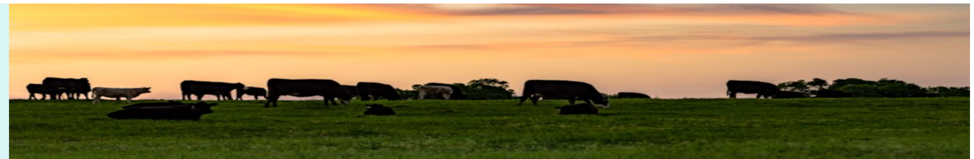
Interbull Meeting 2025

In conjunction with ADSA, Louisville, KY, USA



Reliability Assessment of Single-Milking Fat and Protein Percentages in Dairy Cattle

X.-L. Wu[1], M. J. Caputo[1], C. Donatone[1], A. Miles[2], R. L. Baldwin VI[2], S. Sievert[3], J. Mattison[3], J. Cole[1], J. Burchard[1], J. Joao[1]

[1] Council on Dairy Cattle Breeding; [2] USDA AGIL; [3] National DHIA

COUNCIL ON DAIRY CATTLE BREEDING

Motivations

- High-quality milk and milk component data are crucial for accurate genetic evaluations and daily herd management.
- We recently proposed using intraclass correlation (ICC) as a herd-level metric to assess the consistency of milk components from single milkings, thereby effectively identifying farms with potential data quality concerns.
- In the present study, we proposed a similar metric, namely individual ICC (I-ICC) for milk component data at the cow-day level.

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Data

- Real data: Four Holstein dairy farms (A, B, C, and D) practicing three times daily (3×) milking in three US states
- Simulated data: S0–S9, each containing 48,876 × 3 milking records under 0%–90% data shuffling.

COUNCIL ON DAIRY CATTLE BREEDING

Two types of outliers

- A univariate outlier is an extreme value in a single variable, such as a cow producing over 100 kg of milk daily when the population average is approximately 30 kg with a standard deviation of 10 kg.
- Multivariate outliers are observations that appear unusual only when considering multiple variables simultaneously. For example, a cow may produce a normal fat percentage in one milking but an abnormally low or high fat percentage in the other on the same test day.

COUNCIL ON DAIRY CATTLE BREEDING

Herd-level intraclass correlation (ICC)

- Consider the one-way random effects model:

$$x_{ij} = \mu + b_i + \epsilon_{ij}$$

- Within- and between-group variances are estimated by:

$$\hat{\sigma}_w^2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(x_{ij} - \bar{x}_i\right)^2; \quad \hat{\sigma}_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{x}_i - \bar{x})^2 - \frac{1}{m} \hat{\sigma}_w^2$$

- Population-level ICC (Wu et al., 2025):

$$\text{ICC} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \hat{\sigma}_w^2}$$

GENETICS | SHORT COMMUNICATION · Articles in Press, June 03, 2025 · Open Access Download Full Issue

Consistency assessment of milk fat and protein percentages across 3 daily milkings in Holstein and Jersey dairy herds

Xiao-Lin Wu [1,2] · Malia J. Caputo [1] · Asha M. Miles [3] · ... · John B. Cole [1,5,6] · Javier Burchard [1] · João Dürr [1] ... Show more

COUNCIL ON DAIRY CATTLE BREEDING

Individual-level intraclass correlation (I-ICC)

- The I-ICC replaces the global estimate of within-group variance ($\hat{\sigma}_w^2$) in ICC with its component corresponding to each cow-day group (s_i^2):

$$\text{I-ICC}_i = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + s_i^2}$$

where $s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_i\right)^2$.

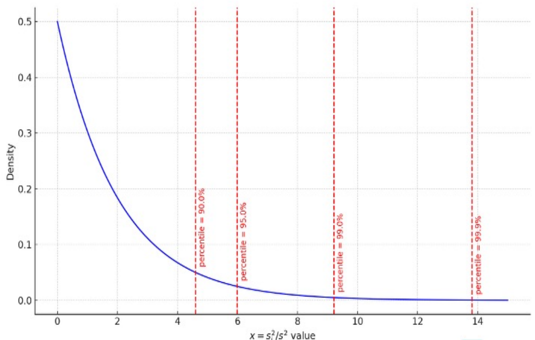
- Challenge: How to determine the cutoff threshold for I-ICC?

