Split-sample correlation of lab data sets upper limit on near infrared correlation

Terry Watkins
Mathematics Department, University of New Orleans, New Orleans, LA 70148, USA.

Joe Montalvo and Bryan Vinyard
USDA, ARS, Southern Regional Research Center, PO Box 19687, New Orleans, LA 70179, USA.

Robert Grimball
1715 Sandra Ave, Metairie, LA 70003, USA.

Steve Buco
Statistical Resources, 7338 Highland Rd, Baton Rouge, LA 70808, USA.

Introduction

Cotton is a hollow fiber whose wall is composed primarily of cellulose. Maturity (wall thickness) and fineness (cross-sectional perimeter) are important indicators of fiber quality in the marketing of cottons. The evolution of cotton near infrared (NIR) reflectance spectroscopy for maturity and fineness has been severely limited by the lack of suitable lab data diagnostics. The standard methods for wall thickness and perimeter, for example, give poor precision and also, the instrument readings drift over time.[1]

A review of the NIR reflectance spectroscopy literature indicates lab data diagnostics has been limited to the standard deviation of the means. Previous work related specifically to the use of split sample methodology for estimating trends and an upper bound for the R^2 between the lab data and NIR does not appear in the statistical literature, although the methods expected to be required appear to be rather standard in mathematical statistics.

The objective of this research is the development of the theory and computational methods required for the use of split samples for (i) estimating the components of variation, including both random and systematic components, encountered in the analysis of cotton fiber property data and (ii) the prediction of an upper bound for the coefficient of determination in regression analysis of that data. Our goal is to develop easy to use, mathematically sound, diagnostic tools that give clear insight into data quality when multiple measurements are available. The methods may also be used effectively for evaluating both existing and new methods of measurement of cotton fiber properties as well as for comparing different methods of measurement.

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996

Significance of research

When limited replicate measures of a dependent variable are made, random and systematic errors may affect the coefficient of determination, R^2, a standard measure of predictability associated with regression analysis. This research provides methods for detecting and estimating systematic errors and also provides estimates of random error. In addition, the research would allow researchers to monitor data quality while work is in progress and, perhaps, to take corrective measures to improve data quality. Since some minimum standard of predictability is usually required, or at least desired, this work, by allowing the researcher to obtain preliminary estimates of an upper bound for the coefficient of determination before completion of data collection, may make it possible for the researcher to modify his/her experimental design to achieve improved results.

It is sometimes the case that a coefficient of determination near 1 does not yield the desired predictability. There are situations when systematic errors in measurement (even small ones) may actually increase the coefficient of determination thus giving deceptive results. In other situations systematic errors do not effect the coefficient of determination at all, but in regression situations produce models which predict the wrong thing. In either case, the usefulness of the regression model is jeopardized without knowledge of the presence, or absence, of systematic error in data. Another application of this research is in the evaluation of new instruments or methods for data collection.

Prior and current research

SRRC has identified three critical parameters that impact the generation of cotton lab data by improved or new analytical instrumentation. These are (i) the random measurement error of the instrument, (ii) the systematic measurement error of the instrument and (iii) the inherent heterogeneity of raw cotton. These errors result in errors in the mean values of measurements of a property of cotton fiber based on a fixed number of replications.

Consider the linear regression of mean values of the dependent variable Y on the independent variable X, where Y is a cotton fiber property and X is either a fibre property or reflectance spectra. In the context of this discussion, the values of X may be either a single determination or the mean of several determinations. When Y is predicted by a linear function of X, the most common measure of the quality of the predictor is given by the coefficient of determination, R^2. We have developed test statistics to probe the error in the mean Y values *a priori* of the linear regression of Y on X. The errors may be random, systematic, or both. These statistics are then used to estimate the maximum possible coefficient of determination, R^2_{MAXREG}, in the regression of Y on X.

