Optimal Data Analysis
Vol. 5 (September 17, 2016), 74-78

2155-0182/10/$3.00

Novometrics *vs.* Correlation: Age and Clinical Measures of PCP Survivors

Paul R. Yarnold, Ph.D. and Charles L. Bennett, M.D., Ph.D.
Optimal Data Analysis, LLC University of South Carolina, College of Pharmacy

The use of ODA to maximize the predictive accuracy of linear Pearson correlation[1] (*r*) models was previously considered.[2,3] The present paper discusses and illustrates the maximum-accuracy alternative to *r*.

The partitioning algorithm (PA) used herein is used in analysis of interrupted and continuous time-series.[4-7] For time-ordered series the PA involves creating a set of dummy codes that indicate if each observation (i.e., data point in the series) was recorded at or before (0), versus after (1), each sequential step in the time-series.[4] Likewise, for value-ordered series the PA involves creating a set of dummy codes to indicate if each observation (i.e., sample data point) had a score equal to or less than (0), versus greater than (1), each sequential value in the class variable measurement scale. In either situation each dummy class variable is employed to model the attribute in a separate novometric analysis, and the model with the largest ESS (if the competing models have the same number of strata), or the smallest D statistic (a normed index of distance separating empirical versus ideal model for the application), is selected as the globally-optimal model of the relationship between class variable and attribute for the sample.[3,8]

This is illustrated here for a sample[9] of 1,473 survivors of PCP pneumonia: if and how each of eight clinical variables all measured at study intake (Table 1) relate to age is assessed using both *r* and novometrics (Tables 2 and 3).

Table 1: Descriptive Summary of Variables

Variable	N	Mean	SD	CV (%)
Age	1,460	38.01	7.92	20.85
Diastolic	1,469	106.62	14.86	13.93
Systolic	1,468	63.68	12.14	19.07
Albumin	1,252	3.00	0.68	22.72
WBC	1,456	6.08	3.26	53.61
Creatinine	1,433	1.12	1.12	99.83
CD4	1,089	46.52	24.72	53.15
AAO2	1,459	33.59	6.36	18.95
HCT	694	230.59	371.23	160.99

Table 2: Clinical Variables and Age: Correlation

Variable	r	r^2	$p <$
Diastolic BP	0.115	0.0133	0.0001
Systolic BP	0.091	0.0083	0.0005
Albumin	-0.062	0.0039	0.0292
White Blood Cells	0.037	0.0014	0.17
Creatinine	0.033	0.0011	0.21
CD4 Count	0.010	0.0001	0.80
AAO2 Gradient	0.085	0.0073	0.0054
HCT	-0.001	0.0001	0.96

2155-0182/10/$3.00

Table 3: Clinical Variables and Age: Novometrics

Variable	Cut-Points	ESS	$p <$
Diastolic BP	56, 60.5	30.78	0.0014
Systolic BP	56, 119.5	25.79	0.012
Albumin	31, 2.55	10.35	0.011
White Blood Cells	54, 4.25	18.92	0.036
Creatinine	53, 1.15	25.23	0.0004
CD4 Count	26, 93.5	18.90	0.14
AAO2 Gradient	52, 43.855	20.50	0.0084
HCT	44, 32.85	8.78	0.046

Creating Age Class Variables

Statistical power analysis indicated that a minimum of N=32 observations should be classified into all model strata to attain 90% power to detect a moderate effect.[3] A total of 24 patients were 24 years or younger, and 41 were 25 years or younger, thus 25 years of age was selected as the minimum age for the PA. A total of 27 patients were 57 or older, and 40 were 56 or older, thus 56 years of age was selected as the maximum age for the PA. Ages in this domain were all represented by data so 32 age class dummy variables were constructed by the PA. All models (Table 3) had two predicted patient strata, each consisting of more than 32 patients.

Age and Diastolic Blood Pressure

Pearson correlation explains a paltry 1.33% of the variation in diastolic blood pressure (DBP) as a positive linear function of age (Table 2). Although minute, this nevertheless was the strongest effect that was identified by correlation analysis in this study. Such a weak model is only able to accurately predict the DBP of observations scoring at or near the sample mean on this variable, thus yielding an ESS statistic close to zero—the level of predictive accuracy expected by chance.[2,3]

As occurred for correlation analysis, the strongest effect identified by novometrics in this study was obtained for DBP. The resulting ODA model was: if DBP $\leq$ 60.5 then predict age $\leq$ 56; otherwise predict age > 56. The model was statistically significant (Table 3), and had ESS = 30.78—an effect of moderate strength.[3,5]

The confusion matrix for this model applied to the data is provided in Table 4. The level of predictive accuracy achieved for patients $\leq$56 years of age with actual DBP$\leq$60.5 (i.e., 51.37%) is close to the level of accuracy expected by chance (50%). In contrast, 27 of 36 (79.41%) patients who were at least 57 years of age had DBP>60.5.