COUNCIL ON DAIRY CATTLE BREEDING

Decide the percentile threshold in two steps

$$s_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_i\right)^2 = \frac{\sigma_w^2}{m-1}\sum_{j=1}^{m}\left(\frac{x_{ij}-\bar{x}_i}{\sigma_w^2}\right)^2 \sim \frac{\sigma_w^2}{m-1}\cdot\chi_{m-1}^2$$

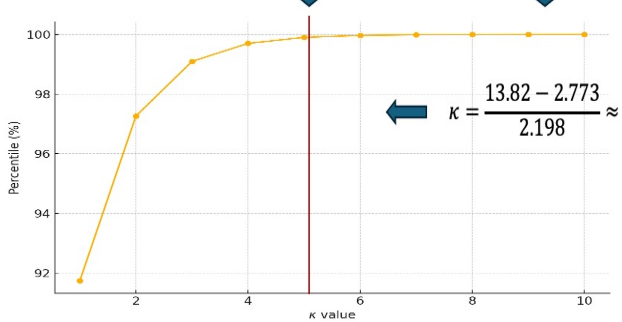
$$\text{I-ICC}_i < \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + s_{(Q99.9\%)}^2} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + (2.773 + 2.198\kappa)\cdot s^2}$$

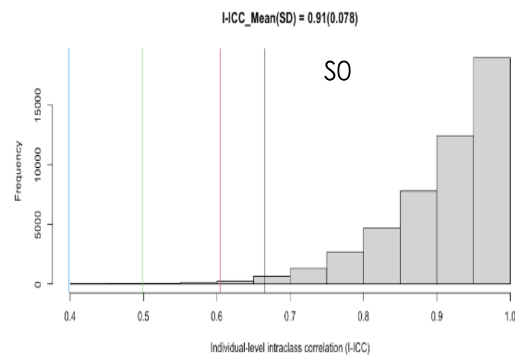


$$F_{\chi^2_{v=2}}\left(\frac{s_i^2}{s^2}\right) = \frac{s^2\cdot Q3 + \kappa\cdot s^2\cdot(Q3-Q1)}{s^2}$$
$$= Q3 + \kappa\cdot(Q3 - Q1)$$
$$= 2.733 + 2.198\kappa$$

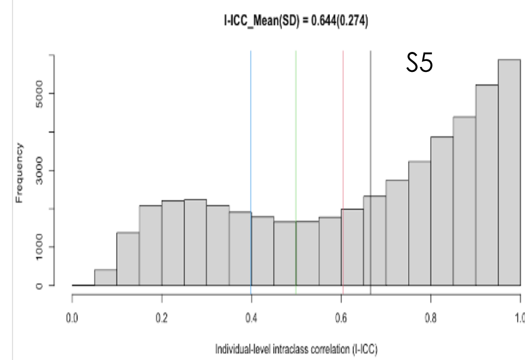
$$2.773 + 2.198\kappa = F^{-1}_{\chi^2_{v=2}}(0.999)$$
$$= 13.82$$

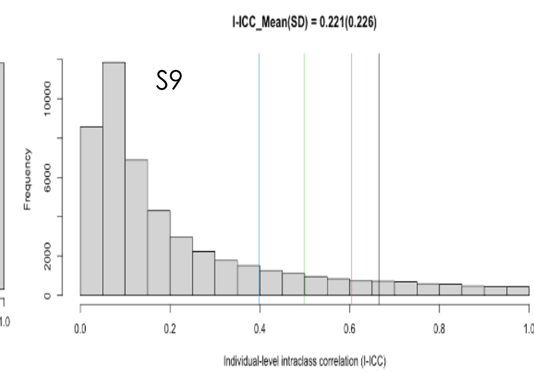
$$\kappa = \frac{13.82 - 2.773}{2.198} \approx 5.03$$





$$r_{(Q99.9\%)} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \left(Q3 + \kappa(Q3-Q1)\right)\cdot s^2}$$
$$= \frac{0.095}{0.095 + (2.773 + 5.023\times 2.198)\times 0.010}$$
$$= 0.407$$





