Validating Mathematical Models of Biological Systems: Application of the Concordance Correlation Coefficient

N. R. St-Pierre

The Ohio State University, 2029 Fyffe Rd., Columbus, OH-43210, USA
E-mail: St-Pierre.8@osu.edu

Abstract: The National Research Council is charged with producing mathematical models of nutrient requirements of domestic animals. In ruminants, protein supply is derived from two sources: a fraction of the feed protein unaltered by ruminal fermentation, and microbial protein (MiN) synthesized by the ruminal micro-flora. Measurements of MiN rely on surgically altered animals and inert markers. The prediction of MiN is based on total digestible nutrients, a function of the uncertain composition of feedstuffs. Both observed and predicted MiNs have errors from measurements, parameter estimates, and structural forms. The question is whether predicted MiN can replace measured values when estimating requirements. The concordance correlation coefficient (ρ^c) has been suggested as an omnibus statistic to jointly assess precision and accuracy. Application to a dataset of 256 measured and predicted values of MiN from 56 published studies shows that predictions and measurements are concordant ($\rho^c = 0.476$), have small scale shift (1.54) and location shift (-0.02), and are accurate (0.913) but that they lack precision (0.522). The deviance (0.573) is composed of a small bias (0.0003), a small scale shift (0.095), and a large imprecision (0.479). Little gain in model precision can be expected until more precise methods of measurements are found.

Keywords: concordance correlation coefficient, precision, accuracy, nutrient requirements

1. INTRODUCTION

Mathematical models are now frequently used to quantify complex biological systems [1, 2]. The validation of such models is done by comparing model predictions to observed data. Various statistical methods have been suggested and used to assess a model's validity: the Pearson correlation coefficient, the paired t-test, the least-square analysis of slope (=1) and intercept (=0), and the coefficient of variation or the intraclass correlation coefficient. None of these can completely assess the desired reproducibility characteristics. The Pearson correlation coefficient only measures precision of a linear relationship, not accuracy. Both the paired t-test and least squares analysis can falsely reject (accept) the hypothesis of high agreement when the residual error is small (large). The coefficient of variation and the intraclass correlation coefficient assume a dependent and an independent variable. More importantly, they fail to recognize the duality (interchangeability) of predictions with observations. Both are mathematical transforms of measurements. Both have random errors from measurements and parameter estimates. And both have structural errors due to the simplification of the complex real world. The relevant question is not whether a model predicts observed data but whether the model and the observation method measure the same thing, whether the methods agree and how good is the agreement. This requires a joint assessment of precision and accuracy.

The Committee on Animal Nutrition of the National Research Council (NRC, [3]) is charged with producing tables of nutrient requirements of various classes of animals. Nutrient requirements are expressed in the form of computerized mathematical models. In a recent publication, the NRC [3] produced a new model for estimating the nutritional requirements of dairy cattle. A key step in the calculation of protein and amino acid requirements is the estimation of the amount of bacterial protein synthesized in the rumen. In ruminants, the net supply of protein and amino acids is derived from two separate fractions: a variable portion of the feed protein not broken down by the ruminal micro-flora passes to the duodenum (small intestine) where it can be digested and absorbed by the animal. The second portion consists of microbial protein synthesized by the ruminal micro-flora using carbon skeletons, ATP, ammonia, amino acids, and short peptides. The quantification of the net supply from each process is very important to the optimal feeding of ruminant animals and for reducing their environment impact from N excretion [4]. The measurements of microbial and undegraded feed protein to the duodenum must rely on surgically altered animals and inert markers [5]. Thus, the measurements of microbial protein (MiN) and non-ammonia-non-microbial protein flows (NANMN) to the duodenum are subject to substantial errors of measurements, plus structural errors (i.e., the non-digestible markers are not perfect markers) and possibly errors in parameter estimates. The prediction of MiN is based on total digestible nutrient intake (TDN) which is a function of the (uncertain) chemical composition of the feedstuffs and their (uncertain) bio-availabilities. Thus, both observed and predicted MiN and NANMN have errors from measurements, parameter estimates, and structural forms. This situation, where predictions and observations are interchangeable is very frequent in biology. The question is whether we can use predictions of MiN and NANMN to replace measured values when estimating nutrient requirements.