Table 4: Confusion Matrix for DBP: Cut-Point $\leq$60.5 mmHg

		Predicted Age		
		$\leq$56	>56	
Actual	$\leq$56	730	691	51.37%
Age	>56	7	27	79.41%

A statistically significant effect of moderate strength emerged: 79.41% (approximately 4 of 5) patients aged >56 years had DBP>60.5.

Age and Systolic Blood Pressure

Pearson correlation explained less than one percent of the variation in systolic blood pressure (SBP) as a positive linear function of age (Table 2), a statistically significant finding.

As occurred for correlation analysis the second-strongest effect identified using novometrics emerged for SBP. The model was: if SBP $\leq$ 119.5 then predict age $\leq$ 56; otherwise predict age > 56. The model was statistically significant (Table 3), and had ESS = 25.79—an effect of borderline moderate strength.[3,5]

The confusion matrix for this model applied to the data is provided as Table 5. The model accuracy achieved for patients $\leq$56 years of age with actual SBP>119.5 (44.12%) is marginally lower than the accuracy expected by chance. In contrast, 1,172 of 1,435 (81.67%) patients $\leq$56 years of age had SBP$\leq$119.5.

2155-0182/10/$3.00

Table 5: Confusion Matrix for SBP: Cut-Point ≤119.5 mmHg

		Predicted Age		
		≤56	>56	
Actual	≤56	1,172	263	81.67%
Age	>56	19	15	44.12%

A statistically significant effect of moderate strength emerged: 81.67% (7 of 8) patients aged ≤56 years had SBP≤119.5.

Age and Albumin

Pearson correlation explained four-tenths of one percent (0.39%) of the variation in albumin level as a negative linear function of age: this result was statistically significant only at a generalized (per-comparison) criterion[3,5] (Table 2).

The optimal model was: if albumin ≤ 2.55 then predict age ≤ 31; otherwise predict age > 31. The model was experimentwise statistically significant for a study with ≤4 "less significant" *p* values (Table 3), and had ESS = 10.35—a relatively weak effect (note the worse-than-chance performance for patients with an actual albumin level >2.55).[3,5] Table 6 is the confusion matrix for this model.

Table 6: Confusion Matrix for Albumin: Cut-Point ≤2.55 g/dL

		Predicted Age		
		≤31	>31	
Actual	≤31	218	39	84.82%
Age	>31	741	254	25.53%

A statistically significant effect of relatively weak strength emerged: 84.82% (7 of 8) patients aged ≤31 years had albumin ≤2.55.

Age and White Blood Cell (WBC) Count

Pearson correlation explained one-seventh of one percent of the variation in WBC count as a positive linear function of age (Table 2). This result was not statistically significant.

The optimal model was: if WBC ≤ 4.25 then predict age ≤ 54; otherwise predict age > 54. This result was only statistically significant at a generalized (per-comparison) criterion (Table 3), and had ESS = 18.92—a relatively weak effect (note the worse-than-chance performance for patients with an actual WBC level ≤2.55). Table 7 presents the confusion matrix for this model.

Table 7: Confusion Matrix for WBC Count: Cut-Point ≤4.25 x 10^3/μL

		Predicted Age		
		≤54	>54	
Actual	≤54	473	929	33.74%
Age	>54	8	46	85.19%

A statistically significant relatively weak effect emerged: 85.19% (7 of 8) patients aged >54 years had a WBC count >4.25.

Age and Creatinine

Pearson correlation explained one-ninth of one percent of the variation in creatinine level as a positive linear function of age (Table 2). This result was not statistically significant.

The optimal model was: if creatinine ≤ 1.15 then predict age ≤ 53; otherwise predict age > 53. The model was statistically significant (Table 3), with ESS = 25.23—a borderline moderate effect (note the chance-level performance for patients with actual creatinine level >1.15). Table 8 is the confusion matrix for this model.

Table 8: Confusion Matrix for Creatinine: Cut-Point ≤1.15 mg/dL

		Predicted Age		
		≤53	>53	
Actual	≤1.15	1,085	285	79.20%
Age	>1.15	34	29	46.03%

2155-0182/10/$3.00

A statistically significant, moderate effect emerged: 79.20% (4 of 5) patients aged ≤53 years had creatinine ≤1.15.

Age and CD4 Count

Pearson correlation explained less than one percent of one percent of the variation in CD4 count as a positive linear function of age (Table 2). This result was not statistically significant. Likewise, novometric analysis found no statistically reliable model for CD4 count (Table 3). This is the only variable of the eight studied without an association with age, when assessed using novometrics.

Age and AAO2 Gradient

Pearson correlation explained three-fourths of one percent (0.73%) of the variation in AAO2 level as a positive linear function of age (Table 2). This result was experimentwise statistically significant for an investigation with nine or fewer "less significant" *p* values.[5]

The optimal model was: if AAO2 ≤ 43.855 predict age ≤ 52; otherwise predict age > 52. The model was statistically significant (Table 3), with ESS = 20.50—a relatively weak effect (note the chance-level accuracy for patients with actual AAO2 level ≤43.855). Table 8 is the confusion matrix for this model.