Examples

$$F_{\nu}(2.773 + 2.198\kappa) = p\_value$$

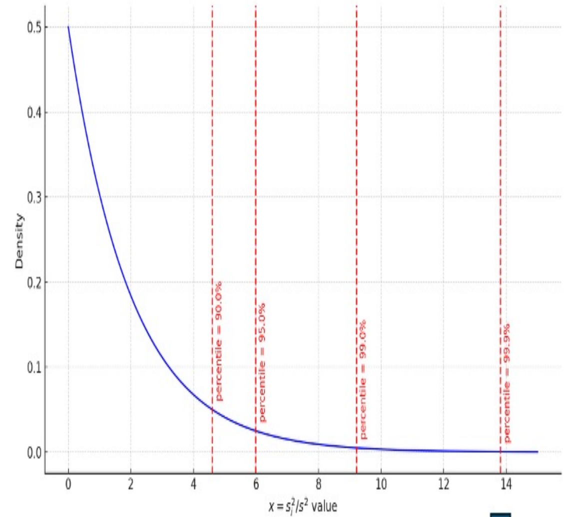
- For $\upsilon = 3 - 1 = 2$,

  90th percentile: $\kappa = 0.834$

  95th percentile: $\kappa = 1.464$

  99th percentile: $\kappa = 2.930$

  99.9th percentile: $\kappa = 5.023$



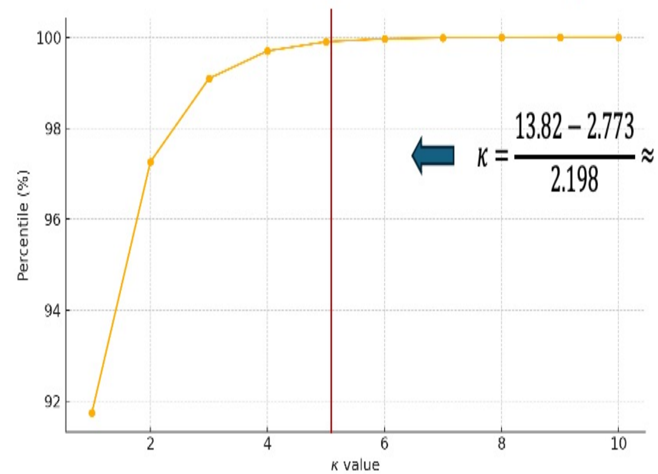
$$F_{\chi^2_{\upsilon=2}}\left(\frac{s_i^2}{s^2}\right) = \frac{s^2 \cdot Q3 + \kappa \cdot s^2 \cdot (Q3 - Q1)}{s^2}$$

$$= Q3 + \kappa \cdot (Q3 - Q1)$$

$$= 2.733 + 2.198\kappa$$

$$2.773 + 2.198\kappa = F^{-1}_{\chi^2_{\upsilon=2}}(0.999)$$

$$= 13.82$$



$$\kappa = \frac{13.82 - 2.773}{2.198} \approx 5.03$$

COUNCIL ON DAIRY CATTLE BREEDING

Baseline methods for comparison

- Z-score approach: standardized deviation for each observed value. Outliers as: $|z_{ij}| > T$ (e.g., T=3)

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{\hat{\sigma}^2_{within}}}; \quad \hat{\sigma}^2_{within} = \frac{1}{n(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} \left(x_{ij} - \bar{x}_i\right)^2$$

- Interquartile range (IQR): a non-parametric method; works when deviates do not follow a normal distribution.

$$d_{ij} < Q1 - 1.5 \cdot IQR \text{ or } d_{ij} > Q3 + 1.5 \cdot IQR; \; IQR = Q3 - Q1$$

- Mahalanobis distance (MD): a multivariate measure accounting for correlations among variables.

$$D_M^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi_k^2; \quad D_M^2(\boldsymbol{x}) > \chi_{k,1-\alpha}^2$$