In this paper, we first review the model used by NRC [3] to predict MiN in dairy cattle and the proper statistical model linking predictions to observations. Results from applying traditional methods of model validation are presented followed by the application of the concordance correlation coefficient (CCC) of Lin [6].

2. METHODOLOGY

2.1 Prediction of microbial protein synthesis by the National Research Council

In high producing ruminants, microbial protein synthesis is primarily determined by the availability of energy to the micro-organisms [7]. Although various expressions of available energy have been proposed and used to express the availability of feed energy for microbial growth, the total digestible nutrient (TDN) system is still favored in the U.S. due to the considerable literature reporting actual measurements in lactating and non-lactating animals. The measurement of TDN is a tedious process and requires urine and fecal collection in a digestibility study performed over several days (generally 5-7) with multiple animals. The TDN of a feed can also be estimated from its proximate composition using the following system of equations [8]:

$$\mathrm{TDN} = \mathrm{tdNFC} + \mathrm{tdCP} + \mathrm{tdFat} + \mathrm{dNDF} - 7, \qquad (1)$$

$$\mathrm{tdNFC} = 0.98 \times (100 - \mathrm{NDFn} - \mathrm{CP} - \mathrm{Fat} - \mathrm{Ash}) \times \mathrm{PAF},$$

$$tdCP = EXP(-1.2 \text{ x } (ADFIP / CP)) \text{ x } CP,$$

$$tdFat = (Fat – 1) \text{ x } 2.25,$$

$$dNDF = 0.75 \text{ x } (NDFn – L) \text{ x } [1 – (L/NDFn)^{0.67}] ,$$

$$NDFn = NDF – NDFIP,$$

where TDN is the estimated total digestible nutrients (%), tdNFC is true digestible non-fiber carbohydrates (%), tdCP is true digestible crude protein (%), tdFat is true digestible fat (%), dNDF is digestible neural detergent fiber (%), NDFn is NDF corrected for NDFIP (%), CP is the crude protein content (%), Fat is the fat content (%), Ash is the ash content (%), PAF is a processing adjustment factor, ADFIP is the acid detergent insoluble N x 6.25 (%), NDFIP is the neutral detergent insoluble N x 6.25 (%), L is the lignin content (%), and NDF is the neutral detergent content (%) of a given feedstuffs. Although the proximate composition (CP, Fat, Ash, etc.) is determined analytically in a laboratory, this is not done without analytical errors, which typically range between 2 and 10% of the true mean depending on the assay and feedstuff involved. Digestibility coefficients (e.g., 0.98, 0.75) are estimates subject to errors. Also, although the structure of the set of equations in (1) was derived mechanistically, it is nevertheless a simplification to the true, unknown, and far more complex system in nature. Thus, TDN values estimated using the system of equations in (1) are subject to measurement errors (feed composition), parameters in the equation are estimates (thus subject to errors), and the functional form itself is an approximation to the complex world.

In NRC [3], estimated TDN values from the set of equations in (1) are used to estimate MiN according to the following equation:

$$MiN = 130 \text{ x } TDN, \quad (2)$$

where MiN is net microbial protein synthesis (expressed in g of N/d). The coefficient 130 was estimated using an independent set of experimental data where both TDN and MiN had been measured. Clearly, it is an estimate also subject to error. By combining Eqs. (1) and (2), the NRC calculates the predicted MiN resulting from a given diet. This prediction is subject to measurement errors (feed composition), as well as errors in estimates of parameters (coefficients in Eqs. (1) and (2)), and errors in functional forms used.

Measurements of MiN are not without errors. Various experimental methods have been suggested in the scientific literature. All have limitations [7]. The prevailing method involves the marking of feeds and fluids with three indigestible markers each associating more predominantly with one of the three major digesta fractions (large particles, small particles, and fluid). Animals must be surgically altered with a large rumen cannula for the infusion or dosage of marker, and a duodenal cannula for sampling digesta leaving the stomach. Multiple samples are taken over time and the concentration of the three markers is then determined in a laboratory for each sample. Assuming first order, steady-state kinetics, forestomach digestibility of feed components can be calculated as well as flow of MiN [9] based on a marker of microbial protein (e.g., purines). It is clear that measured MiN are subject to considerable errors resulting from true measurement errors (concentrations of indigestible markers, concentration of microbial marker) as well as errors in parameter estimates, and error in the functional form (first-order, steady-state kinetics).