In brief, the replicate Y values on each sample are split into halves, the means computed for each half and the corresponding means correlated to produce two measures of split sample correlation, R_{SPLIT} and M_{SPLIT}. R_{SPLIT} is the Pearson correlation coefficient and M_{SPLIT} is a measure of the fit of the paired means to the line $Y = X$. (A more complete description of M_{SPLIT} is given in the Appendix.) R^2_{MAXREG} is estimated from the split sample correlations and is an estimate of the maximum possible coefficient of determination, assuming an ideal regression model and no errors in the X values. In effect, R^2_{MAXREG} sets the upper limit on the coefficient of determination between the lab data sample means and the independent variable.

The theory has been confirmed by computer simulations and testing on cotton data. Some of the computer simulation results are presented in Table 1 for $N = 1000$ cotton samples in each of four data sets with hypothetical wall thickness and perimeter values. First, "true" values were simulated for wall thickness and perimeter to represent the independent X variable. Next, the "real" replicate Y values were simulated to give 20 values with both random error and systematic error

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996

for each corresponding true value. The replicates represent the dependent variable whose mean values are regressed on the true values (the X variable).

The replicates are then split into halves and R_{SPLIT} and M_{SPLIT} computed. From these test statistics R^2_{MAXREG} is computed. R^2_{TRUE} is computed by regressing the sample mean real values on the true values. The theoretical (asymptotic) value of R^2_{MAXREG} should exceed the theoretical value of R^2_{TRUE}, so that ideally, $R^2_{TRUE} \leq R^2_{MAXREG}$. However, due to the variance of the estimators, it is possible that R^2_{TRUE} may exceed R^2_{MAXREG} by some small amount. In our simulation we see excellent agreement between predicted and observed coefficients of determination for both wall thickness and perimeter (See Table 1 below.). The value of R^2_{TRUE} is slightly larger than expected in the data set 22 but this is probably explained by the large systematic errors in that case.

Finally, based on spectral modeling, the true values were transformed to an equivalent multivariable NIR spectra. Note that when the sample mean real values are regressed on the NIR spectra using a linear regression technique called partial least squares (PLS, PRESS by one-out-rotation), R^2_{NIR} is approximately equal to R^2_{TRUE}, as it should be, since the spectra are free from error. Again, R^2_{MAXREG} is in close agreement with R^2_{NIR} (see Table 1.). Table 1 also shows that when there is a systematic trend in replicates, then M_{SPLIT} is significantly less than R_{SPLIT} (data sets 12 and 22) and when there is no systematic trend then $M_{SPLIT} \approx R_{SPLIT}$.

Table 1. Results on simulated cotton data (N = 100) with error in wall thickness and perimeter values.

Data Set	Error[a]	R^2_{SPLIT}	M^2_{SPLIT}	R^2_{MAXREG}	R^2_{TRUE}	R^2_{NIR}
Wall thickness (T)						
11	A	0.695	0.694	0.907	0.898	0.896
12	A+B[b]	0.668	0.061	0.9	0.9	0.898
21	A+C[c]	0.785	0.785	0.563	0.56	0.555
22	A+B+C[d]	0.796	0.218	0.519	0.549	0.537
Fiber perimeter (P)						
11	A	0.986	0.985	0.996	0.996	0.996
12	A+B[b]	0.985	0.915	0.995	0.996	0.996
21	A+C[c]	0.992	0.992	0.611	0.596	0.594
22	A+B+C[d]	0.99	0.942	0.592	0.6	0.599

[a]Error code: A = random, B = horizontal trend and C = vertical trend.
[b]B = 0.05 μm between reps for T and P (0.05 × 100/3.34 = 1.5% and 0.05 × 100/47.52 = 0.1%) where 3.34 μm and 47.52 μm are the population mean values, respectively, for wall thickness and perimeter.
[c]C = 0.001 μm between samples for T and 0.005 μm between samples for P.
[d]B = 0.05 μm between reps for T & P; C = 0.001 μm between samples for T and C = 0.005 μm between samples for P.