Table 8: Confusion Matrix for AAO2: Cut-Point ≤43.855 mmHg

		Predicted Age ≤52	>52	
Actual	≤52	525	499	51.27%
Age	>52	20	45	69.23%

A relatively weak, statistically significant effect emerged: 69.23% (7 of 10) patients aged >52 years had AAO2 >43.855.

Age and HCT

Pearson correlation explained less than one percent of one percent of the variation in HCT level as a negative linear function of age: this was not statistically significant (Table 2).

The optimal model was: if HCT ≤ 32.85 then predict age ≤ 44; otherwise predict age > 44. This result was statistically significant at a generalized criterion (Table 3), and had ESS = 8.78—a relatively weak effect (note the worse-than-chance performance for patients with an actual HCT level ≤32.85). Table 9 gives the confusion matrix for this model.

Table 9: Confusion Matrix for HCT: Cut-Point ≤32.85 mg/dL

		Predicted Age ≤44	>44	
Actual	≤44	521	654	44.34%
Age	>44	101	183	64.44%

A statistically significant, weak effect emerged: 64.44% (13 of 20) patients aged >44 years had HCT >32.85, compared with 55.66% (654 of 1,175) patients aged ≤44 years.

Comments

Exploring journal stacks of a research library unveils seemingly endless literature featuring innumerable correlation coefficients that, while statistically significant, nevertheless explain a meaningless or trivial proportion of variance. Additionally, an incalculable treasure trove exists of published correlations that failed investigator criteria for statistical significance. However, as seen presently, the absence of a strong and/or statistically significant *r* between two variables does not mean that no *linear relationship* exists between the two variables.

2155-0182/10/$3.00

For example in this study involving eight variables—all essentially uncorrelated with age, novometric analysis identified two-strata parses (defined by linear inequalities) for seven of the eight variables examined, and for three variables the predictive accuracy achieved by the optimal models was in the moderate range. Discovery of between-group effects such as reported herein is crucial in causal inference research, for example to assess comparability of different groups[10] and identify variables to include in propensity score development.[11,12] Further research comparing *r* versus novometric models of association for ordered class variables is clearly warranted.

References

[1]Pearson K (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*, 240-242.

[2]Yarnold PR, Bryant FB, Soltysik RC (2013). Maximizing the accuracy of multiple regression models via UniODA: Regression *away from* the mean. *Optimal Data Analysis, 2*, 19-25.

[3]Yarnold PR, Soltysik RC (2016). *Maximizing predictive accuracy*. Chicago, IL: ODA Books. DOI: 10.13140/RG.2.1.1368.3286

[4]Linden A, Yarnold PR (2016) Using machine learning to identify structural breaks in single-group interrupted time series designs. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12544

[5]Yarnold PR, Soltysik RC (2005) *Optimal data analysis: A Guidebook with Software for Windows* Washington, DC: APA Books.

[6]Yarnold PR (2013). The most recent, earliest, and *K*th significant changes in an ordered series: Traveling backwards in time to assess when annual crude mortality rate most recently began increasing in McLean County, North Dakota. *Optimal Data Analysis, 2*, 143-147.

[7]Yarnold PR (2013). Ascertaining an individual patient's *symptom dominance hierarchy*: Analysis of raw longitudinal data induces Simpson's Paradox. *Optimal Data Analysis, 2*, 159-171.

[8]Yarnold PR, Soltysik RC (2014). Globally optimal statistical classification models, I: Binary class variable, one ordered attribute. *Optimal Data Analysis, 3*, 55-77.

[9]Yarnold PR, Soltysik RC, Bennett CL (1997). Predicting in-hospital mortality of patients with AIDS-related *Pneumocystis carinii* pneumonia: An example of hierarchically optimal classification tree analysis. *Statistics in Medicine, 16*, 1451-1463.

[10]Yarnold PR (2016). UniODA *vs*. chi-square: Describing baseline data from the National Pressure Ulcer Long-Term Care Study (NPULS). *Optimal Data Analysis, 5*, 24-28.

[11]Linden A, Yarnold PR (2016). Using machine learning to assess covariate balance in matching studies. *Journal of Evaluation in Clinical Practice*. DOI: 10.1111/jep.12538

[12]Linden A, Yarnold PR (2016). Combining machine learning and propensity score weighting to estimate causal effects in multivalued treatments. *Journal of Evaluation in Clinical Practice*.

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported. Dr. Yarnold's work was partially funded by a grant from the National Cancer Institute (1R01CA165609-01A1).

Mail: Optimal Data Analysis, LLC
6348 N. Milwaukee Ave., #163
Chicago, IL 60646