In this context, observations and predictions play a symmetric role because they are both functional transforms of other variables. This situation is actually quite frequent when modeling biological systems. The symmetric role of observations and predictions, however, has been largely ignored when models are being validated

2.2 Statistical Model

The following model, which naturally models comparison studies when both observations and predictions are subject to multiple errors, is commonly known as errors-in-variables regression [10, 11]:

$$\begin{aligned} X_i &= \xi_i + \delta_i, \\ Y_i &= \eta_i + \varepsilon_i, \qquad i = 1, \ldots, n, \\ \eta_i &= \alpha + \beta\xi_i, \end{aligned} \tag{3}$$

where X_i is the prediction from the mathematical model and Y_i is the observed value of the i^{th} observation, ξ_i and η_i are the unobserved mean parameters ("true values") of X_i and Y_i respectively, δ_i and ε_i are the errors on the predicted and observed values (generally assumed to be independent, bivariate Gaussian), α is the overall bias of the prediction model, and β is the linear scale difference between the predicted and the observed values. The variance of the two errors, σ^2_δ and σ^2_ε, are the precision parameters for the predictions and observations, respectively. With known or estimable σ^2_δ and σ^2_ε (or more accurately, an unbiased estimate of $\lambda = \sigma^2_\delta / \sigma^2_\varepsilon$), the maximum likelihood estimate of β is [11]:

$$\beta = \frac{S_{YY} - \lambda\, S_{XX} + ((S_{YY} - \lambda\, S_{XX})^2 + 4\,\lambda\, S^2_{XY})^{½}.}{2\, S_{XY}} \tag{4}$$

An estimate of σ^2_ε can be calculated from experimental data. Because of the nonlinearity of the system of equations in (1), an analytical estimate of σ^2_δ does not exist. Numerical methods could possibly be used but would require knowledge about the variances and covariances of all random variables in the equation. This information is currently not available.

2.3 Concordance correlation coefficient

Lin [6] proposed a statistic termed the concordance correlation coefficient (CCC) to evaluate the agreement (reproducibility) between two readings. In short, the degree of concordance between pairs of sample (Y_{i1}, Y_{i2}), i = 1, 2, …, n, can be characterized by the expected value of the squared difference, i.e.,

$$E(Y_1 - Y_2)^2 = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2 + 2(1 - \rho)\,\sigma_1\sigma_2, \tag{5}$$

where ρ is the Pearson correlation coefficient. This expectation also represents the expected squared perpendicular deviation from the 45° line, multiplied by 2. Standardizing both sides, we get:

$$\frac{E(Y_1 - Y_2)^2}{2\,\sigma_1\sigma_2} = \frac{(\mu_1 - \mu_2)^2}{2\,\sigma_1\sigma_2} + \frac{(\sigma_1 - \sigma_2)^2}{2\,\sigma_1\sigma_2} + (1 - \rho)\,, \tag{6}$$

which has a sample equivalent:

$$\frac{E(Y_1 - Y_2)^2}{(n\text{-}1)\ 2\ s_1s_2} = \frac{(Y_1 - Y_2)^2}{(n\text{-}1)\ 2\ s_1s_2} + \frac{(s_1 - s_2)^2}{(n\text{-}1)\ 2\ s_1s_2} + (1 - r). \tag{7}$$

In (7), a form that has been called deviance analysis, the total deviance, represented by the left-hand side is partitioned into three right-hand side components: bias (first term), scale difference (second term), and imprecision (third term). The deviance is equal to zero when all (non-negative) terms on the right-hand side are exactly zero, i.e., when the two means are equal, the two variances are equal, and the correlation is equal to 1.

The CCC is defined as follows:

$$\rho^c = 1 - \{E(Y_1\text{-}Y_2)^2 / E[(Y_1\text{-}Y_2) \mid Y_1, Y_2 \text{ are uncorrelated}]\}, \tag{8}$$

$$\rho^c = 2\ \sigma_{12} / [\sigma^2{}_1 + \sigma^2{}_2 + (\mu_1 - \mu_2)^2], \tag{9}$$