CDCB COUNCIL ON DAIRY CATTLE BREEDING

Outlier rates (%) of milk fat percentage

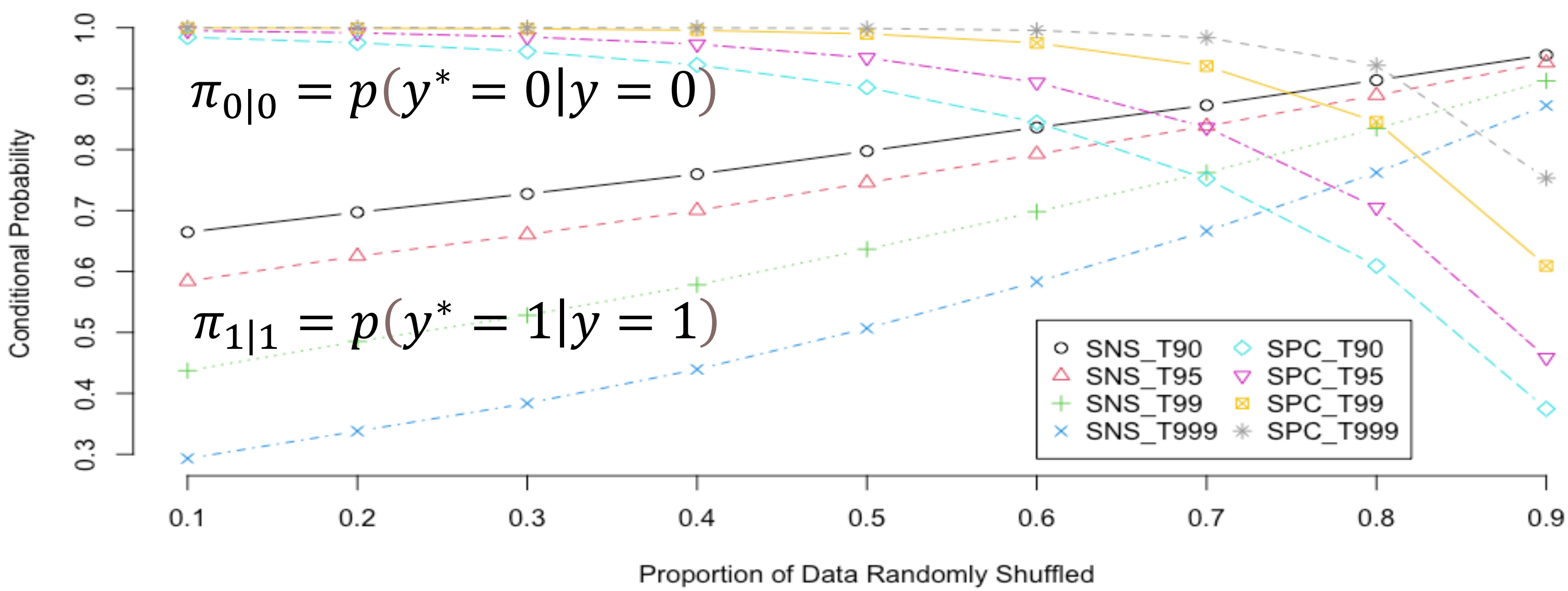
Dataset	N	Mean	$\hat{\sigma}_b^2$	$\hat{\sigma}_w^2$	ICC	I-ICC	Outliers (%)						
							Z-score	IQR	MD	I-ICC90	I-ICC95	I-ICC99	I-ICC999
A	15,995	4.04	0.274	0.244	0.529	0.681 (0.233)	0.75	1.14	4.55	4.55	2.96	1.37	0.54
B	13,336	4.05	0.308	0.154	0.666	0.801 (0.208)	0.48	0.73	2.81	2.41	1.49	0.73	0.29
C	8,363	4.51	0.256	0.248	0.508	0.699 (0.239)	0.97	1.40	4.63	5.03	3.50	1.94	1.05
D	11,182	4.15	0.368	0.283	0.565	0.717 (0.236)	1.09	1.69	6.55	4.24	2.55	1.03	0.33
S0	48,876	4.15	0.308	0.211	0.593	0.667 (0.200)	0.01	0.06	1.12	0.60	0.13	<0.01	<0.01
S1	48,876	4.15	0.278	0.242	0.534	0.630 (0.215)	0.16	0.33	2.60	2.45	1.10	0.24	0.04
S2	48,876	4.15	0.246	0.274	0.473	0.590 (0.228)	0.32	0.63	4.15	5.23	2.77	0.78	0.15
S3	48,876	4.15	0.215	0.305	0.413	0.546 (0.238)	0.47	0.92	5.70	9.19	5.36	1.78	0.41
S4	48,876	4.15	0.184	0.336	0.354	0.500 (0.245)	0.61	1.19	7.17	14.5	9.16	3.53	0.98
S5	48,876	4.15	0.153	0.367	0.294	0.448 (0.248)	0.76	1.47	8.71	22.0	15.1	6.77	2.35
S6	48,876	4.15	0.121	0.399	0.233	0.389 (0.246)	0.92	1.76	10.2	32.0	23.9	12.6	5.39
S7	48,876	4.15	0.091	0.429	0.174	0.323 (0.237)	1.05	2.04	11.7	44.9	36.1	22.3	11.7
S8	48,876	4.15	0.059	0.461	0.113	0.243 (0.217)	1.21	2.34	13.3	61.6	53.6	39.2	25.6
S9	48,876	4.15	0.029	0.491	0.055	0.148 (0.175)	1.37	2.62	14.8	80.6	75.6	65.4	53.3