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996

Examples

The following are examples of what has been found to date with cotton fiber property data intended for use as the dependent variable in linear regressions:

1. Collaborator: Henry Perkins, ARS, Clemson. Fiber property: stickness in cotton. Results: split sample correlations and preliminary estimates of an upper bound for the coefficient of determination, R^2_{MAXREG}, were extremely poor. Henry is attempting to reduce the error in the sample mean values by additional measurements on each sample to achieve improved results.
2. Collaborator: Devron Thibodeaux, ARS, New Orleans. Fiber properties: maturity and fineness by image analysis. Results: preliminary estimates of R^2_{MAXREG} indicated the need to revise the methodology to provide improved data.
3. Collaborator: Stuart Gordon, ARS, New Orleans. Fibre properties: maturity (wall thickness) and fineness (perimeter) by the Advanced Fiber Information System (AFIS) and the Shirley Development Limited Fineness and Maturity Tester (FMT, Micromat model).

FMT results were based on six replicates on each sample ($N = 80$). Differences between R_{SPLIT} and M_{SPLIT} did not suggest a significant drift in instrument readings during data collection and was confirmed by examination of data from a quality control cotton dispersed in the sample set. For perimeter, $R^2_{SPLIT} = 0.928$, $M^2_{SPLIT} = 0.927$, $R^2_{MAXREG} \approx 0.981$, and $R^2_{NIR} = 0.939$. The R^2_{NIR} was the result of the PLS algorithm, PRESS with one-out-rotation. In another experiment with fresh specimens from the same samples, FMT results were based on 20 replicates on each sample. Differences between R_{SPLIT} and M_{SPLIT} did not suggest a significant drift in instrument readings during data collection. For wall thickness, $R^2_{SPLIT} = 0.981$, $M^2_{SPLIT} = 0.981$, $R^2_{MAXREG} \approx 0.995$ and $R^2_{NIR} = 0.96$. The R^2_{NIR} was the result of the PLS algorithm, PRESS with one-out-rotation.

AFIS results were also based on 20 replicates on each sample ($N = 80$). Differences between R_{SPLIT} and M_{SPLIT} did not suggest a drift in instrument readings during data collection and was confirmed by examination of data from a quality control cotton dispersed in the sample set. R^2_{MAXREG} was ≈ 0.95 indicating wall thickness and perimeter sample means not confounded by error. However, the sample means gave very poor correlations with the NIR, observed $R^2 < 0.7$, suggesting the AFIS measures different fineness and maturity characteristics of the fiber compared to the FMT.

Some preliminary data diagnostics

Based on the very limited experience with the split sample methodology obtained to this point, we feel that the following statements reflect, in a general way, conclusions regarding trends in the lab data, the upper boundary for R^2 and regression model clues. (In the inequality statements below < and > refer to significant differences.)

1. If $M_{SPLIT} \approx R_{SPLIT} \approx 1$ and $R^2_{MAXREG} \approx R^2_{NIR} \approx 1$, then the regression model is ideal for prediction.
2. If $R^2_{MAXREG} < R^2_{NIR}$, there are two possible explanations. First, a trend between samples will decrease R^2_{MAXREG} so that, in fact, if no such trend exists then R^2_{MAXREG} is negatively biased. (R^2_{MAXREG} is computed from the split sample methodology.) Second, overfitting of the NIR algorithm will make $R^2_{MAXREG} < R^2_{NIR}$.
3. If $R^2_{MAXREG} < 1$, then there is significant error in the lab data sample means that will confound the predictability of the model.
4. If $M_{SPLIT} \approx R_{SPLIT} \approx 1$ and $R^2_{MAXREG} \approx 1 > R^2_{NIR}$, there are two possible explanations. There may be error in the NIR spectra or the regression model is inadequate (i.e. wrong algorithm or wavelength range).
5. If $M_{SPLIT} < R_{SPLIT}$ and $R^2_{MAXREG} \approx R^2_{NIR}$, then there is a significant trend between replicate observations and the regression model is unreliable for prediction even if $R^2_{NIR} \approx 1$.

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996

6. If $M_{SPLIT} \approx R_{SPLIT}$ and $R^2_{MAXREG} \approx R^2_{NIR} < 1$, then if the estimate of the trend between samples is significantly different from 0, then the regression model is unreliable for prediction even if $M_{SPLIT} \approx R_{SPLIT}$ and $R^2_{MAXREG} \approx R^2_{NIR}$.