$$\rho^c = \rho_{12}\,\chi_{12}, \tag{10}$$

where $\mu_1 = E(Y_1)$, $\mu_2 = E(Y_2)$, $\sigma^2{}_1 = Var(Y_1)$, $\sigma^2{}_2 = Var(Y_2)$, and $\sigma_{12} = Cov(Y_1, Y_2) = \sigma_1\ \sigma_2\ \rho_{12.}$ The CCC is a product of two components: precision (ρ_{12}) and accuracy (χ_{12}), where $\chi_{12} = 2\ \sigma_1\ \sigma_2 / [\sigma^2{}_1 + \sigma^2{}_2 + (\mu_1 - \mu_2)^2] = [(\nu_{12} + 1/\nu_{12} + u^2{}_{12}) / 2]^{-1}$, with $\nu_{12} = \sigma_1 / \sigma_2$ representing scale shift, and $u_{12} = (\mu_1 - \mu_2) / (\sigma_1\ \sigma_2)^{1/2}$ representing location shift relative to the scale. The CCC is an omnibus statistic used to test simultaneously and jointly for accuracy and precision.

2.4 Dataset

The data used are described at length in the NRC publication [3]. In short, feed composition and measured MiN were gathered from 56 published, peer-reviewed studies of which 27 involved growing cattle and 29, lactating dairy cows. In total, the dataset comprised 256 records of observed MiN (oMiN, g/d) and predicted MiN (pMiN, g/d).

3. RESULTS

3.1 Pearson correlation

The Pearson correlation, which measures the degree of linear association (relationship) between two random variables has been used for comparing mathematical model predictions to observed values. In our application, this correlation is equal to: $r_{pMiN,\ oMiN} = 0.52$, $P < 0.0001$. This statistic shows that oMiN and pMiN have a significant association. The Pearson correlation, however, is invariant to location and scale. Agreement is a much more stringent concept than correlation because both the scale of the measurements and the slopes are important. Also, observations are not random samples from a population (i.e., the sample of observed and predicted values was not drawn at random from the population of all cows in the world). Thus, the Pearson correlation coefficient fails to determine whether pMiN and oMiN are equivalent.

3.2 Paired t-test

Applying the paired t-test on the data (mean oMiN = 244.91, mean pMiN = 246.36, $SE_{diff} = 4.16$, $t_{255} = -0.35$, $P = 0.73$), we conclude that there is no significant difference between the mean oMiN and the mean pMiN. This test provides information only for the overall bias (location shift). Because of its structure, the t-test can falsely reject the null hypothesis of high agreement when the residual error is small. That is, the larger the precision, the more likely you are to conclude that the two methods are not equivalent.

3.3 Least-squares analysis

The linear regression of oMiN on pMiN is presented in Figure 1. The model:

$$\text{oMiN} = B_0 + B_1\,\text{pMiN} + e \tag{11}$$

is theoretically incorrect because both oMiN and pMiN have errors. Under least-squares analysis, the null hypothesis is that the two methods are concordant. Thus, small datasets will generally lack power resulting in the conclusion that the two methods are concordant. Likewise, large datasets will result in rejecting the null hypotheses for the intercept ($B_0 = 0$) and the slope ($B_1 = 1$) when differences are relatively trivial. This is what occurs with the dataset at hand where the two null hypotheses are rejected. A casual inspection of the regression line in Figure 1 reveals the trivial difference between the regression line and the line of unity when the spread of the data points from either line is considered.

