicted–reference| or (predicted–reference)^2] of the regression model from the cross-validation set or the validation set became the random variable in ANOVA. The individual samples were treated as blocks (random effect) and the pre-treatments (fixed effect) were tested collectively for significant differences. Analytes that demonstrated a significant difference among pre-treatments were further tested using the means comparison procedure, with a probability level for type I error (α) set to 0.05. Because the number of pair-wise comparisons increases nearly as much as the square of the number of pre-treatments, the occurrence of at least one type I error is also greatly enhanced. In general, statistical corrections, such as those attributed to Bonferroni, are often imposed for the purpose of making

the comparison of means test more conservative. With the Bonferroni correction, the probability of a type I error is maintained at the probability level for the main effect of the ANOVA by dividing α by the number of means.[18] In this study, we also explored the use of the Bonferroni correction on the lignin content models.

Two general forms of the means comparison routine are available. The first is embedded in a large program that also reads spectral and chemical data, applies spectral pre-treatments, and performs PLS regression. The second is a stand-alone routine, in which the regression trials are required to have been performed beforehand. As such, the second program reads in the model predictions for each sample, for each pre-treatment.

For the forages dataset, 22 pre-treatments were compared. These included all combinations of derivative order (none, first, or second), multiplicative scatter correction[5] (MSC, applied or not applied) and gap width for the central difference-based derivative approximations (4, 8, 14, 32 or 64 points). For the wheat datasets, the following pre-treatments were examined: none, MSC, standard normal variate[6] (SNV) and detrend (with detrend being a baseline correction formed by subtracting a least squares quadratic regression function from each SNV-transformed spectrum), Savitzky–Golay (SG) smooth (quadratic polynomial), SG first derivative (cubic polynomial) and SG second derivative (quadratic polynomial). These were the same pre-treatments used in our earlier paper.[11] Briefly, two window widths, 11-point and 21-point, were used for each type of SG convolution. Mean centring was performed on all pre-treatments immediately before the PLS procedure. Altogether, ten pre-treatments were examined, with the "none" pre-treatment acting as the control to which all other pre-treatments were statistically compared in our previous work. In the current work, the comparison of *RMSD*s-by-ANOVA procedure allows for the comparison of all pairs of pre-treatments, instead of the former restriction of one pair member always being the control. This procedure was separately used on the cross-validation and validation sets.

Results and discussion

Absolute value of/or squared residuals

Initial efforts were carried out using either the squared differences between the predicted and reference values or the absolute value of the difference. Results (data not shown) indicated that the squared differences were better at evaluating calibrations and were used in the results that follow. Using the absolute value weights and all analyte values equally can produce very different results than might be expected by examining plots of predicted-versus-reference values for calibrations because of the influence of extreme values in the domain which can greatly affect slope. Using a weighted difference [|(Predicted–Reference)|/Reference] similarly gave unexpected results. However, comparisons between results produced using various modifications to the basic residual expression (Predicted–Reference) might be useful and a subject of future efforts.

The modelling results for wheat protein content are summarised in Table 2. As a whole, the PLS regression equations

Table 2. Summar of pre-treatment trials for PLS regression modelling of wheat protein content ($N \times 5.7$ as a percentage).

	Cross-validation			Validation			
Pre-treatment	R^2	RMSD	Group	r^2	Bias	*RMSD*	Group
21-pt 1st der	0.997	0.111	A	0.997	0.002	0.132	A
11-pt sm	0.998	0.110	A	0.997	–0.010	0.125	AB
21-pt sm	0.998	0.110	A	0.997	0.008	0.125	AB
None	0.998	0.109	A	0.997	0.013	0.125	AB
21-pt 2nd der	0.998	0.107	A	0.997	–0.005	0.127	AB
11-pt 1st der	0.998	0.098	B	0.997	0.013	0.117	B
11-pt 2nd der	0.998	0.095	BC	0.997	0.008	0.120	B
SNV w/det	0.998	0.092	BC	0.998	0.012	0.099	C
MSC	0.998	0.086	CD	0.998	0.015	0.099	C
SNV	0.999	0.081	D	0.998	0.006	0.098	C

Pre-treatment abbreviations
x-pt = number of points in a Savitzky–Golay convolution window
None = no pre-treatment
SNV = standard normal variate transformation
w/det = SNV with detrending
MSC = multiplicative scatter correction
Groups with the same letter are deemed not statistically different at the $P = 0.05$ level.

for this analyte were extremely accurate and probably more so than would normally occur during calibration development. This is because of, first, the careful attention paid to equilibration of the samples to a common ambient humidity level before scanning and, second, the very high precision (repeatability) of the reference, combustion procedure, having a standard deviation of repeated measurement of 0.05 to 0.10 protein percentage units. A comparison of the $RMSD_{xval}$ values among the pre-treatments indicated the superiority of the SNV transformation with respect to all other pre-treatments except MSC. These two pre-treatments, along with SNV and detrend, were also superior in the comparison of $RMSD_{val}$ values. When compared with correlated variances test,[11] the three pre-treatments were consistent in being significantly different from the control (i.e. no transformation) pre-treatment. A visual representation of the two extreme pre-treatments (none and SNV) is shown in the cross-validation residual graphs in Figure 1. The overall similarity of the two graphs, with one pre-treatment showing just a slightly tighter band around zero, is indicative of the subtleness of the means comparison test.