Future work

The theoretical aspects of the research is expected to continue. As some of the anticipated theoretical results are of an asymptotic nature, the limits of those results will be explored via computer simulation. Although calculations for individual lab data sets are expected to be relatively easy and require little time, the extensive simulation required is expected to require a considerable amount of both programming time and computer time. Actual lab data from various sources will be collected to provide realistic tests of the performance of these test statistics. Requests for data have been made to numerous researchers in order to demonstrate the broad applicability of the split sample methodology.

References

1. J.G. Montalvo Jr and S. Faught, *Determination of maturity/fineness by FMT and Diode-Array HVI. Part I. FMT (Micromat Model) procedure optimization*, Ed by D.J. Herber and D.A. Richter. Proc. Beltwide Cotton Confs., Natl. Cotton Council Am., Memphis, TN, **2,** 1276 (1995).

Appendix

Suppose that an even number of measurements are made on each of a number of subjects (cotton samples). Then the sample from each subject may be split into two groups of equal size and the mean for each group computed. This leads to a collection of points (x_i,y_i), $i = 1, 2, \ldots, n$, where (x_i,y_i) represents the means of the split sample for the i th subject. One would typically expect that the points (x_i,y_i) would lie near the line $Y = X$. Consistent patterns of deviation from that line would suggest that systematic measurement errors may be present, while random deviations from the line would suggest that measurement errors are more likely to be purely random. As an initial measure of the relationship between X and Y, one is likely to consider the Pearson correlation coefficient. One quickly notes, however, that large systematic errors may actually increase the Pearson correlation while at the same time causing the points (x_i,y_i) to fall

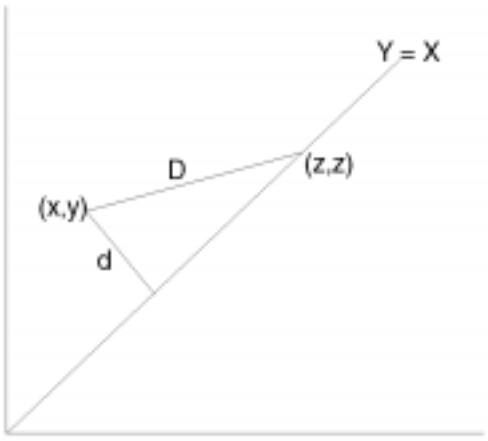

Figure 1.

© IM Publications Open LLP 1996

farther from the line $Y = X$. This suggests that an alternative measure of the relationship of the points (x_i,y_i) to the line $Y = X$ may be appropriate. The proposed alternative is described below. Let d_i denote the distance from the point (x_i,y_i) to the line $Y = X$ and let D_i denote the distance from (x_i,y_i) to the point (z,z), which lies on $Y = X$, where z denotes the mean of all observations over all subjects. Figure 1 illustrates the geometry. When (x_i,y_i) lies on the line $Y = X$ then $d_i = 0$, and when (x_i,y_i) lies on the line through (z,z) that is perpendicular to $Y = X$ then $d_i = D_i$. Otherwise, $0 < d_i < D_i$. The proposed measure of the relationship of the points (x_i,y_i) to the line $Y = X$ is defined by:

$$M = 1 - 2(\Sigma\, d_i^2)(\Sigma\, D_i^2)$$

One should note that $-1 \leq M \leq 1$, and thus, in this regard is similar to a correlation coefficient. Also, if $M = 1$ then all points lie on the line $Y = X$ and if $M = -1$ then all points lie on the line through (z,z) that is perpendicular to $Y = X$. It can be shown that, in addition to being intuitively appealing, M has a solid foundation in mathematical statistics and is, in fact, a correlation coefficient. Also, based on initial investigations, M is quite sensitive to systematic measurement errors.

From Near Infrared Spectroscopy: The Future Waves
© IM Publications Open LLP 1996