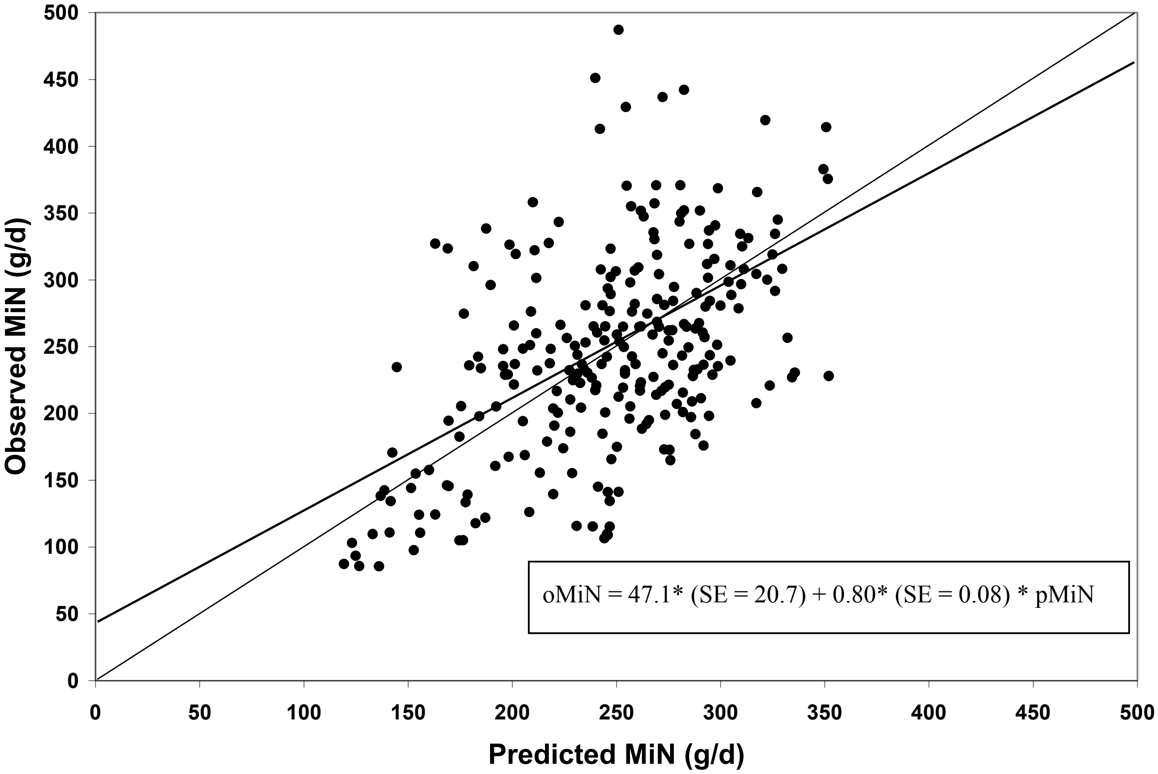

Figure 1. Linear regression of observed microbial flow to the duodenum (oMiN) on predicted microbial N flow (pMiN) using the National Research Council model.

This is quite clear when the differences between oMiN and pMiN are plotted against pMiN as in Figure 2. This plot, however, raises the legitimate question as to which variable should be used on the X-axis?