The results of the pre-treatment comparison for the durum wheat amylose models are contained in Table 3. Unlike protein content, there was a noticeable difference between cross-validation and validation error terms. Recalling that the cross-validation and validation samples were from different harvest years and also measured one year apart, it is not surprising that biases as large as 10 percentage units occurred. Because of such differences, pre-treatments that tended to show the lowest error in cross-validation, such as the control (no pre-treatment), produced some of the poorer validation set errors. Conversely, the pre-treatment with the highest cross-validation error, a 21-point second derivative, produced a validation error that was the third lowest of the ten pre-treatments.

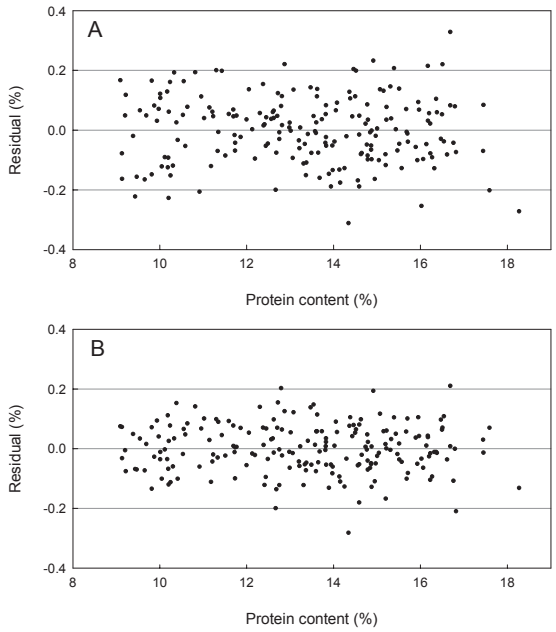

Figure 1. Plots of cross-validation residuals of PLS regression models for wheat protein content. (A = no pre-treatment and B = standard normal variate transformation pre-treatment).

Proc Mixed with and without Bonferroni correction

The Bonferroni (BON) adjustment places more stringent criteria for two means to be found to be different. Results for lignin, with and without this option (ADJUST = BON),

Table 3. Summary of pre-treatment trials for PLS regression modelling of durum wheat amylost fraction (amylose to total starch, as a percentage.

	Cross-Validation			Validation			
Pre-treatment	R^2	*RMSD*	Group	r^2	Bias	*RMSD*	Group
21-pt 2nd der	0.969	2.07	A	0.954	4.63	5.14	F
11-pt 2nd der	0.974	2.13	B	0.962	1.57	2.55	H
11-pt 1st der	0.977	2.05	BC	0.957	6.31	6.67	E
SNV w/det	0.976	2.02	BC	0.939	2.89	3.84	G
21-pt 1st der	0.977	2.01	BC	0.952	3.56	4.22	G
SNV	0.977	1.97	BC	0.946	10.66	10.94	A
21-pt sm	0.978	2.00	CD	0.934	9.92	10.30	B
MSC	0.981	1.95	CD	0.959	6.75	7.07	DE
11-pt sm	0.980	1.94	CD	0.947	8.38	8.74	C
None	0.981	1.92	D	0.951	7.04	7.45	D

Pre-treatments defined in Table 2
Groups with the same letter are deemed not statistically different at the $P = 0.05$ level.

demonstrated (data not shown) that the use of the BON adjustment made the criteria so stringent that all the calibrations were often found to be statistically the same, while without the adjustment, this was rarely true. While results were not always found to be dramatically different for pre-treatments with and without the BON option, they were also rarely the same and varied greatly on the nature of the random split into calibration and validation sets. Only more efforts with many datasets can determine which method should be used.