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Outlier rates (%) of milk protein percentage

Dataset	N	Mean	$\hat{\sigma}_b^2$	$\hat{\sigma}_w^2$	ICC	I-ICC	Outliers (%)						
							Z-score	IQR	MD	I-ICC90	I-ICC95	I-ICC99	I-ICC999
A	15,995	3.09	0.087	0.007	0.927	0.943 (0.085)	0.44	0.60	1.69	1.83	1.33	0.79	0.50
B	13,336	3.19	0.088	0.010	0.897	0.929 (0.110)	1.03	1.32	3.27	3.65	2.88	1.80	1.14
C	8,363	3.25	0.114	0.012	0.903	0.940 (0.118)	1.73	2.11	5.02	4.27	3.42	2.33	1.34
D	11,182	3.18	0.103	0.015	0.875	0.926 (0.118)	1.39	1.78	4.53	4.25	3.13	1.91	1.14
C0	48,876	3.17	0.095	0.010	0.902	0.910 (0.077)	0.03	0.10	0.97	1.01	0.28	0.01	<0.01
C1	48,876	3.17	0.086	0.020	0.813	0.867 (0.149)	2.49	3.06	7.26	8.05	6.27	4.40	2.93
C2	48,876	3.17	0.076	0.029	0.721	0.820 (0.195)	5.00	6.08	13.6	15.9	13.2	9.76	6.76
C3	48,876	3.17	0.067	0.039	0.633	0.768 (0.229)	7.44	9.03	19.9	24.6	20.9	16.0	11.5
C4	48,876	3.17	0.057	0.048	0.543	0.710 (0.255)	9.93	12.0	26.2	34.1	29.6	23.4	17.6
C5	48,876	3.17	0.048	0.058	0.451	0.642 (0.276)	12.4	15.0	32.5	44.8	39.7	32.3	25.4
C6	48,876	3.17	0.038	0.067	0.359	0.565 (0.288)	15.0	18.1	38.8	56.4	51.2	42.9	35.2
C7	48,876	3.17	0.028	0.077	0.270	0.476 (0.290)	17.4	21.0	45.1%	68.5	63.6	55.3	47.1
C8	48,876	3.17	0.019	0.086	0.180	0.368 (0.275)	19.9	24.0	51.4	80.9	77.0	69.9	62.2
C9	48,876	3.17	0.010	0.096	0.090	0.229 (0.229)	22.3	27.0	57.7	92.2	90.2	86.0	81.0