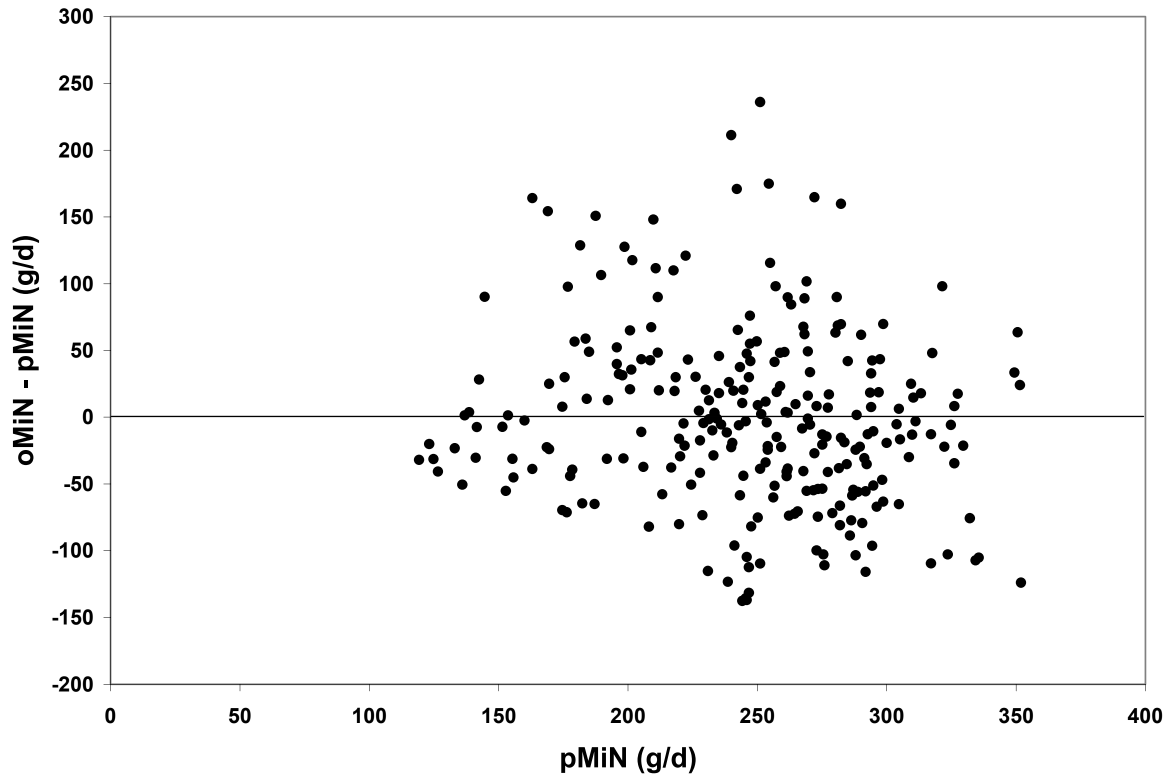

Figure 2. Plot of residuals vs. predicted microbial N flow to the duodenum (pMiN)) using the National Research Council model to calculate predicted microbial N flow (pMiN).

Recall that both oMiN and pMiN are measurement with errors. In Figure 2, pMiN was chosen on the X-axis because this is the correct variable to use in residual plots when the independent variable is assumed to be errorless, as in the linear regression paradigm [12]. Because of the duality of oMiN and pMiN, one could have chosen oMiN for the X-axis, resulting in a different conclusion regarding the presence or absence of bias (Figure 3).

Recognizing this problem, Altman and Bland [13] suggested using the mean of oMiN and pMiN for the X-axis (Figure 4). In fact, this is the correct axis if, and only if the precisions of both methods are equal (i.e., when $\sigma^2_\delta = \sigma^2_\varepsilon$, or simply that $\lambda = 1$). In the data at hand, however, the precision of pMiN is unknown. Thus the correct residual plot lies somewhere between the two extremes presented in Figures 2 and 3. Unless a satisfactory estimator for σ^2_δ can be identified, residual plots will invariably lead to the paradox depicted in Figures 2, 3, and 4, where one cannot decide whether a linear bias is present or not.