Multiple tests based on randomly generated datasets

Running calibrations many times on randomly selected calibration/validation set splits avoids (1) problems associated with atypical validation sets if only one is randomly selected, (2) complaints that selecting samples by clustering methods to produce matched calibration and validation sets predisposes the process to produce the best possible results, but calibrations which may not be the best for new samples and (3) problems similar to the second point, if all the samples available are used in the calibration set. By running the results many times, calibrations should be selected that will work on new samples as well as can be determined based on the samples at hand.

```
DATA PREDICTACTUAL;
INFILE "COMPLETECOMBINEDPREDICTACTUALRUN5.SAV";
INPUT CHEM ID $ DER $ MSC $ MC $ GAP CALVAL $ PRED ACT RUNNUM;
CARDS;

DATA NEW; SET PREDICTACTUAL;
IF CALVAL = "VAL";
DIFF = (PRED - ACT);
SQRDIFF = DIFF*DIFF;

PROC SORT DATA = NEW;
   BY RUNNUM CHEM DER MSC MC GAP;

PROC TRANSPOSE DATA = NEW OUT = SQUARED;
   BY RUNNUM CHEM DER MSC MC GAP;
   VAR SQRDIFF;

DATA OUTDIFF; SET SQUARED;
FILE "OUTDIFFS.SAV";
PUT RUNNUM 1-3 CHEM 5-7 DER $ 9-11 MSC $ 13-15 MC $ 17-19 GAP 21-24
      COL1-COL80 RUNNUM;

DATA ALL; INFILE "OUTDIFFS.SAV";
INPUT RUNNUM 1-3 CHEM  $ 5-7 TREATMENT $ 9-24@;
   DO BLOCK = 1 TO 80;
   INPUT SQRDIFF @;
   OUTPUT;
   END;

PROC SORT DATA = ALL;
   BY CHEM TREATMENT;

PROC MIXED DATA =  ALL; CLASS TREATMENT BLOCK;
   BY CHEM;
   MODEL SQRDIFF = TREATMENT;
   RANDOM BLOCK;
   LSMEANS TREATMENT / PDIFF ADJUST = BON;
   ODS OUTPUT DIFFS=PPP LSMEANS=MMM;
   ODS LISTING EXCLUDE DIFFS LSMEANS;

%INCLUDE 'PDMIX800.SAS';
%PDMIX800 (PPP, MMM,α=.01, SORT = YES)
   RUN;
```

Figure 2. Typical code for performing "least squares means multipole comparison testing" using Proc Mixed on squared residuals from predicted minus reference.

The statistical assumptions of the ANOVA are not strictly satisfied when the same sample appears several times in the analysis as it would for repeated random draws. Some investigations (not shown) suggest that the approach still gives useful answers in this case, but further work is probably needed before it can safely be recommended.

SAS code for Proc Mixed

Figure 2 contains the SAS program used in this study. This program assumes that the cross-validation or validation sample predictions from the various pre-treatments have been developed beforehand. For use with different data sets, the primary changes needed relate to the input dataset's pre-treatments. That is, output from the SAS PLS program contains information on the spectral pre-treatment used, which other input data may not contain. The rest of the code is considered standard except for the "PDMIX800". This is an SAS macro program made available on the web.[17]

Conclusions

A method has been developed for grouping calibrations by use of the squared residuals from the predicted versus reference. This method can be applied to residuals from any calibration method to statistically compare calibrations as to being statistically the same or different. Finally, results have demonstrated the method to work on residuals compiled from multiple runs of the same spectral pre-treatments, but using randomly selected calibration/validation data sets.

References

1. U.G. Indahl and T. Næs, *J. Chemometr*. **12**, 261 (1998).
2. E.V. Thomas, *J. Chemometr*. **17,** 653 (2003).
3. U. Indahl, *J. Chemometr*. **19,** 32 (2005).
4. H.R. Cederkvist, A.H. Aastveit and T. Næs, *J. Chemometr*. **20,** in press (2006).
5. P. Geladi, D. MacDougall and H. Martens, *Appl. Spectrosc*. **39,** 491 (1985).
6. R.J. Barnes, M.S. Dhanoa and S.J. Lister, *Appl. Spectrosc*. **43,** 772 (1989).
7. A. Savitzky and M.J.E. Golay, *Anal. Chem*. **36,** 1627 (1964).
8. P.J. Van Soest and R.H. Wine, *J. Assoc. Off. Anal. Chem*. **51,** 780 (1968).
9. J.B. Reeves, III, *J. Dairy Sci*. **71,** 143 (1988).
10. J.B. Reeves, III, *J. Dairy Sci*. **71,** 388 (1988).
11. S.R. Delwiche and J.B. Reeves, III, *J. Near Infrared Spectrosc*. **12,** 177 (2004).
12. J.B. Reeves, III and S.R. Delwiche, *J. Near Infrared Spectrosc*. **11,** 415 (2003).
13. J.B. Reeves, III and S.R. Delwiche, *NIR news* **15,** 10 (2004).

14. Anonymous. http://www.nirpublications.com/software/index.html
15. R.C. Littell, G.A. Milliken, W.W. Stroup and R.D. Wolfinger, *SAS System for Mixed Models*. SAS Institute Inc., Cary, NC, USA, 633 pp. (1996).
16. A.M. Saxton, *Proc. 23rd SAS Users Group Intl*, Nashville, TN, 22–25 March. SAS Institute Inc., Cary, NC, USA, pp. 1243–1246 (1998).
17. Anonymous. http://animalscience.ag.utk.edu/faculty/saxton/software.htm
18. P.H. Westfall, R.D. Tobias, R.D. Wolfinger and Y. Hochberg, *Multiple Comparisons and Multiple Tests using the SAS System*, SAS Institute Inc., Cary, NC, USA, 416 pp. (1999).

Received: 22 December 2005
Revised: 6 July 2006
Accepted 8 July 200
Web Publication: