Optimal Data Analysis
Vol. 5 (September 19, 2016), 83-85

2155-0182/10/$3.00

Novometrics *vs.* Regression Analysis: Literacy, and Age and Income, of Ambulatory Geriatric Patients

Paul R. Yarnold, Ph.D.
Optimal Data Analysis, LLC

A convenience sample of 293 ambulatory women patients, all older than 65 years of age, were surveyed in a general medicine clinic.[1] Correlation (*r*), multiple regression analysis (MRA), and novometric analysis were used to model the relationship of scores (even integers) on the TOFHLA literacy measure[2] (the dependent or class variable) with age (recorded to two significant digits to the right of the decimal) and income (measured as 1 to 8, inclusive, integer annual increments of $10,000). Regression- and novometric-based findings are contrasted.
.

Table 1 presents a descriptive summary of the ordered variables investigated herein.

Table 1: Descriptive Summary: Study Variables

Variable	N	Mean	SD	CV (%)
TOFHLA	360	49.87	21.08	42.28
Age	450	77.62	6.81	8.78
Income	416	2.96	1.72	22.95

Creating Literacy Class Variables

Statistical power analysis indicated that N≥32 observations are needed for model strata (endpoints) for 90% power to detect a moderate effect with $p<0.05$.[3] A total of 34 patients had TOFHLA scores≤18, selected as the minimum score for the partitioning algorithm (PA).[4-6] A total of 23 patients had TOFHLA scores>70, and 74 patients had scores≥70, so 70 was the maximum score for PA. Data populated every TOFHLA score in this domain, so PA created a total of 27 literacy class dummy variables.

Literacy and Age

The *r* between literacy and age, -0.349, reveals 12.2% of the variance in literacy scores is explained as a negative linear function of age ($p<0.0001$). Models having such weak R^2 values are only able to accurately predict the scores of observations scoring at or near the sample mean on the dependent variable—and thus yield an ESS statistic close to zero, the level of predictive accuracy expected by chance.[3,7] For this model ESS=2.70, D=1,227.1—an extremely weak effect.[3]

2155-0182/10/$3.00

For novometrics the optimal model was: if age≤84.5 then predict TOFHLA>20; otherwise predict TOFHLA≤20. This model was statistically significant (p<0.0001) and produced a moderate effect: ESS=39.33, D= 3.08. Table 4 presents the confusion matrix for this model used in training (total sample) analysis.

Table 4: Confusion Matrix: Age Model

		Predicted TOFHLA		
		≤20	>20	
Actual	≤20	20	20	50.00%
TOFHLA	>20	27	226	89.33%

Thus a statistically significant, moderate effect emerged in training analysis: 89.33% (9 in 10) patients aged ≤84.5 years had a TOFHLA score>20, compared to 50.00% (5 in 10) of the patients aged >84.5 years. Model sensitivity for actual TOFHLA score≤20 fell to 47.50% in jackknife analysis (ESS=36.83).

Literacy and Income

The r=0.419 for literacy and income reveals 17.6% of the variance in literacy scores is explained as a positive linear function of income (p<0.0001). For this model ESS=2.30, D=1,441.7—an extremely weak effect.

For novometrics the optimal model was: if income≤2.5 then predict TOFHLA≤46; otherwise predict TOFHLA>46. This model was statistically significant (p<0.0001): ESS=41.65, D= 2.80, indicates a moderate effect. Table 5 gives the confusion matrix for the model (predictive accuracy was stable in jackknife analysis).

Table 5: Confusion Matrix: Income Model

		Predicted TOFHLA		
		≤46	>46	
Actual	≤46	76	31	71.03%
TOFHLA	>46	47	113	70.62%

A statistically significant effect of moderate strength emerged: 7 of 10 patients with an annual income ≤$20,000 had a TOFHLA score ≤46; and 7 of 10 patients with an annual income >$20,000 had a TOFHLA score>46.

Literacy, and Age and Income

For MRA analysis TOFHLA score was treated as the dependent measure and modeled as a simple main-effects function of the age and income independent variables.[8] The model (coefficients are reported to two significant digits to the right of the decimal) was: score = 114.16 – 1.02 * age + 4.76 * income. The model explained 27.90% of the variation in patient TOFHLA scores: $F(2,264)=51.1$, p<0.0001. These findings indicate a statistically significant linear association of moderate strength (not accounting for chance) exists between TOFHLA score and age and/or income. The source table for individual independent variables included in the MRA model (sum of squares for variable-entered-last method[8]) revealed a statistically significant negative association (with TOFHLA score) of age, and a statistically significant positive association of income (p's<0.0001). However, for this MRA model ESS=1.63 and D=2,046.8 in training analysis, an extremely weak effect.

Novometric analysis revealed that no multivariable model existed for the present data. That is, the best (globally optimal) model with the lowest D statistic was already identified, using income to construct a binary parse of age. The next lowest D statistic was obtained for a four-strata model that involved a three-branch parse of income, with age loading on the middle branch: for that model, D=4.307.

Examination of the confusion table for the optimal model for age (Table 4) reveals that the number of *misclassified* patients having an actual TOFHLA score ≤20 (N=20), and also having an actual TOFHLA score >20 (N=27), are both too small to justify further statistical analysis—on the basis of inadequate statistical

2155-0182/10/$3.00

power.[3] Likewise, for income the number of misclassified patients with actual TOFHLA score ≤46 (N=31) is too small to justify adding additional attributes due to inadequate power to test *a priori* hypotheses.

In the present sample identifying an optimal model that used both attributes necessitated doubling the number of model endpoints, indicating that adding more attributes to either single-attribute model is statistically unjustified. Using a more granular, precise measure of income may yield a more accurate model with greater ESS and lower D than was obtained using the integer measure.[3,9] Of course, a more accurate model would further reduce the number of misclassified observations and leave even less residual opportunity for other attributes, such as age, to enter a multivariable model to predict literacy.

References

[1]Arozullah AM, Lee SD, Khan T, Kurup S, Ryan J, Bonner M, Soltysik RC, Yarnold PR (2006). The Roles of low literacy and social support in predicting the preventability of hospital admission. *Journal of General Internal Medicine, 21*, 140-145.

[2]Parker RM, Baker DW, Williams MV, Nurss JR (1995). The test of functional health literacy in adults: A new instrument for measuring patients' literacy skills. *Journal of General Internal Medicine, 10,* 537-541.

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

[4]Yarnold PR, Linden A (2016). Novometric analysis with ordered class variables: The optimal alternative to linear regression analysis. *Optimal Data Analysis, 5*, 65-73.

[5]Yarnold PR, Bennett CL (2016). Novometrics *vs*. correlation: Age and clinical measures of PCP survivors. *Optimal Data Analysis, 5*, 74-78.

[6]Yarnold PR, Bennett CL (2016). Novometrics *vs*. multiple regression analysis: Age and clinical measures of PCP survivors. *Optimal Data Analysis, 5*, 79-82.

[7]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.

[8]Licht MH (1995). Multiple regression and correlation. In: Grimm LG, Yarnold PR (Eds.), *Reading and Understanding Multivariate Statistics*. Washington, DC: APA Books, 1995, pp. 19-64.

[9]Yarnold PR (2014). "Breaking-up" an ordinal variable can reduce model classification accuracy. *Optimal Data Analysis, 3*, 19.

Author Notes

The study analyzed de-individuated data and was exempt from Institutional Review Board review. No conflict of interest was reported.

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