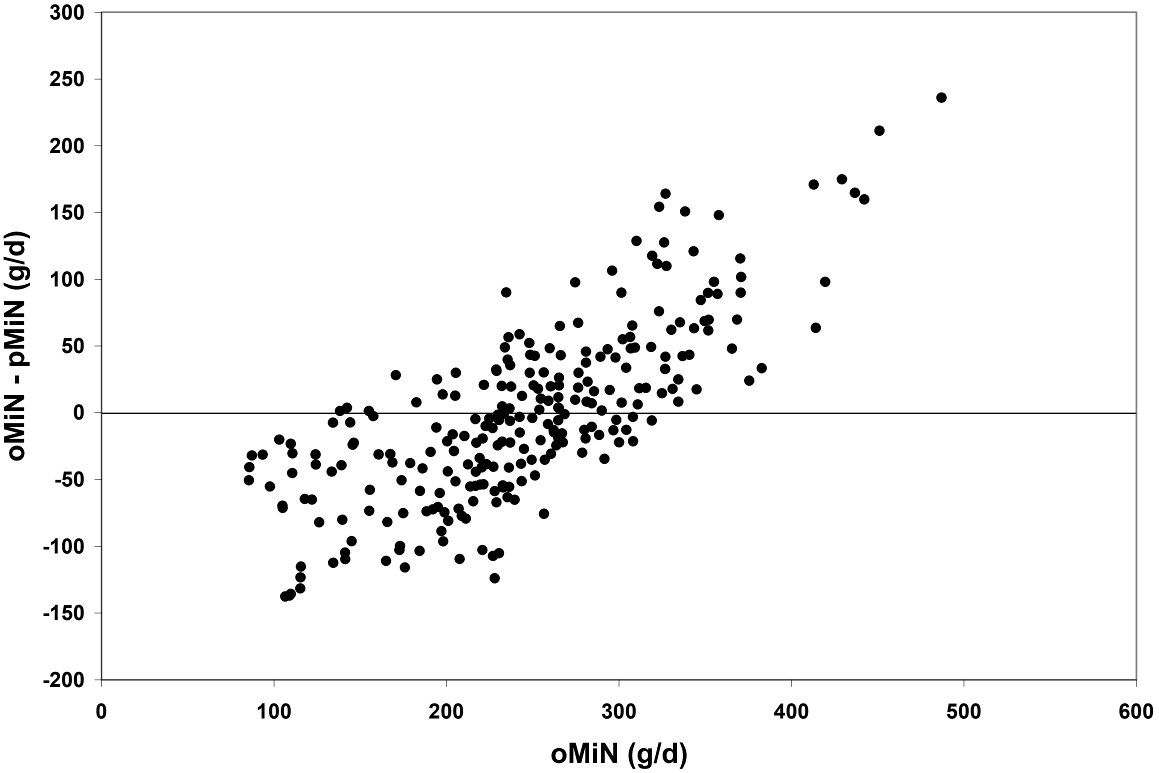

Figure 3. Plot of residuals vs. observed microbial N flow to the duodenum (pMiN) using the National Research Council model to calculate predicted microbial N flow (pMiN).

3.3 Deviance analysis

Application of equation (7) using the following estimates: s_1 = 50.16, s_2 = 77.23, s_{12} = 2020.2, mean (Y_1) = 246.4, mean (Y_2) = 244.9, and r = 0.522 (where the subscript 1 refers to pMiN and the subscript 2, to oMiN) results in the following:

$$\begin{array}{ccccccc} 0.5733 & = & 0.0003 & + & 0.0945 & + & 0.4785 \\ \text{Deviance} & = & \text{Bias} & + & \text{Scale difference} & + & \text{Imprecision} \end{array} \qquad (12)$$

The deviance is composed of a very small bias (0.0003; or 0.05% of the deviance), a small scale shift (0.0.95; or 16.5% of the deviance), and a large imprecision (0.479; or 83.5% of the deviance). Thus, it is clear that most of the deviance is the result of imprecision. The expression of deviance in (7) is in the form of the mean of squared deviations standardized by the product of standard deviations. The unit for deviance does not correspond to the unit of the physical variables being measured or predicted. Thus, although the method is useful, biologists struggle with the physical interpretation of the analysis. However, biologists are very familiar with the Pearson correlation coefficient, so that the expression of deviance re-scaled with a lower bound of -1 and an upper bound of 1 is certainly appealing. In essence, this is what is accomplished by the CCC.