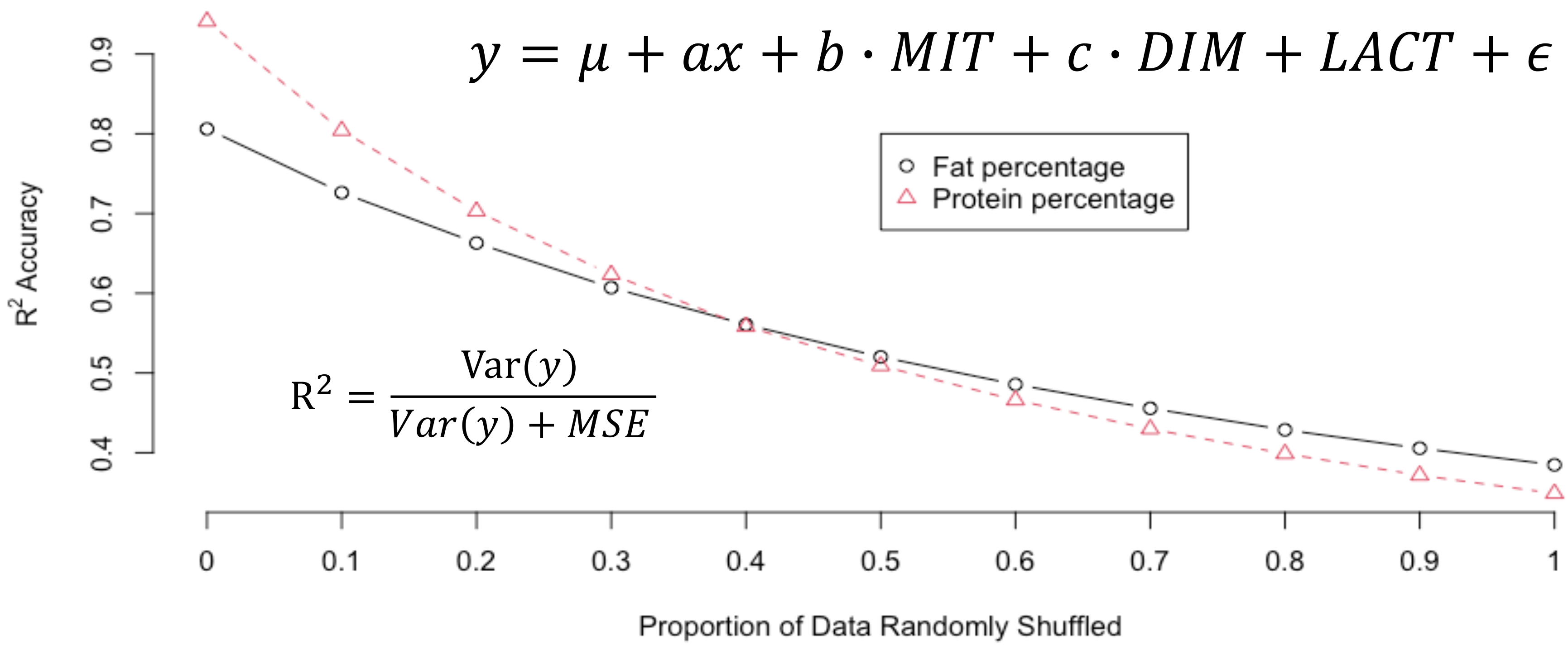
Sensitivity and specificity: protein percentage



COUNCIL ON DAIRY CATTLE BREEDING

Effects of data shuffling on

$$y = \mu + ax + b \cdot MIT + c \cdot DIM + LACT + \epsilon$$



$$R^2 = \frac{\mathrm{Var}(y)}{Var(y) + MSE}$$

Take-home messages

- Conventional univariate and multivariate outlier detection methods are not well-suited for identifying outliers of milk components.
- We proposed a new metric, namely I-ICC, which proved informative and practical for evaluating milk component data quality at the cow-day level.
- Record shuffling, like many other milking record errors, compromised the accuracy of estimated daily fat and protein percentages.

CDCB
COUNCIL ON DAIRY CATTLE BREEDING

Questions and comments?

CDCB
COUNCIL ON DAIRY CATTLE BREEDING