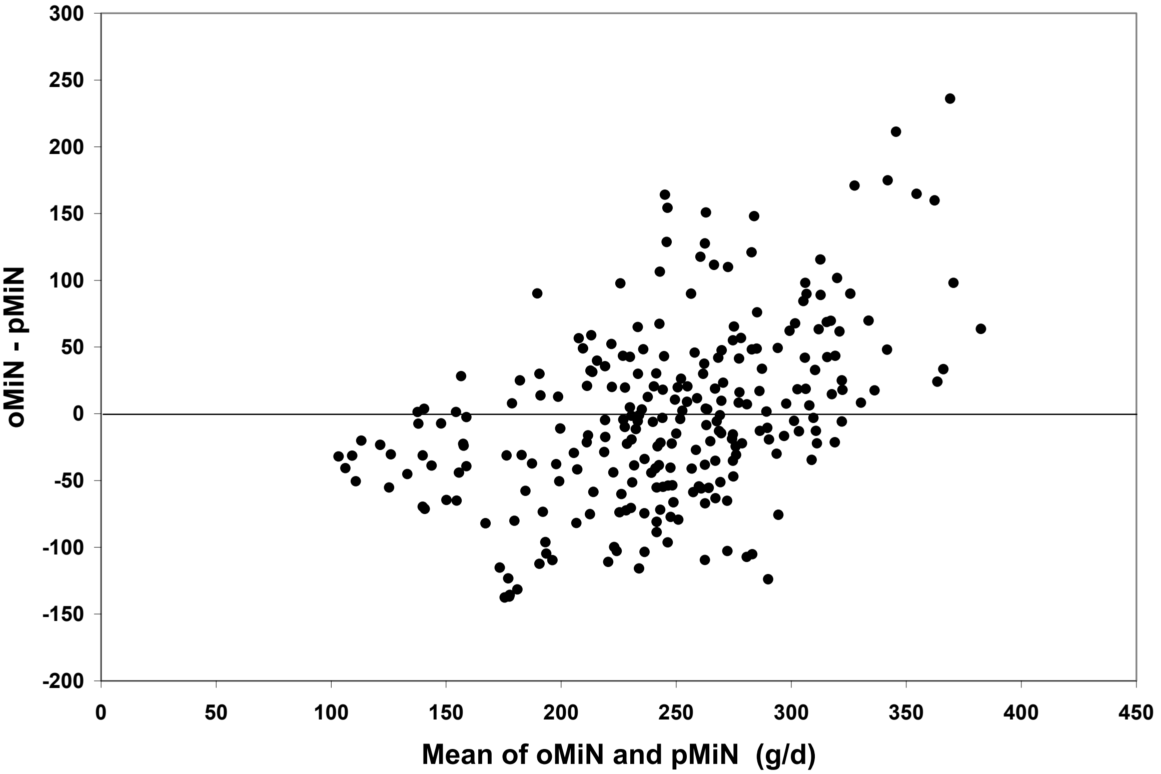

Figure 4. Plot of residuals vs. the mean of observed microbial N flow to the duodenum (pMiN) and predicted microbial N flow (PMiN) using the National Research Council model.

3.3 Concordance correlation coefficient

Application of equation (10) to our dataset results in $\rho^c = 0.476$. Using the inverse hyperbolic tangent transformation (or Z-transformation) suggested by Lin [6], and under the assumption of asymptotic normality, one concludes that predictions and measurements are concordant ($P = 0.22$). The accuracy statistic (χ_{12}) is equal to 0.913, whereas the precision statistic (ρ_{12}) is equal to 0.522. Recalling that $\rho^c = 0.476 = 0.913 \text{ x } 0.522$, it becomes evident that precision and not accuracy is the issue. The CCC is equal to 1 when there is no location differential, no scale differential, and perfect correlation between the two variables. It is an omnibus statistic that tests jointly precision and accuracy. In our application, measurements are too imprecise to allow the development of a model with acceptable prediction error. Thus, gains in the prediction of MiN can only be achieved with the development of superior methods of measurements, with much greater precision than the methods currently in use.

4. CONCLUSIONS

The validation of quantitative biological models is not a simple problem. Methods must account for the multiplicity of errors in both the observed and the predicted values. That is, methods must recognize the symmetric role of observations and predictions because both are algebraic transforms of other variables. The CCC shows potential in this regard.

REFERENCES

1. J. France, and J. H. M. Thornley. *Mathematical Models in Agriculture*. Butterworth, London, 1884.
2. J. D. Murray. *Mathematical Biology*. Springer, Berlin. 1993.
3. National Research Council. *Nutrient Requirements of Dairy Cattle.* 7th Ed. National Academy of Sciences, Washington. 2001.
4. N. R. St-Pierre, and C. S. Thraen. Animal grouping strategies, sources of variation, and economic factors affecting nutrient balance on dairy farms. *J. Dairy. Sci.* 82 (Suppl. 2):72-83, 1999.
5. G. J. Faichney. Digesta Flow. In J. M. Forbes, and J. France, editors, *Quantitative Aspects of Ruminant Digestion and Metabolism*, pages 53-86. CABI, Wallingford (UK), 1993.
6. L. I. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255-268, 1989.
7. J. L. Firkins, M. S. Allen, B. S. Oldick, and N. R. St-Pierre. Modeling ruminal digestibility of carbohydrates and microbial protein flow to the duodenum. *J. Dairy Sci.* 81:3350-3369, 1998.
8. W. P. Weiss, H. R. Conrad, and N. R. St-Pierre. A theoretically-based model for predicting total digestible nutrient values of forages and concentrates. *Animal. Feed Sci. Tech.* 39:95-110, 1992.
9. J. France, and R. C. Siddons. Determination of digesta flow by continuous marker infusion. *J. Theoretical Biology* 121:105-119, 1986.
10. G. Casella, and R. L. Berger. *Statistical Inference*. Wadsworth, Pacific Grove, CA, 1990.
11. C. Y. Tan, and B. Iglewicz. Measurement-methods comparisons and linear statistical relationship. *Technometrics* 41:192-201, 1999.
12. N. R. St-Pierre. Reassessment of biases in predicted nitrogen flows to the duodenum by NRC 2001. *J. Dairy Sci.* 86:344-350, 2003.
13. D. G. Altman, and J. M. Bland. Measurement in medicine: the analysis of method comparison studies. *The Statistician* 32:307-317, 1983.